Predictive Modelling Approach to Data-Driven Computational Preventive Medicine

Aldraimli, M.

This is a PhD thesis awarded by the University of Westminster.

https://doi.org/10.34737/w4ww2

بسم الله الرحمن الرحيم

Predictive Modelling Approach to Data-Driven Computational Preventive Medicine

Doctor of Philosophy Thesis

Mahmoud Aldraimli

Department of Computer Science and Engineering
College of Design, Creative and Digital Industries
University of Westminster

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Data Science at the University of Westminster 2017 – 2023
For more information, please contact:

University of Westminster
School of Computer Science and Engineering
115 New Cavendish Street
London UK
W1W 6UW
E-mail: m.aldraimli2@westminster.ac.uk

This research programme was completed within the Health Innovation Ecosystem, the Health and Social Care Modelling Group at the University of Westminster and the Radiotherapy Machine Learning Network at the University of Manchester.

Printed in the United Kingdom of Great Britain

UNIVERSITY OF
WESTMINSTER

brief contents

contents

abstract

This thesis contributes novel predictive modelling approaches to data-driven computational preventive medicine and offers an alternative framework to statistical analysis in preventive medicine research. In the early parts of this research, this thesis presents research by proposing a synergy of machine learning methods for detecting patterns and developing inexpensive predictive models from healthcare data to classify the potential occurrence of adverse health events. In particular, the data-driven methodology is founded upon a heuristic-systematic assessment of several machine-learning methods, data pre-processing techniques, models' training estimation and optimisation, and performance evaluation, yielding a novel computational data-driven framework, Octopus.

Midway through this research, this thesis advances research in preventive medicine and data mining by proposing several new extensions in data preparation and preprocessing. It offers new recommendations for data quality assessment checks, a novel multi-method imputation (MMI) process for missing data mitigation, a novel imbalanced resampling approach, and minority pattern reconstruction (MPR) led by information theory. This thesis also extends the area of model performance evaluation with a novel classification performance ranking metric called XDistance.

In particular, the experimental results show that building predictive models with the methods guided by our new framework (Octopus) yields domain experts' approval of the new reliable models' performance. Also, performing the data quality checks and applying the MMI process led healthcare practitioners to outweigh predictive reliability over interpretability. The application of MPR and its hybrid resampling strategies led to better performances in line with experts' success criteria than the traditional imbalanced data resampling techniques. Finally, the use of the XDistance performance ranking metric was found to be more effective in ranking several classifiers' performances while offering an indication of class bias, unlike existing performance metrics.

The overall contributions of this thesis can be summarised as follow. First, several data mining techniques were thoroughly assessed to formulate the new Octopus framework to produce new reliable classifiers. In addition, we offer a further understanding of the impact of newly engineered features, the physical activity index (PAI) and biological effective dose (BED). Second, the newly developed methods within the new framework. Finally, the newly accepted developed predictive models help detect adverse health events, namely, visceral fat-associated diseases and advanced breast cancer radiotherapy toxicity side effects. These contributions could be used to guide future theories, experiments and healthcare interventions in preventive medicine and data mining.

declaration

I declare that this thesis is composed entirely by myself. It was not submitted in whole or part to any previous application for a degree. The presented work is my own and was fully funded by the Quintin Hogg Trust and partially by the Science and Technology Facilities Council. This research was developed as joint research with collaborators from the UK Biobank, REQUITE Consortium, various national and international research institutions, universities and the UK National Health Services (NHS), except where I state otherwise by a reference or acknowledgement. The opinions expressed in this thesis do not reflect the view of the UK Biobank, the National Health Services (NHS), or the Department of Health in the United Kingdom.

Mahmoud Aldraimli

research impact

Parts of this research-related work were disseminated in scientific journals and conferences proceedings:

- ❒ Aldraimli, M., Soria, D., Parkinson, J., Whitcher, B., Thomas, E.L., Bell, J.D., Chaussalet, T.J. and Dwek, M.V., 2019, September. *Machine Learning Classification of Females Susceptibility to Visceral Fat Associated Diseases*. In Mediterranean Conference on Medical and Biological Engineering and Computing (pp. 679-693). Springer, Cham.

- ❒ Aldraimli, M., Soria, D., Parkinson, J., Thomas, E.L., Bell, J.D., Dwek, M.V. and Chaussalet, T.J., 2020. *Machine learning prediction of susceptibility to visceral fat associated diseases*. Health and Technology, 10(4), pp.925-944.

- ❒ Aldraimli, M., Nazyrova, N., Djumanov, A., Sobirov, I. and Chaussalet, T.J., 2020, October. *A Comparative Machine Learning Modelling Approach for Patients' Mortality Prediction in Hospital Intensive Care Unit.* In The International Symposium on Bioinformatics and Biomedicine (pp. 16-31). Springer, Cham.

❒ Aldraimli, M., Osman, S., Grishchuck, D., Ingram, S., Lyon, R., Mistry, A., Oliveira, J., Samuel, R., Shelley, L.E., Soria, D. and Dwek, M.V., 2022. *Development and Optimization of a Machine-Learning Prediction Model for Acute Desquamation After Breast Radiation Therapy in the Multicenter REQUITE Cohort.* Advances in Radiation Oncology, 7(3), p.100890.

❒ Aldraimli, M., Soria, D., Grishchuck, D., Ingram, S., Lyon, R., Mistry, A., Oliveira, J., Samuel, R., Shelley, L.E., Osman, S. and Dwek, M.V., 2021. *A data science approach for early-stage prediction of Patient's susceptibility to acute side effects of advanced radiotherapy*. Computers in biology and medicine, 135, p.104624.

Further impact of this authentic interdisciplinary research and work was featured in news articles:

❒ Medical Research News – The Medical Express
Medicalxpress.com. 2021. *Data science approach helps oncologists predict which patients will suffer side effects from radiotherapy*. [online] Available at: <https://medicalxpress.com/news/2021-07-science-approach-oncologists-patients-side.html> [Accessed 21 September 2022].

❒ The University of Westminster News
Westminster.ac.uk. 2019. *Westminster PhD student leads experts at the STFC Global Challenge Network event in Advanced Radiotherapy*. [online] Available at: <https://www.westminster.ac.uk/news/ westminster-phd-student-leads-experts-at-the-stfc-global-challenge-network-event-in-advanced> [Accessed 21 September 2022].

- The University of Westminster News
  Westminster.ac.uk. 2020. Health Innovation Ecosystem Organised Machine Learning Bootcamp ‘*Exordium’ to bust myths of Artificial Intelligence*’. [online] Available at: <https://www.westminster.ac.uk/ news/health-innovation-ecosystem-organised-machine-learning-bootcamp-exordium-to-bust-myths-of-artificial> [Accessed 21 September 2022].

- The University of Westminster News
  Westminster.ac.uk. 2021. *Westminster researchers use data science approach to assess whether cancer patients will suffer side effects from radiotherapy*. [online] Available at: <https://www.westminster.ac.uk/ news/westminster-researchers-use-data-science-approach-to-assess-whether-cancer-patients-will-suffer-side-effects-from-radiotherapy> [Accessed 21 September 2022].

- The University of Westminster News
  Westminster.ac.uk. 2020. *Research Collaboration receives Top Award at International Symposium on Bioinformatics and Biomedicine*. [online] Available at: <https://www.westminster.ac.uk/news/research-collaboration-receives-top-award-at-international-symposium-on-bioinformatics-and> [Accessed 21 September 2022].

The development of this interdisciplinary thesis conceived valuable efforts in bridging the gap between data science and multiple other scientific communities, including life sciences, healthcare, business, arts and media, and cities and architectures, in conferences, symposia, seminars and workshops:

❒ XIV Mediterranean Conference on Medical and Biological Engineering and Computing. Coimbra, Portugal
Aldraimli, M., Soria, D.(presenter), Parkinson, J., Whitcher, B., Thomas, E.L., Bell, J.D., Chaussalet, T.J. and Dwek, M.V., 2019, September. '*Machine Learning Classification of Females Susceptibility to Visceral Fat Associated Diseases'*. In Mediterranean Conference on Medical and Biological Engineering and Computing (pp. 679-693). Springer, Cham.

❒ International Symposium on Bioinformatics and Biomedicine. Burgas, Bulgaria
Aldraimli, M., Nazyrova (presenter), N., Djumanov, A., Sobirov, I. and Chaussalet, T.J., 2020, October. '*A Comparative Machine Learning Modelling Approach for Patients' Mortality Prediction in Hospital Intensive Care Unit'*. In The International Symposium on Bioinformatics and Biomedicine (pp. 16-31). Springer, Cham.

❒ University of Westminster Doctoral Conference. London, UK
M. Aldraimli (presenter), D. Soria, M. V. Dwek, and T. Chaussalet, *'Building Machine Learning Models for Non-communicable Diseases Risk Prediction'* at the University of Westminster, Faculty of Science and Technology Doctoral Conference, 2018, New Cavendish Street Campus, London, UK

❒ The University of Westminster Student-Staff Research Conference. London, UK
M. Aldraimli (presenter), D. Soria, M. V. Dwek, and T. Chaussalet, *'Machine Learning Classification of Females Susceptibility to Visceral Fat Associated Diseases'* at the University of Westminster, School of Computer Science &

Engineering Student-Staff Research Conference, 2019, New Cavendish Street Campus, London, UK

❒ University of Westminster Researcher Network Third Annual Conference. London, UK
M. Aldraimli (presenter), D. Soria, M. V. Dwek, and T. J. Chaussalet, *‘Am I Fat Inside’* at the University of Westminster, Researcher Network Third Annual Conference, 2018, Marylebone Campus, London, UK

❒ University of Westminster Health and Innovation Ecosystem. London, UK
M. Aldraimli (instructor), and T. J. Chaussalet, *‘Exordium, The Machine Learning Bootcamp’* at the University of Westminster, Health and Innovation Ecosystem Staff and Researchers Workshop, 2020, Online, London, UK

❒ University of Westminster Health and Innovation Ecosystem. London, UK
M. Aldraimli (presenter), and T. Chaussalet, ‘*Machine Learning Takes on Difficult Health Screening and Treatment Tasks’* at the University of Westminster, Health Innovation Ecosystem (HIE) Seminar Spring, 2020, Online, London, UK

❒ University of Westminster Health and Innovation Ecosystem. London, UK
M. Aldraimli (presenter), and T. J. Chaussalet, *‘Gearing you up to dive into the unknown world of missing data in your research’* at the University of Westminster, Health Innovation Ecosystem (HIE) Seminar Spring, 2021, Online, London, UK

The journey of this interdisciplinary research was acknowledged with multiple awards on various occasions:

❒ Best Oral Presentation Award in International Symposium on Bioinformatics and Biomedicine

M. Aldraimli, N. Nazyrova (presenter), A. Djumanov, I. Sobirov, and T. J. Chaussalet, presented to ‘*New Machine Learning Modelling Approach for Patients’ Mortality Prediction in Hospital Intensive Care Unit*’ at International Symposium on Bioinformatics and Biomedicine, 2020 Burgas, Bulgaria.

❒ First Place Award for short presentations in Computer Science and Engineering sessions

M. Aldraimli (presenter), D. Soria, M. V. Dwek, and T. J. Chaussalet, ‘*Machine Learning Classification of Females Susceptibility to Visceral Fat Associated Diseases*’ at the University of Westminster, School of Computer Science & Engineering Student-Staff Research Conference, 2019, New Cavendish Street Campus, London, UK.

❒ Second Place Award for short presentations in Computer Science and Engineering sessions

M. Aldraimli (presenter), D. Soria, M. V. Dwek, and T. J. Chaussalet, ‘*Building Machine Learning Models for Non-communicable Diseases Risk Prediction*’ at the University of Westminster, Faculty of Science and Technology Doctoral Conference, 2018, New Cavendish Street Campus, London, UK

acknowledgement

I praise all those who go the extra mile to help others without obligations or reasons and give without expecting anything in return. I give my appreciation to those who talk the talk and walk the walk with their best intentions. Finally, I pass all the credit for completing this work to those scientists who contributed through their actions and knowledge in all parts of this research journey.

I want to acknowledge the Quintin Hogg Trust (QTH) research award no. 165435391-2017 for funding and supporting this research, the UK Radiotherapy Machine Learning Network (RTML) funded through the Advanced Radiotherapy Challenge+ by the Science and Technology Facilities Council (STFC), the REQUITE consortium study funded from the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 601826, the UK Biobank provided quality data to achieve the results with high confidence under agreement no. 23889, the University of Manchester and the Health Innovation Ecosystem at the University of Westminster for hosting the national interdisciplinary workshops that conceived this work.

Mahmoud Aldraimli

dedication

Dedicated to the memory of my father, Ahmed, who always believed in my ability to be successful in the academic arena. Sadly, you are gone, but your belief in me has made this journey possible. You always were a teacher by passion.

about this thesis

I can only show you the door. You're the one that has to walk through it.

Morpheus, *The Matrix*

When reading the table of contents, you will notice the diversity of the covered topics in this thesis. This interdisciplinary thesis provides researchers with a little bit of every consideration the author crossed to complete this research and its practical work—enough to get you and others to continue this research. Data mining is a very wide field, and adding it to preventive medicine makes it even wider, so wide indeed that a thesis ten times the size of this one wouldn't be able to cover it all. For each chapter, we picked a different aspect of the work we considered in the development cycle of this research. To narrow down the content, some hard decisions had to be made to keep this thesis from collapsing any bookshelf!

I hope it serves as a continuum point to this research field—your doorway into the exciting world of interdisciplinary research.

Thesis roadmap

Chapters 1 to 3 offer the theoretical background research, heuristically and systematically examining available data mining methods in the data science community and their acceptability in modelling real-world problems, notably healthcare. The deep understanding of such methods and their applications confine critical considerations and decisions taken to formulate our data-driven framework in chapter 4 necessary to execute the practical case studies in chapters 5 and 6 and finally conclude the contribution to the knowledge of this thesis in chapter 7.

❐ ***Chapter 1*** is an introduction that describes this research, the essential abstraction of its domain and its objectives in preventive medicine.

❐ ***Chapter 2*** starts a research review journey to formulate the data-driven modelling flow adopted to build the later chapters. It critically reviews the data cleaning, preparation, and transformation approaches available to be integrated into the formulated approach. In much of its content, it presents various debates regarding the effectiveness of their use and their deployment sequence.

❐ ***Chapter 3*** reviews and depicts a selection of machine learning algorithms considered for our predictive data-driven approach. It also presents a justification of choice and various debates regarding their effectiveness and adjustments based on reported experiments in the literature. We thoroughly review a variety of performance metrics available for evaluating classification modelling. In addition, the chapter analyses critical debates regarding each metric's suitability for assessing modelling outputs.

❐ ***Chapter 4*** formulates our new data-driven framework for modelling susceptibility to adverse health events. The formulation of our data-driven framework is based on an informed decision reached by systematic and heuristic considerations for the methods reviewed in the previous chapters. This chapter integrates multiple methods and provides justifications concerning their selection and sequence of application where needed. Within the framework, the chapter introduces a new approach to handling missing data, a new concept of information theory-driven resampling technique to enhance imbalanced learning and proposes a new performance metric design for imbalanced binary classifiers.

In Chapters 5 and 6, we apply the formulated data-driven framework to model two different case studies. Then we conclude our new contributions to knowledge in Chapter 8 of this thesis:

❒ ***Chapter 5*** presents the predictive modelling of visceral fat-associated diseases case study by applying the newly formulated framework. This is demonstrated by the collaborative development of new machine-learning prediction models approved by life sciences experts as potential inexpensive tools to screen patients for susceptibility to such diseases.

❒ ***Chapter 6*** applies the new framework to predict advanced radiation therapy acute skin toxicity. This is supported by the collaborative development of new machine-learning prediction models approved by clinicians and physicians as potential inexpensive tools to predict susceptibility to acute desquamation (skin toxicity).

❒ **Chapter 7** concludes this thesis and touches on the met objectives, limitations and implications of the work in a more-or-less independent and impartial manner, opening the door wide for future research opportunities.

Whom this thesis is for

This interdisciplinary thesis is a cross-contribution to data science and healthcare. Seasoned data scientists may see that we only scratch the surface of some of the topics. For other readers from other domains, there may be some prerequisites to fully grasp the thesis. A minimal understanding of statistics and machine learning is recommended before diving into this thesis's practical projects.

Therefore, this thesis is written in a language suited to a wide range of academic audiences to attract a larger interest from the healthcare and data science communities. Such an interest encourages new researchers from both domains to continue with this research. Therefore, any recommendations for improvements a reader can add to this thesis deserve the author's ultimate appreciation. Apart from the case studies, this thesis write-up lacked reviews. The author solely developed the narrative of this thesis with no one's guidance.

Models and data resources downloads

To enhance the interdisciplinary collaboration with healthcare experts, we acknowledged that coding is going to be a barrier to full involvement in this research; therefore, the author opted to use the Waikato Environment for Knowledge Analysis (WEKA) for practical examples in this research journey and thesis.

Over the past two decades, WEKA has developed into a much respected and stable machine learning environment, enhancing the reproducibility of data mining outputs and being widely used by academics.

Besides offering detailed steps for reproducibility, all the developed models are uploaded to the GitHub repository. These shall be made accessible to the readers; access can be granted upon making an online case application to the author.

The data is bounded by data transfer agreements, prohibiting sharing or transferring such resources outside the boundaries of approved researchers. However, the data can be accessed upon research application approvals. Applications are made directly to the UK Biobank and European REQUITE Consortium.

Chapter 1
Introduction

This chapter covers

- *Thesis statement*
- *Research Aim*
- *Research Objectives*

Machine Learning (ML), a subtopic of Artificial Intelligence (AI), provides trend discovery, similarities recognition and anomaly detection techniques used in multiple domains that require the discriminative power it offers. A domain of which is medical research. In medical research, ML models various health and clinical attributes with their combined contribution to enhancing patients' Quality of Life (QoL). It is also applied in treatment planning research and predicting the risk of developing non-communicable diseases in preventive medicine [1].

Moreover, it is applied in overall patient management research, such as hospital capacity forecasting and intelligent monitoring in Intensive Care Unit (ICU) admissions and mortalities [2]. However, there is still a debate about whether integrating ML methods in healthcare systems can enhance medical specialists' decision-making and improve diagnostics and healthcare quality [3].

ML model-based schemes provide entities for outcome inference from patient data. Patterns are extracted from experts' knowledge and the available electronic health records (EHR) to construct decision support systems. In some cases, depending on the model's complexity, specialists may face a problem interpreting the contributing factors that lead to the model's decision [3].

Decision support systems use symbolic ML techniques, such as supervised learning, to embed learning into the system while producing a systematic description of the health variables, ultimately categorising the health condition. Supervised learning uses a set of patients' cases, AKA examples, to systematically describe the clinical variables attributing to a health condition's determination [1]. A model's systematic description of variables can vary from simple rules in decision stumps to weights and coefficients (parameters) of complex functions. This type of learning is then taken to predict health conditions for new uncategorised cases. Such models may improve existing systems' performance for categorising existing conditions or embed new capabilities (knowledge) to classify new health conditions. The latter models can be considered initial hypotheses to drive further experimentation. In a research setting, learning from patients' data has several challenges [4]. The data are characterised by incompleteness, errors, noise and inexactness (inappropriate transformation). Various statistical and ML methods can deal with such challenges; some are proven more efficient than others. This adequacy is often associated with improved medical decision-making [4].

In health care, there is always the need for more effective early screening for health conditions. Some medical domains, like preventive medicine, promote preventive healthcare to improve patient well-being. The ultimate goal is to prevent disease, disability and death. In a way, it seeks ways to intercept the potential development of future adverse health events so that suitable countermeasures can be implemented to prevent such events [5]. The rising demand for preventive healthcare services has reshaped how modern medicine treats diseases. Preventative healthcare is rapidly emerging as the most inexpensive means of saving lives [6]. The first step to saving costs and lives can be achieved by identifying those susceptible to diseases and complications. Recent publications explored the potential of machine learning algorithms to carry out such tasks [7]. This interest also resulted in the publishing of additional guidance for healthcare professionals on models' evaluation and interpretation [7].

Although preventive medicine strategies may seem new, they have been examined historically for years [8]. The key problem is that many machine learning studies in this field, in general, are exploratory as it comes with a huge variety of methods. Also, the developed models may not have reliable predictions [9][10]. While these machine learning models show high accuracy on their own test datasets, their performance is questionable in real-world settings where the input data can vary drastically, resulting in unreliable prediction results [10].

This thesis aims to produce a new methodology and predictive models in the field of machine learning for preventive medicine. We believe that in order to have ML models' performances approved by healthcare experts, there is a need to reshape their contribution to the data mining processes. And in order to meet the healthcare research objectives, we will incorporate their expertise appropriately into a new methodology. They will be involved in the case studies conceptualisation, data acquisition, preparation, features selection, features engineering, modelling and performance evaluation.

Therefore, our research integrates both machine learning and healthcare experts together, leveraging their expert knowledge and sound judgement to develop new models that effectively address new inexpensive potential solutions to real-world preventive healthcare challenges.

1.1 Thesis Statement

This thesis uses machine learning (ML) approaches to develop a data-driven methodology to produce new models for predicting adverse health events in preventive medicine. With the hope that the new models would gain approval from the domain experts as an inexpensive step to a solution. Thus, this thesis investigates the following statement:

> “*A reliable data-driven framework can be formed to build tools for preventive medicine to predict adverse health events, approved by experts in true interdisciplinary case studies.*”

1.2 Research aim

The aim of this thesis is to develop new reliable ML models for preventive medicine by formulating a data-driven methodology. The new models are to be assessed by healthcare experts as potential new tools to detect patients' susceptibility to adverse health events. The new models may be used to predict potential deterioration in individuals’ health and Quality of Life (QoL).

1.3 Primary research objectives

A. Investigate and assess challenges, scientific debates and considerations of applying various data mining methods and ML algorithms to integrate them into a new data-driven methodology.

B. Formulate a new methodology from the selected methods for Machine Learning predictive modelling. The methodology is to be assessed by attempting to solve preventive medicine case studies systematically. This should show the combinations of chosen methods achieving a desired data-driven predictive outcome. The logic behind these methods and choices is to be made clear and justified. The applied methodology/framework must demonstrate the structural workflow of using the collection of methods to model the new preventive healthcare problems.

C. Apply the newly formulated methodology/framework to new case studies in the field of preventive medicine by building new predictive models to predict susceptibility to adverse health events.

D. Assess the new models and performances as potential new tools for early screening alongside healthcare domain experts.

1.4 Secondary research objectives

To investigate issues affecting the performance of the new predictive models. From cited research and recommendations, try to create novel methods to tackle gaps in the field and enhance understanding. Such novel contribution can be in any area within the scope of this thesis statement, including data preparation, modelling, evaluation, etc.

Finally, while reading this thesis, the reader will observe various recommendations and contributions throughout this research journey. Any novel methods within the scope of this thesis are applied as proof of concept where an opportunity arises. We hope the readers see our humble contributions as new opportunities for other researchers to take our findings and progress them further.

Chapter 2

Data Preparation Literature Review

This chapter covers and reviews

- *The data-driven modelling flow*
- *Modelling conceptualisation*
- *Data quality assessment*
- *Scales of measurements*
- *Exploratory Data Visualisation*
- *Data preparation and feature engineering*

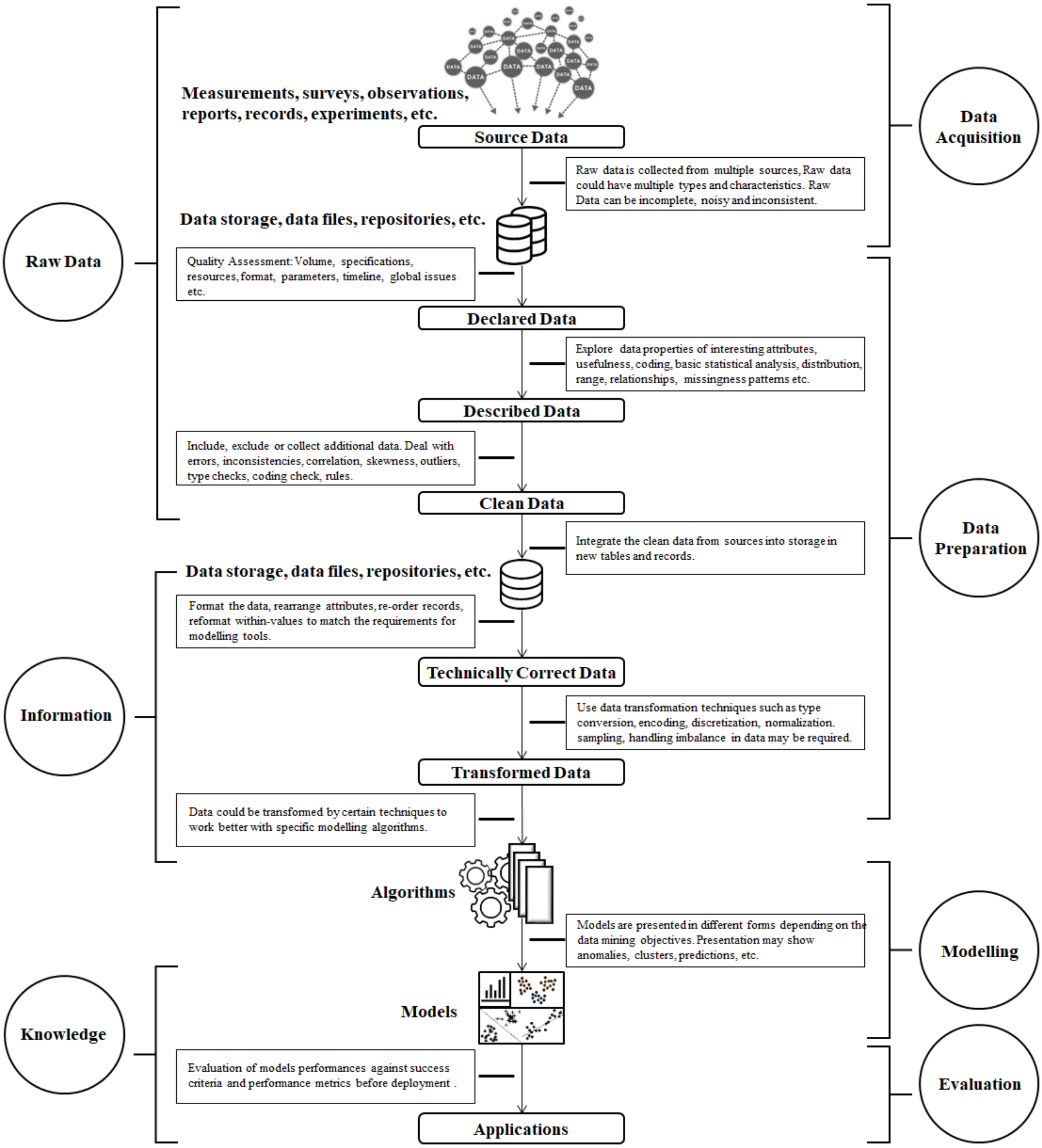
Data mining is a process of preparing and modelling data to extract knowledge, propose conclusions, and enhance decision-making in real-world applications. Most data mining methods focus on the modelling phase of the data mining goal and assume that the collected data is correct and suitable for analysis. However, in the real world, there is a massive inheritance to date of methods and algorithms that were developed starting from the 1950s. Each algorithm's use is bound to the data mining problem at hand. Predictive modelling, i.e., supervised learning, is a function that maps an input to an output based on example input-output pairs in a cohort of patients based on a feature vector of recorded responses, measurements or readings that belong to each patient. Models' evaluation is the final phase before deployment into real-world applications. Evaluating the models comes with many metrics; each has an interpretation of the model's performance. Using any of these metrics relies upon their potential to satisfy the domain experts' success criteria within the data mining project to produce the desired outcomes. In this chapter, we review various literature available to briefly explain and assess the use of various common data exploration and preparation methods. We also review the effectiveness of their application under various settings in various studies. This chapter also identifies gaps in the literature regarding the use of such methods, starting from problem conceptualisation, data quality assessments (DQA) at source, and possible methods to declare, visualise, clean and engineer the data.

Following the recommendation presented in this chapter can lead to important decisions that impact the formulation of this thesis methodology and its execution on predicting adverse events in preventive medicine case studies.

2.1 The data-driven modelling flow

The sequence of the data mining process can be illustrated in Figure 2.1. From Figure 2.1, the data mining process goes through different phases: collecting readings and measurements, storing them adequately in data repositories (sources), a sequence of investigations and explorations followed by data modifications, and passing through algorithms generating data models deployed in applications that produce insights.

Through the data mining flow, we recognise a challenge presented in the vast inheritance of algorithms [11], making it harder to choose among them; however, defining the modelling objectives and the data type for the task at hand narrows down such a choice. It is also worth mentioning that no ultimate algorithm performs the best in all data mining tasks. One must remember that developing data-driven models is not about approving or disapproving an algorithm for a specific job. Instead, it is about the adjustments that led a model closer to meeting domain experts' success criteria.



Raw data is collected from multiple sources, Raw data could have multiple types and characteristics. Raw Data can be incomplete, noisy and inconsistent.

Quality Assessment: Volume, specifications, resources, format, parameters, timeline, global issues etc.

Explore data properties of interesting attributes, usefulness, coding, basic statistical analysis, distribution, range, relationships, missingness patterns etc.

Include, exclude or collect additional data. Deal with errors, inconsistencies, correlation, skewness, outliers, type checks, coding check, rules.

Integrate the clean data from sources into storage in new tables and records.

Format the data, rearrange attributes, re-order records, reformat within-values to match the requirements for modelling tools.

Use data transformation techniques such as type conversion, encoding, discretization, normalization. sampling, handling imbalance in data may be required.

Data could be transformed by certain techniques to work better with specific modelling algorithms.

Models are presented in different forms depending on the data mining objectives. Presentation may show anomalies, clusters, predictions, etc.

Evaluation of models performances against success criteria and performance metrics before deployment .

Fig. 2.1 Sequence of data mining tasks and phases

2.2 Modelling conceptualisation

The initial step for new data mining projects is constructing a clear concept of the targeted outcome. Defining the initial objectives leads to determining the type of modelling task. Data modelling tasks fall within two areas descriptive vs predictive [12]. Descriptive models detect the relationships in the data and discover patterns of the data studied, for example, Clustering, Summarisation, Association rule, Sequence discovery etc. Predictive modelling has been defined in many different ways [12]. One study [13] by Delen & Demirkan describes it as having the data as a prerequisite when making authoritative predictions via forecasting and simulation to address the occurring event and its causes. Another study by Lechevalier, Narayanan, & Rachuri [14] describes predictive modelling as using statistical, machine learning, and data mining to discover facts to predict unknown future events when investigating a specific domain problem. A predictive model anticipates unidentified data values using the identified values, for example, Classification, Regression, Time series analysis etc.

Based on the aims and objectives to create models to predict patients' development developing adverse health events from patients' health records, the scope meant that this thesis adopts a data-driven predictive modelling approach.

The Health and Safety Executive regulatory body (HSE) in the UK defines the Risk assessment as a method of identifying the cause of injury or illness, deciding how likely it is that someone could be harmed and how seriously in order to take action to eliminate the cause, or if this isn't possible, control the risk [15]. Research carried out by Perkin and Balbus stated that efforts to protect persons at elevated risk, however, are hindered by varying interpretations of the term "susceptible" [16]. They show numerous variations in definitions of susceptibility. The definitions not only tend to be different in scientific and policy settings but also vary between and within disciplines and professions. Concepts of susceptibility focus on the ability of an individual to resist harm, the probability that an individual will react to a specific exposure, the comparison of an individual's susceptibility to that of the majority of the population, or the variation of individual states of vulnerability within a population. Definitions or descriptions of susceptibility sometimes embody more than one of these concepts, with the dominant one affecting how susceptibility is addressed in risk assessment and risk management processes [16].

Over time, individuals in the same discipline share interpretations of terms and develop a commonly accepted definition that may never be formally recognised. For our modelling conceptualisation, this thesis does not describe the varying definitions of susceptibility within various domains except those within the scope of patients' health. Therefore, we focus on the general, medical, biological, and epidemiological definitions from Perkin and Balbus' research [16].

In English, the term "susceptible" is once defined as "accessible or especially liable or subject to some influence". This definition shows that "susceptible" describes an inherent characteristic of some entity, which may or may not be a living being, and states that an external factor interacts with that entity. Beyond this relational concept, however, this definition remains broad.

Moreover, among all other definitions in various dictionaries, there is a consensus that the term implies that an external factor changes the entity somehow. The Oxford and Webster definitions further describe that the external factor influence may be neutral, beneficial or adverse events such as disease or injury. Webster's dictionary adds that the extent of influence can be acute, chronic, or delayed effects combined with the dependence on a portal of entry, dose, and virulence or toxicity of the agent, the natural and acquired resistance of the host, and lifestyle. In the medical and biological domain, susceptibility has a variety of descriptions as the state of being open to disease or infection, the state of being readily affected or acted upon; diminished immunity to disease, especially to an infection; the likelihood of the individual to develop ill effects from an external agent [16], the latter is found in Stedman's Medical Dictionary. Further definitions were also presented in more depth; susceptibility is the net influence of a large set of variables. Another goes into more detail, stating that susceptibility is the enhanced responsiveness regardless of cause, and susceptible individuals are those who respond to toxic or carcinogenic substances at doses significantly lower than that to which the general population responds. The previous definition naturally led to a new sub-term of susceptibility known as “Hyper-Susceptibility”; given a high enough dose, everyone is susceptible to a toxic reaction to an environmental agent, and those who react abnormally to the greatest extreme are “hypersusceptible.” In Epidemiology, Susceptibility is defined as the relative propensity of an individual or population to develop dysfunction or disease or an individual’s intrinsic or acquired traits that modify the risk of illness.

Perkin and Balbus' paper [16] surveys the definition of susceptibility in depth across domains and in relation to other terms, including sensitivity. They found little agreement about the appropriate terminology for the physiological state, noting that susceptibility, hyper-susceptibility, high-risk, sensitivity, and hyper-sensitivity are used almost interchangeably [16]. They recommended that the term susceptibility requires an urgent formal or even a legal definition for risk assessments as there is no consensus on the definition. In addition, they suggested when susceptibility is addressed to develop risk assessment methods or conduct a risk assessment for a specific agent, one of these definitions should be cited, or a consensus definition should be set to ensure that the interdisciplinary audience will clearly understand the scope of susceptibility intended.

Their proposed definition to be used in interdisciplinary research is: “Susceptibility is a capacity characterisable by intrinsic and extrinsic factors that modify the impact of a specific exposure upon risks/severity of outcomes in an individual or population.” [16]

2.3 Data Quality Assessment

Often, once the data mining objectives are perceived, data is required. Data can be collected from measurements, surveys, observations, reports, records, experiments, devices, images etc. Data can also be obtained from repositories and data showcases. Some sources are unrestricted (open access) [17], while access fees and legal agreements restrict others [18] [19]. Finding a source for large sets of health data representing a large population is proven challenging in many aspects; some are related to privacy and access controls, demographics, population characteristics, cost and expertise in collecting as accurate data as possible. Besides issues like missing observations, noise and errors in datasets, other problems may arise; in particular, clinical and medical datasets often come with many technical challenges, such as the curse of dimensionality, bias, and the inherent limitations of observation study and the inability to test causality resulting from residual confounding and reverse causation [20][21].

Also, never undermine data quality issues; the evaluation of data missingness, descriptive statistics and format do not fully provide a complete picture of the dataset's quality as described by the data source. The manual review of quality is a resource-intensive task, and there are no evidence-based or community-driven metrics for assessing research data quality [22]. There are many frameworks for Datasets Quality Assessment (DQA) [23], but they are objectives-centric and context-dependent. This issue can be problematic since clinical research datasets are often phenotype-based, require a unique patient cohort, or is specific to a set of participating medical centres. In addition, research investigators frequently develop ad-hoc metrics specific to their study, hence cannot be replicated [21].

2.4 Scales of Measurements

Medical research modelling studies [24][25] are often criticised for not declaring how data types are handled when used to build their statistical models. Many medical research statistical models [26-31] treat data as either categorical or numeric in encoding, imputation and feature engineering.

In 1946, Stanley Smith Stevens was the first to raise the issue of measurement scales in statistics, as he created a taxonomy of measurement [32]. Stevens stated that measures are defined as the assignment of numerals to objects or events according to rules. Such assignments' rules lead to different kinds of scales and other types of measurements. This statement signified the importance of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale [32]. As a result, Stevens created four data measurement scales: Ratio, Interval, Ordinal and nominal. Over time, other researchers criticised Stevens' classification of scales, adding more scales to the topology of measurements [33].

The machine learning domain inherited parts of Steven's topology and introduced multiple measurement scales in various interpretations. Some statisticians [34] merged ratio and interval types into quantitative and referred to ordinal measurements as ordered categorical values. The categorical variables are called qualitative or discrete [34]. Others combine ratio and interval in a numerical pot and use a topology of three scales; numerical, ordinal and categorical [35]. Some online-active data scientists classified the measurements as continuous, ordinal and nominal [36], while others scientists went for discrete vs continuous values. Therefore, interpreting data dictionaries from various data sources may seem confusing and inconsistent. Different domains adopt one topology over another. Thus, the absence of a consistent framework for thinking and communicating data types [37] to domain experts can hinder interdisciplinary ML studies. The problem worsens in data transformation and may slow down data cleaning. However, for a start, we can safely say that machine learning data scales measurement types are ultimately split between categorical and numerical. Besides, other data types, such as binary and date-time, should not be left out. In all cases, we believe that converting features to the correct data types results in conceptualising the data mining problem, enhancing data preparation, selecting applicable modelling ML algorithms, making the modelling process faster and likely improving models' performances and interpretability.

We emphasise the importance of having a uniform strategy for classifying data types within a data science project that closely conforms to the data source description and appeals to domain experts' understanding. Thus, we present six measurement scale types with a strategy to account for their processing in the modelling phase: Nominal, Binary, Ordinal, Count, Time and Interval.

The nominal type of categorical value is discrete and holds no arithmetic significance. Their groups do not have a specific order; hence the man and median are meaningless to describe them. Ordinal values are also discrete but can be ranked or sorted; however, the mathematical significance of their quantification is not known, i.e., the distance between any two measurements is not defined. If the ordinal is assumed to have equal magnitudes, it can be encoded as discrete numeric intervals similar to Likert scales [38]. Another option is to consider ordinal responses as nominal data where each category has no mathematical relationship with the other [36]. A common type of measurement scale is Binary data; it is a special categorical case of one out of two responses, indicating a count, an interval or a nominal label. This measurement type can be accounted for as of own kind in machine learning modelling [36]. On the other hand, Counts are discrete numeric natural numbers similar to interval values; they hold mathematical significance and follow a Poisson distribution [39]. Exact measurements are known as Interval data. Such reading precision does not matter when modelled with machine learning; they may hold discrete or zero values on a continuous scale of measurements. Finally, Time data is a repetitive type of continuous data that can describe any period. Such data is relevant to Time Series models with a time component that is very crucial to the model. There are various ways of engineering time-date data types into features.

There is no uniform topology in statistics that can be used across all data mining projects to assign value types. Besides Steven's, in 1977, Mosteller and Tukey proposed a new seven-level typology [40], and in 1998, Nicholas R. Chrisman expanded the list of levels of measurement to ten [41].

In practice, addressing data measurement types in data-driven modelling depends on the capabilities of the tool of choice. Various modelling tools have built-in perspective strategies to handle specific data measurement types before consuming a variable by the modelling algorithm. Stevens's typology also raised the debate of having a meaningful statistical description of attributes characteristics, i.e., central tendency and dispersion, since it differs for each scale. Table 2.1 shows methods of describing the attributes based on their data types [42].

Table 2.1. Descriptive statistics per the classification of variable measurement

Type	**Classification**	**Variability**	**Descriptive Statistics**	
			Central tendency	*Dispersion*
Interval	*Numeric*	*Continuous*	*Mode, Median, Arithmetic mean*	*Range, Standard Deviation*
Ratio	*Numeric*	*Continuous*	*Geometric mean, Harmonic mean, Arithmetic mean, Mode and Median*	*Studentised range, coefficient of variation*
Time	*Numeric*	*Continuous*	*Index of Dispersion (Variance-to-mean ratio)*	
Count	*Numeric*	*Discrete*	*Index of Dispersion (Variance-to-mean ratio)*	
Ordinal	*Numeric|Nominal*	*Discrete*	*Median, Mode (if numeric)*	*Range (if numeric)*
Nominal	*Nominal*	*Discrete*	*Mode*	*Frequencies*
Binary	*Binary|Nominal*	*Discrete*	*None*	*Frequencies*

2.5 Exploratory Data Visualisation

Using data-driven insights to create actionable strategies and deploy valuable initiatives is essential. The graphical representations of information (charts) play a prominent role in data exploration. It assists in interpreting gathered data from real-world applications by exposing its structure and exploiting its complex sets of numerical figures.

From an analytics point of view, the use of charts and their complexity vary depending on the dimensionality of the data and the type of analysis performed, univariate, bivariate or multivariate. There are many examples of conventional graphical representation, including bar charts, column charts, line graphs, area graphs, surface charts, scatter plots, pie charts and spider (Radar) plots. The higher the number of variables, the higher the complexity of analysing all possible relationships among the variables on a single two-dimensional. One approach to simplify the complexity of visualising multiple dimensions is to reduce the dimensionality of the dataset to eigenvectors that preserve the portion of pattern within the data most, known as Principal Component Analysis (PCA), developed by Karl Pearson in 1901[43]. PCA compresses the data dimensionality by projecting each data point only onto the first few PCA vectors (Principal Components). Some of the benefits of PCA are the 2D compressed projection of the data, the reduction of modelling time and the removal of correlated features in large dimensional data, which improves the modelling performance for algorithms impacted by multi-collinearity.

However, the disadvantages may outweigh the benefits. Through the transformation to the PCA components plane, the original visual representation of the collected variables disappears, resulting in low visual interpretability and increasing numeric interpretability, which in turn prevents domain experts from observing the association and contribution of each of the originally collected attributes individually [44]. This transformation could make it harder to draw conclusions or deduce further hypotheses for further studies on specific attributes in the collected data. Also, PCA makes additional assumptions that may be considered disadvantageous. PCA represents a transformation of data features that captures the features' linear relationships. At the same time, this is not the general case for all features; further transformation of non-linear features into linear may be required, such as log transform. Keeping in mind that PCA is pruned to favour features with larger magnitudes of scale, which means further transformations may be necessary, such as scaling. Also, PCA sensitivity to scale imposes outliers' removal or mitigation before analysis. Despite adding Kernel-PCA extension to PCA to deal with the non-linearity problem of features, dimensionality reduction often results in loss of information. PCA components try to cover the maximum variance among the variables in the dataset. Still, if the number of Principal Components is not carefully selected, they may miss some information compared to the original set of predictors [44].

Unlike PCA compression of features into effectively new transformed features, Independent Component Analysis (ICA) is an alternative approach to visualise higher dimensional data [45]. ICA decomposes a multivariate feature into independent components (mixtures). Each predictor feature can be extracted from a set of feature mixtures by taking the inner product of a weight vector and those feature's mixtures where this inner product provides an orthogonal projection of the feature mixtures, then finding such a weight vector. One method for doing so is known as the Projection Pursuit (PP), proposed by Kruskal in 1969 [46]. PP locates the projection from high- to low-dimensional space, revealing the data set's most detailed structure while reserving the original features' presence. Once a projection of interest out of a set of projections is found, existing structures (such as clusters) can be extracted and analysed separately. Friedman and Tukey first implemented PP in 1974 with automatic machine-chosen interesting low-dimensional projections of a high-dimensional point cloud [47]. Therefore, the major advantage offered by the PP method implementation is that it bypasses the "curse of dimensionality". It was extended with additional practical features supplementing the automatic machine-selected (blind) view of most interest. The new features were proposed in 2007 by Joe Faith, known as Targeted Projection Pursuit (TPP) [48] and were later implemented in 2011. TPP implementation is user interactive. It allows exploring complex datasets with up to 100s of features to find patterns (projections) of interest. TPP exploits PP's ability to bypass the "curse of dimensionality" by recognising that most of the feature space is empty. Thus, TPP allows for exploring the space of projections by manipulating individual data points or groups of points directly in a multidimensional scatter plot. Once the user chooses a projection, existing structures (such as clusters) can be extracted and analysed separately [48]. Historically, one drawback of TPP methods is the high demand for computational resources, and nowadays, this is no longer a barrier since it can be performed on a personal device. Figure 2.2 illustrates the difference between data visualisation of non-linear data with PCA's two principal components and the TPP of all features.

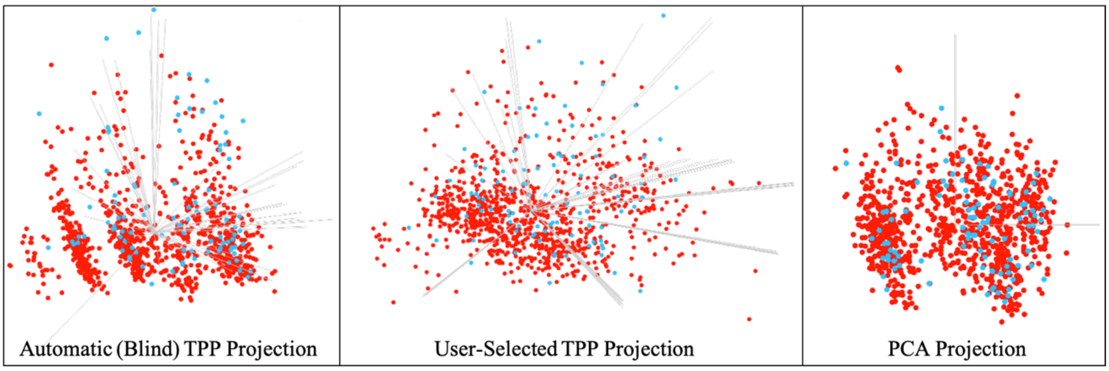


Fig. 2.2 Illustration of TPP advantage over PCA projections of a non-linear dataset

Figure 2.2 illustrates the automatic (blind) TPP projection of the binary classes among 122 predictors, shows a better (more interesting) view of the data, almost showing five different clusters within the dataset with observed high inter-cluster dissimilarity and intra-cluster similarity. In contrast, the user-selected visualisation shows two clusters of interest, one of which seems to have a low intra-cluster similarity. However, the earlier projection offers a more detailed insight. Here, the conventional (also automatic) PCA appears to show 2~3 clusters of patterns. The additional benefits of visualising the blind TPP projections are demonstrated by preserving the original features when PCA seems to have considerably suppressed the number of inconsistent data points (i.e., outliers and extreme values).

TPP may not be the perfect answer to a multi-dimensional visualisation of data points. Yet, it seems the simplest in users' eyes to interpret and take control of and supports the delivery of the objectives of our projects. We live in a three-dimensional world, but everything we see is first recorded on our retinas in only two dimensions. Therefore, visualising data points in higher dimensions has always been a challenging topic in research due to our sight limitations to seeing rendered images in 2D views.

Data scientists in the industry addressed this problem in many different creative ways, from visualising the data in a matrix of pairs, using parallel coordinates, kernel density plots, and adding depth, hue, size, and shape to the plots. Adding more dimensions makes it harder to go around the limitation of the two-dimensional rendering device to visualise our data.

2.6 Data preparation and feature engineering

Raw data often come with multiple problems, such as value errors, missingness, incorrect coding, incorrect formatting, etc. Data preparation aims to improve the data content to provide better-extracted knowledge from the modelling process. Therefore, such data requires selective extraction, organisation and formatting and once transformed, the data may show indications, raise questions or provide answers. Specific tasks, like data imputation and outlier detection, are conventional data preparation methods that could impact the modelling outputs, often known as data cleaning techniques. In some frameworks [49] data preparation also includes pre-processing that transforms the data so that machine learning algorithms can consume (algorithm-understandable format); it also consists of any enrichment, reduction, enhancement and feature engineering.

We consider all the steps between data collection and modelling, including any statistical analysis and data transformation, to be data preparation steps because tasks between the two phases deliver the transition from raw data to information [49] [50].

Since data preparation directly impacts the models' success rate, it is vital to be performed in a reproducible flow in the data mining environment [51]. Therefore, any data pre-processing should adopt the following sequence of steps: (1) perform data pre-processing on the training dataset; (2) record the statistical parameters required for pre-processing the training dataset; (3) Perform the same data pre-processing methods on the test dataset in isolation by using the same statistical parameters learnt from the pre-processing of the training dataset [51]. These recommendations are also based on empirical studies by Weiss & Provost (2001) behind a machine learning predictive model (classifier) performance being affected by the class distribution of the training data [52].

Historically, some modelling algorithms, including various Machine Learning techniques, showed preferences among applied methods to transform data [53]. Depending on the data type, strategies could be used, such as conversion, discretisation, normalisation, sampling, etc. Additional transformation techniques are proven beneficial when handling class imbalance in datasets [53]. Some preparation steps are known to improve the model's performance. Previously, various methods have been applied to vary each of these pre-processing steps. However, there are no standard systematic procedures to compare these methods' impact on the modelling performance [54] [55]. Furthermore, data pre-processing (Figure 2.3) is grouped into two types; supervised and unsupervised data pre-processing;

each type can be performed at two levels; instance (individual records) and attribute (variables) levels [56].

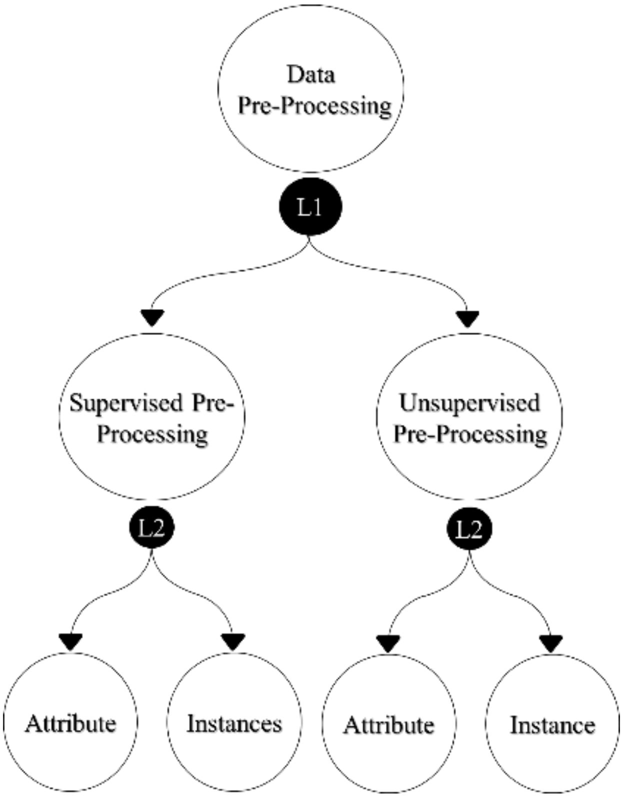


Fig. 2.3 Data pre-processing types

Data pre-processing is considered unsupervised if the response variable does not govern the data transformation per attribute or instance. Therefore, any conversion to the values held within an attribute or a record is performed independently from the class. For attributes, removing useless features, converting data type, unsupervised discretisation methods, treating missing observations, outlier detection, feature scaling, binarisation, and others are examples of unsupervised data pre-processing techniques [57]. For instances, randomising records, removing duplicate records, resampling records, removing outliers, and others are considered unsupervised pre-processing tasks [57]. However, if the response variable governs pre-processing, the latter becomes supervised because any transformation to the values held within the attribute or a record depends on the class [58]. For attributes, supervised discretisation, feature selection, and others are examples of supervised data pre-processing techniques. Resampling techniques addressing imbalanced learning are examples of supervised pre-processing steps applied at the instances level [58]. There are a wide variety of data pre-processing methods; however, we only can review a handful of them.

2.6.1 *Records randomisation*

Shuffling dataset records is advised before performing pre-processing tasks, such as splitting the dataset into training test subsets. This technique randomly reorders the instances with a random number generator that is set with a new seed value whenever it processes a new set of cases. Its effect is known to prevent the records selection bias and provides assurances against any accidental bias. As a result, comparable groups are produced with eliminated bias if any was present at the source or in data collection phases [59].

2.6.2 *Useless features elimination*

This technique removes variables that do not vary or vary too much. All constant nominal and numeric attributes are considered useless and eliminated before any analysis based on their distinct recorded readings, including nominal variables that exceed the maximum variance percentage setting [60]. Nominal variables that do not vary or vary too much are considered useless data. Nominal data that varies too much, storing almost a category per record, could map a one-to-one relationship to the output variable, compromising the learning process. Here we note extreme cases rather than having a large number of categories. All constant attributes that do not vary are useless because they hold no learnable differences to distinguish the endpoint. The maximum variance percentage for a variable(χ) is calculated by:

$$\boldsymbol{Maximum\ variance\ percentage} = \frac{\boldsymbol{number\ of\ distinct\ values}}{\boldsymbol{Total\ number\ of values}} \times \mathbf{100}$$

If the attribute's maximum percentage of variance is greater than allowed, i.e., 99%, it is considered useless [60].

2.6.3 Feature binarisation

This task constructs new features by transforming a raw data attribute of any entity into vectors of binary numbers. Even the most complex concepts can be transformed into the binary form [61]. In the usual label encoding, potentially when categorising records, the categorical values also proportionally increase as the number of distinct records increases. Standard categorical encoding (labelling) assumes that the larger the entries of a categorical value, the better the category; this assumption causes bias towards certain groups in categories with larger records in variables. One hot encoding resolves such a problem [61]. It is a binarisation process by which categorical variables are converted into a form that could be provided to ML algorithms to improve prediction performance [62]. It converts a nominal variable into multiple attributes. This method is also used for variables that have a large number of categories. The categorical value (label) represents a group of entries in the dataset; the values start from 0 and go up to N-1 categories. One-Hot encoding binarises each category, converting every categorical value into a new feature.

2.6.4 Missing data investigation

Missing data is a popular area that should be handled with care in crucial decision-making models. The recommended strategies for missingness remedies vary depending on the missing data pattern, missingness levels, missingness type, the domain where it originated from, the problem we want to model, etc [63]. In the real world, raw data collection usually comes with multiple cavities in the data; this is also the case for this research's clinical and healthcare data. There are two types of caveats; one is missing a set of variables, and the other is missing observations' values for present variables. In most cases, data missingness refers to missing values for one or more variables per record which will be discussed in this subsection. There are many reasons for missing data; some are natural, like when participants opted out of a study, preferred not to disclose some information, the subject's death, etc. However, it also can be systematic due to collection procedures and design, improper collection (instability), data management and devices issues, etc. (An area where data repositories, such as the UK Biobank, are very explicit about [64]).

In that sense, missing data can be unknown, unrecorded, undisclosed or irrelevant, etc. There may be a good reason why the attribute's value is unknown. Perhaps it is based on a decision made, on the evidence available, not to collect a particular measurement; that might imply some information about the record other than the fact that the value is simply missing [63]. A scenario emerges when doctors and clinicians analyse clinical datasets provided for healthcare study; in some cases, a diagnosis could be identified from a range of tests, procedures or treatments a patient underwent, regardless of the test's output. In such a scenario, the missing outcome of tests can be ignored.

The previous scenario emphasises that only medical experts from the specialist domain who are familiar with the data can make an informed judgement about whether a missing observation has some significance or should be considered an ordinary missing observation. That scenario raises the importance of a systematic review of missing readings from medical datasets from four different angles before embarking on a missing value handling strategy [65].

Feeding variables directly to machine learning algorithms and relying on their general embedded methods to handle missingness in medical datasets may not be entirely appropriate. This is because machine learning schemes run under the assumption that there is no particular significance to the fact that a record or a group of records contain missing observations in a variable [65]. Therefore, when observing domain experts' activities in healthcare research and before embarking on a strategy to handle missingness, four areas should be investigated to assess the meaning of missing values. These areas are the *causal missingness of observations*, the *type of non-response*, the *missingness pattern* and the *proportion of missing data* [65].

A) Causal missingness distinguishes between random versus selective loss of data. When reviewing the literature on this particular issue, we find that in 2001, Paul D. Alison divided the cause of missingness into three types [66] depending on the potential cause of data missingness: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). MCAR missing values are unrelated to a process causing this missingness; hence it is considered random. Assume a dataset with two variables, a dependent variable $\boldsymbol{Y}$ with missing response $\boldsymbol{R}$ and a dependant variable $\boldsymbol{X}$. The probability of missing value $\boldsymbol{R}$ does not depend on anything. $\boldsymbol{X}^{\mathrm{obs}}$ and $\boldsymbol{X}^{\mathrm{mis}}$ are any observed and missing values for $\boldsymbol{X}$, respectively. $\boldsymbol{\psi}$ represents any other value. MCAR condition is given by [66]:

$$\boldsymbol{P(R|Y,X) = P\left(R\middle|Y,X^{obs},X^{miss}\right) = P(R|\psi)}$$

MAR missing observations may be due to a relationship between the missing data and recorded observations for other variables; however, one cannot be sure due to a cavity in the declared data or missing a related variable. This cause of missingness can be defined by [66]:

$$\boldsymbol{P(R|Y,X) = P\left(R\middle|Y,X^{obs},\psi\right)}$$

MNAR is the most severe type of missingness. It may be due to a relationship between the missing data and observed data or missing values for other present or missing variables. The missingness is expressed [66] in relation to an unknown probability P(A):

$$\boldsymbol{P(R|Y,X) = P(A)}$$

Since the missing values are unknown, one cannot be entirely sure about the probability of missingness. Besides, many missing data exploration methods assume that the missingness cause is MCAR or MAR [67]. For example, techniques like the T-test may establish an MCAR type [68], but it is not totally accurate. Heckman-type selection models explore MNAR data [69] [70].

B) Missingness Non-response is a common research survey issue of missingness that investigates the subject's (unit) preferences in answering questions in a survey. We find this topic is widely also spread in the literature. For example, in 2010, Yan and Curtin described the types of non-response in a survey [71].

They describe two types of non-response missingness: unit non-response and item non-response [71]. Unit non-response refers to the complete missingness of a record's responses. Where Item non-response is selective missingness at a record level for a subject that relates to the absence of answers to specific questions, an example is a subject who answered most of a questionnaire but decided to skip a few unanswered. Historically, for such a case, different statistical treatments, adjustments and causes were described [72][73]. Unit non-response has a larger spread in a dataset than item non-response. There are many techniques to address unit non-response, such as reweighting methods proposed by Pérez-Duarte et al. [74] and weighting estimates by the Horvitz-Thomson Estimator [75]. Item non-response severely threatens data quality since it reduces dataset size if only completed cases are used in any modelling. Imputation techniques can be used to substitute for missing values to avoid the shrinkage of sample size. However, it is still problematic if an item non-response represents non-ignorable missing data. Non-ignorable missing data occur when the missing data pattern is correlated with the values of the dependent variable (output variable) [76]. Various imputation and modelling methods are developed to compensate for item non-response [76].

C) Missingness patterns can assist scientists in identifying the type of missingness in the dataset, leading to more suitable strategies for handling missingness. Van Buuren's book "Flexible Imputation of Missing Data" describes three key patterns of missingness in a dataset. There are multiple patterns of missingness patterns; Monotone vs Non-monotone (Arbitrary) missing data patterns, Univariate vs Multivariate and Connected vs Unconnected [77].

A monotone missingness pattern is systematic in the data. For example, assume having a dataset with variables $\boldsymbol{X_1}$,$\boldsymbol{X_2}$,$\boldsymbol{X_3}$, ..., $\boldsymbol{X_p}$; To have a monotone pattern is specific to the event for a variable value $\boldsymbol{X_j}$ is missing that implies all subsequent variables values are missing $\boldsymbol{X_k}$ are missing where $\boldsymbol{k>j}$. Alternatively, when $\boldsymbol{X_j}$ is observed for a subject, it is assumed that all previous $\boldsymbol{X_k}$ for $\boldsymbol{k<j}$ are also observed for the same subject. Monotone missingness occurs in longitudinal studies with drop-out. If the pattern is not monotone, it is called *non-monotone* or Arbitrary. A non-monotone pattern indicates that missingness is random [77].

A missing data pattern is said to be univariate if there is only one variable with missing data. Missingness that occurs everywhere is knowns as multivariate missing data. A missing data pattern is considered connected if any absent data point can be reached from any other present data point through a sequence of horizontal or vertical moves in the dataset and is considered unconnected otherwise [77].

Generally, one cannot be sure whether data are missing at random or whether the missingness is related to unobserved predictors or the missing data themselves. The fundamental difficulty is that these potential "lurking variables" are unobserved-by definition, so one can never rule them out. We generally must make assumptions or check with reference to other studies. Another practical approach is to try to include as many predictors as possible in a model so that the "missing at random" assumption is reasonable. For example, a strong assumption may be that non-response to the earnings question depends only on sex, race, and education. However, this is much more plausible than assuming that the probability of non-response is constant or depends only on one of these predictors [78].

D) Missingness proportions for each variable and each record must be computed to help decide which variables or records should be considered candidates for removal or data imputation [79].

2.6.5 Missing data mitigations

Missing observations treatment depend not only on the reason for missingness or possible pattern of missing values but also on the spread of missingness, the dataset size, the volume of affected records and the required analyses on the dataset. One must remember that the primary goal of missingness remedies is not guessing a missing value itself; instead, it is for obtaining reasonable parameter estimates about the relationship between variables to provide the best guess. In practice, we try to include as many predictors as possible in a model to make the "missing at random" assumption reasonable [78]. Missing data treatments fall within three main groups: deletion, single imputation and model-based methods [79]. Different imputation techniques introduce different values to complete the data; hence are expected to perform differently on various datasets. The level of performance may be assessed on how close the completed values are to the known hidden (intentionally missing) values; the technique that best recovers the actual data "wins"; this approach can only be applied in simulation and not in the real world; this method was proposed in 1975 as proposed by Gleason and Staelin [80]. Another recent approach compares selected metrics such as the Sum of Square Error (SSE) to differentiate imputation techniques performance (excluding deletion) over incremented proportions of missing data in a sample of the original data [79]. Whichever the case, one must never forget that the missing data requiring imputation is unknown in the real world. Thus, we cannot quantify the quality of a missingness treatment based on how well it can recreate the actual data.

Imputation is not a prediction [77]; in simulation, imputing with predicted values may lead to realistic imputations if the predicted values are close to the precise true hidden values; the method may have the ability to reconstruct the missing values from the available data. But this will only indicate that no information was missing in the first place; it was only concealed in a different form of coding or within another value in a variable, i.e., a derived variable [77].

In recent years, with the development of machine learning, many algorithms embedded models-based, deletion (by ignoring) or single imputation methods of handling missing data within; Early ML algorithms such as ID3 (Iterative Dichotomiser 3) decision tree by Quinlan (1979) simply ignores the missing numeric values or treats nominal missing values as a new category [81]. In Gaussian processes, Gharamani and Jordan (1994) published a method to handle missing values as hidden variables and impute them with a variant of the Expectation Maximisation (EM) algorithm [82]. For Kernel methods, Smola and Hofmann (2005) described how to handle missing observations with a variant of Support Vector Machines (SVM) [83]. Chen and Guestrin, in 2016, described Extreme Gradient Boosting (XGB) trees that come with extensions to handle missing observations, so they can process the data without imputing missing values via a sparsity awareness extension [84].

Finally, we believe that Machine Learning algorithms that treat the missing values by ignoring them or accounting for them as a separate group, in a sense, are not handling missing values. To handle missing values correctly, one must ensure that the estimated data points are not excessively utilised in the predictive model (inducing bias) to produce predictions. Also, one should ensure that most of the model learning, decisions and parameter tuning are made on observed data points. For example, we should avoid a scenario where an estimated value is used in evaluating a node split in a decision tree; this is another demonstration of machine learning algorithms misestimating the importance of a missing value in a record. And here is a brief background review of various methods of missing data treatment [66]:

A) Deletion is one way to eliminate the problem of missing data. Objects are retained as per completeness in target data for the specific analysis and, ultimately, modelling.

Deletion has three types: Case-wise deletion, Pair-wise deletion and Variable dropping**.** Case-wise deletion**,** also knowns as complete case analysis, deletes any record (case) with a minimum of one missing observation and retains those without any missing values across all variables. Pair-wise deletion*,* also known as available case analysis, depending on the case study, records are retained as per completeness in target variables for the specific analysis; other records with missingness will be discarded. This deletion method typically results in bespoke instances kept in a dataset for each different analysis. Variable dropping discards a whole variable for all records in the dataset. In some data sources, i.e., the UK Biobank [19], raw data has a property of stability; In some cases, the number of records could be small and growing where the data stability for some variables is referred to as Accruing, ongoing or updatable. If the portion of available observations in such variables is small compared to the remaining variables, the complete variable is removed from the analysis. In healthcare research, case-wise deletion may be used as a remedy for missing data only if the proportions of missing data (all observations) are below 5% (a rule of thumb). Also, case-wise and pair-wise deletion should not have certain patients' groups, i.e., those who were identified as very vulnerable to a disease or those identified as immune to a disease specifically if missingness was suspected to be lost to follow-up or opt-out of a trial [85].

The 5% threshold also assumes that the missing data is MCAR or MAR (Little's test can also be performed) [86]. Therefore, if the missing observations are MCAR/MAR (hard to prove), the case-wise deletion shall suffice, and the modelling on the observed remaining data will not be biased

[87]; otherwise, the analysis will be at risk of biased results, and no statistical method can absolutely account of the potential bias [88]; Weighting-Case Analysis can be used to weight the complete-cases by modelling the missingness to reduce the potential bias introduced in the available-case [79]. In the medical domain, a clinical trial is said to be confirmative if missingness within any predictor variable is below 40% and treated with considered unbiased imputation technique. Variable dropping should be exercised on predictor variables whose missing values exceed the 40% threshold; otherwise, the analysis can only be used for hypothesis-generating results [89].

B) Simple imputation methods substitute missingness with a new (predicted) observation. Some of these methods are simple such as single value substitution, and others are more complex. Examples of single imputation include mean and median, hot-deck and cold-deck, last observations carried forward (LOCF), baseline observation carried forward (BOCF), indicator variable substitution, information-based imputation, logical rule imputation and random imputation. Mean, Median or Mode imputation replaces missing values by the mean (marginal average), the median (the midpoint) or the mode (the most repeated value) of a frequency distribution. It is thought that the median imputation is more robust when in the presence of data outliers; However, all single imputation techniques could potentially severely distort the distribution [79]. Although such methods are straightforward, they reduce the strength of associations between variables, pulling any present correlations towards zero due to the reduction of the data's variability (change). Furthermore, if such a replacement happened at a large enough scale, it would bias the data analysis toward that single replacement value [78] [88].

Hot deck imputation replaces missing observations with a randomly selected value from other present observations for the same attribute in the dataset. The missing observations are known as recipient units, and the values of the present observations providing the values for imputation are called donor units [90].

Cold deck imputation is similar to the hot deck imputation except that the pool of donor records is held in a different dataset used for the same purpose of analysis [90]. Both methods require grouping similar observations in a dataset, then selecting the donor units from the same imputation cell as the recipient units makes the imputation random with replacement in the case of hot deck or without in the case of cold deck [91]. This technique is used by the US Census Bureau [92].

LOCF, also known as the sample-and-hold approach, recommends replacing the missing values in the treatment outcome variable with observations from pre-treatment measures. In clinical trials and longitudinal studies, this method assumes that patients' health for those who were given treatment improved better compared with the health of those who didn't receive treatment, hence encouraging replacing missing values with pre-treatment measures. [79]

BOCF accounts for partial treatment or medical trial dropouts. So, it imputes values for subjects who had at least a single exposure to a drug or treatment and dropped out by not completing the entire course of treatment [79]. After receiving partial treatment, the expectations in observations for subjects see an improvement, deterioration, or no change in the post-baseline measurements for a subject. Subjects who do not have a value at the endpoint (baseline outcome) because they dropped out are assumed (imputed) to have had "no change". The assumptions run by LOCF and BOCF are often unrealistic and can lead to underestimating the actual effect of treatment [93].

Indicator variable substitution is used for nominal and numeric variables with missing observations. In the case of nominal variables, missing values are substituted with an additional dummy category. In continuous variables, an extra indicator is used to identify missing values, such as zero or the mean without a preference. This approach is widely used in social sciences. However, the absence of interaction between the indicator and values in other predictors could yield biased coefficient estimates in statistical models such as regression, where the slope is forced to remain the same across groups with missing observations [78].

Information-based imputation technique imputes values based on information stored in other variables [78]. For example, the annual salary is a common variable with high missing observations. It is logical to find the household occupier profession and substitute missing observations with the average salary for that profession from the population; these imputations may propagate logically valid measurement errors.

Logical-rule imputation does not rely on assumptions and requires the missing-data mechanism to be known. Suppose the data acquired from a data source is well-declared. In that case, it is worth reviewing the data collection rules and survey designs that potentially lead to a better understanding of the data transaction [78].

In simple random sampling imputation (SRSI), samples are randomly drawn from the dataset to impute the missing value. SRSI is reported to distort the distribution of the independent variables with high missingness [94].

C) Model-based imputation methods use a predictive model to estimate the missing observations. In this case, the dataset is split into two subsets for the variable under evaluation: one subset holds the present observations only used as a training set for the model, and the other contains all missing values requiring estimation [66]. Many modelling methods can be used, such as Regression, K-Nearest Neighbour (KNN), Maximum Likelihood Imputation (MLI), Expectation Maximization Imputation (EMI), Decision Tree-based Missing Value Imputation (DMI) and Multiple Imputation (MI).

In general, both prediction and imputation models provide an estimated value that is hidden or unknown. However, imputation cannot be regarded as a prediction. Imputation indicates that the missing values cannot be accessed, and its estimation cannot be validated. Unlike imputation, prediction modelling assumes access to the true unknown value can be either instantly, in the case of simulation (hidden) or in the future (awaiting to occur). Therefore model-based imputation estimation should not be assessed in a similar way to prediction models.

In Regression, the variables in the dataset are used to create a regression model using the available observations in the variable to impute as an output. There are various types of regression, such as linear, stochastic and logistic [95]. Linear regression uses all variables to build a regression model where the variables subject to imputation are considered an output. Stochastic regression imputation is a deterministic regression imputation with an added random error component [96]. The logistic regression technique can be used for imputing missing data. It produces a probability-based model [97]. The model estimates the probability of an observation occurring depending on the values of the other categorical or numerical independent variables. Regression model imputation, in general, benefits from accounting for the relationship between variables. Despite linear and logistic regression models' prediction of the most likely value of the missing observation, they do not supply uncertainty about that value which, by turn, overestimates the model's fit and correlations between variables. Stochastic regression is used for numeric and categorical missing values imputation and uses the same procedure as standard regression imputation. It introduces uncertainty by adding an extra step of augmenting each predicted value with a residual term. It aims to preserve the variability in the data; however, it tends to underestimate the standard error leading to increased false positive predictions [98].

KNN completes missing values with the mean of the k values coming from the k most similar complete observations [99]. The similarity of observations is determined using a distance function, i.e., Euclidean, Manhattan, etc. The present observations in the dataset must be normalised before imputation since this method is pruned to bias towards variables with larger magnitudes. KNN can reasonably impute missing values and estimate both categorical and numerical observations. The similarity-based estimation takes into account the correlation between variables. However, the choice of the parameter k is experimental. There are multiple methods to solve the problem of the k parameter [100]. The model could become unstable if k is too small or too large.

Maximum Likelihood Estimation (MLE) fits an optimal distribution to the data by comparing multiple models to determine the best fit. Fitting a distribution to the data results in a better generalisation of the data to predict further data points [101]. There are different types of distributions, such as normal, exponential, beta, gamma, Bernoulli and others [102]. Fitting a distribution to the data requires determining the shape of the required distribution and shifting the location of the centre of the distribution multiple times to obtain the best likelihood of observing the recorded measurements. Maximum Likelihood estimation uses the log-likelihood function to estimate parameters for the model. This method shifts the distribution to determine the mean and the standard deviation location, maximising the likelihood estimate of the observed data points. There are multiple methods to compare various fit distributions (models). One approach uses the Kullback-Leibler Information Theory [103] to quantify the loss of information when using a probability density function to approximate (model) another probability distribution function. Another approach to selecting the best-fit model was described by Hirotugu Akaike and known as Akaike Information Criterion [104]. It imposes a penalty on models for using a higher number of parameters and penalises more complex models. Therefore, if the criterion is small, the model will have a higher likelihood of being the best-fit model. Although the National Research Council of the National Academies advisory panel recommends methods that provide valid type I error rates (False positive), such as Maximum Likelihood under explicitly stated assumptions [105], current commercially available methods are only available for the continuous numeric type of variables. There are two types of Maximum Likelihood imputation: A Direct Maximum Likelihood and Full Information Maximum Likelihood. There remains a conservative limitation to using MLE as it relies on having strong predictor variables to produce an efficient imputation of missing values. Furthermore, it assumes that all variables that are relevant to the problem are present. Conventional maximum likelihood estimation does not work well in the presence of latent (hidden or not observed) variables [106].

Unlike MLE, Expectation Maximization Imputation (EMI), an emerging technique from maximum likelihood estimation, performs imputation in the presence of latent variables [107]. It runs two cycles. First, it estimates the values for the latent variables; this step is known as the estimation step (E-Step). The second step, the maximisation step (M-Step), optimises the model to maximise the parameters of the model explaining the data. There remain some limitations for EMI; it can be very slow, even on the fastest computer. It works best when you only have a small proportion of missing data (MAR type), and the dimensionality of the data isn't too big. Although EM accounts for latent variables, to our knowledge, the current commercially available methods are only available for the continuous numeric type of variables [107].

The practical limitation of EMI towards imputing numeric type variables is approached by creating a hybrid imputation technique known as Decision Tree-based Missing Value Imputation (DMI) [108]. Since EM is expected to perform better for a data set with higher correlations than a data set with lower correlations, DMI exploits the correlation behaviour by finding horizontal segments within a data set where there are higher correlations than the correlations over the whole data set. Finding such segments is done via a decision tree algorithm (C4.5), EMI is then applied within various horizontal segments to the numerical variables, and the imputation of categorical variables missingness is handled by the decision tree algorithm using majority class values within the leaves [108]. DMI was reported to outperform the use of EMI alone over multiple missing data patterns; however, it is not widely trailed in various studies [108].

In the 1970s, Donald B Ruben developed a powerful statistical Multiple Imputation (MI) technique. This method has three steps. First, imputation with any method of choice leading to M > 2 completed datasets (some say that 5–10 sets are usually sufficient); however, the literature suggests that the number of imputations needed in MI is a function of the rate of missing information in a data set [109]. Rubin, in 1987 provided a formula to compute the relative efficiency of M to break the infinite number of iterations. A data set with a large amount of missing information requires more imputations [110]. These M imputed datasets have the same observed values, but the imputed values are different, reflecting the uncertainty about imputation. Second, analyse each completed dataset; each parameter set estimate will differ somewhat because the data differs slightly. Finally, pool the results by calculating the variation in parameter estimates [110]. MI overcomes the problem of having small standard errors produced by statistical imputation techniques, such as regression, as it incorporates the uncertainty inherent in the imputation process. However, MI is not always quick or easy. Efficient MI requires that the missing data be ignorable, and it requires a very good imputation model. A good imputation model depends on having strong predictor variables estimating the missing values [111]. There are three types of MI: single-value regression analysis, monotonic imputation and the Markov chain Monte Carlo (MCMC) method [112].

There is a consensus that the proportion of missing data is directly related to the quality of statistical inferences. Hence, there is an agreement on the issues introduced with missing data and the occurrence of bias which specific methods can conceive in the statistical analysis if applied at different missingness proportions. The consensus is that avoiding missing data is every scientist's first call. However, there is no consensus on the exact levels (thresholds) at which missingness conditions allow a specific method of missingness treatment to be applied. Additionally, there is no established cut-off from the literature regarding an acceptable percentage of missing data in a data set for valid statistical inferences [113]. The previously mentioned thresholds of 5% for ignorability in MCAR and MAR data and 40% for variable dropping are only recommendations based on various studies [86] [89].

Also, one of Scheffer's letters in 2002 [114] highlighted a slightly different ignorability threshold of a maximum of 10% missingness levels to conclude MCAR/MAR missingness type that allows for likelihood-based imputation. Therefore, Multiple Imputation, EM imputation and regression imputation are all valid, provided the missingness mechanism is not NMAR and the percentage of missing data is not too great [114]. Scheffer thresholds are based on observing the combinations of proportion and type of missingness versus preserving the variance structure (mean and standard deviation) towards the complete data. Scheffer recommended that case-wise deletion is very wrong except in MCAR and noted that single imputation methods could work for MAR type, but only if less than 10% of the data is missing. However, if the variance structures in the data are important [115], one should not use these methods, specifically the mean imputation method, where the missing data proportion is greater than 5%.

Finally, Scheffer highlighted that MI works well, up to 25% on MNAR data. However, many modern missing data methods (e.g., MI, EMI, MLE) assume the MAR type of missingness. Thus, including variables in the statistical inferential modelling that could explain the missingness makes the MAR condition more plausible [113]. Hence, the literature often recommends adding more variables in a statistical model to make the missing data ignorable [116].

2.6.6 Feature scaling

Data comes in different distributions and magnitudes. Centring shifts numeric input attributes in the given dataset to zero-mean. Centring any attribute requires deducting the mean from all data points for that attribute. Monotonic scaling converts the features from having different dimensional units and magnitudes of measurement into unified dimensionless features (common scale) [117]. This type of scaling is also known as non-dimensionalization. Centring and monotonic scaling are considered types of feature engineering (monotonic transformation) [117].

A monotonic transformation transforms one set of observations into another set of values to preserve the order relevance of the observations [117]. It was reported that feature scaling could vary your modelling results while using certain machine learning algorithms or may have a minimal or no effect on others [118 – 120]. In particular, feature scaling is needed when a dataset contains highly varying magnitudes, units and range features. For example, the problem may emerge when using an instance-based machine learning algorithm, K-Nearest Neighbour, that utilises a distance function between two data points in their computations for modelling. KNN considers the magnitudes of features and neglects the units of measurement; hence predictive modelling results would vary greatly among different units with enormous magnitudes. The features with higher magnitudes will weigh more in the distance calculations than features with lower magnitudes. In general, algorithms that are distance based, such as KNN, require feature scaling. Feature scaling may be necessary for all curve fitting machine learning algorithms (linear/non-linear regressions), logistic regression (LR), Support Vector Machine (SVM), Artificial Neural Networks (ANN), clustering algorithms, the like of k-means clustering etc. Machine Learning algorithms which do not rely on the above conditions do not require feature normalisation (scaling). Graphical-model-based classifiers, such as Naive Bayes (NB), Decision Trees and tree-based ensemble methods, such as Random Forest (RF) and C4.5, don't require their features to be scaled. Still, it might be beneficial to standardise (scale) the data because NB calculates the probability function of events for individual variables in its prediction. At the same time, the rest utilise rules such as impurity and entropy measures and hence do not require scaling to be applied [121]. There are many techniques of feature scaling. The effect of feature scaling on modelling is also a wide area of research. In this thesis, we only can review a handful of these methods, including the Z-score, minimum-maximum, mean, unit-length and L-Norm normalisation techniques.

A) Standardisation, also known as normalisation, the Z-score scales the values of each feature in the data to have a zero mean and standard deviation of 1. Every i^{th} Recorded observation in the j^{th} numeric variable, x_{ij} is converted into a unified unit of standard deviation σ_j, where $\overline{x_J}$ is the mean of the j^{th} predictor and $\acute{x}_{ij}$ is the new standardised observation [118]. The Z-score is given by:

$$\acute{\boldsymbol{x}}_{\boldsymbol{ij}} = \frac{\boldsymbol{x}_{\boldsymbol{ij}} - \overline{\boldsymbol{x}_{\boldsymbol{J}}}}{\boldsymbol{\sigma}_{\boldsymbol{j}}}$$

This method of scaling is linear and assumes that the distribution of the variable follows the normal distribution. Therefore, a preceding transformation may be required for a variable to a normal distribution where such a transformation is often non-linear and may distort associations with other variables [118].

B) Minimum-maximum normalisation is an interval-based scaling method [118]. This method usually scales a feature's values to the range [0, 1] or [-1, 1].

$$\mathbf{A}' = \left(\frac{\boldsymbol{A - A_{min}}}{\boldsymbol{A_{max} - A_{min}}}\right) \times (\boldsymbol{D} - \boldsymbol{C}) + \boldsymbol{C}$$

A' is the new scaled min-max normalised data, *C* is the new minimum value, *D* is the new maximum value, and [*C*, *D*] is a predefined interval. This technique is more suitable for variables that follow a non-normal distribution. But it is sensitive to the presence of outlier observations [118].

C) Scaling to unit-length normalisation is widely used in machine learning to scale the components of a feature vector such that the complete vector has a length unit [118]. Assume a random predictor *X* with a set of observations x_1 to x_n, the vector length is defined by $\|X\|$. Normalised observation $\grave{x_\imath}$ where $x_i \in X$ in the n-dimensional Euclidean space $\mathbb{R}^n$ can be calculated by,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \\ \vdots \\ \boldsymbol{x_n} \end{bmatrix} \Rightarrow \|\boldsymbol{X}\|_2 = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i^2}} = \sqrt{\boldsymbol{x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2}} \Rightarrow \boldsymbol{\grave{x_\imath}} = \frac{\boldsymbol{x_i}}{\|\boldsymbol{X}\|_2}$$

This method may be best suited when dealing with variables with hard boundaries whose values are always capped with a valid range controlled by a device, such as sensor data. However, in health applications, this may not be suitable in cases where measurements in a variable don't have boundaries. Also, the effect of having outliers could be overpowering in this type of scaling [118].

D) Mean scaling can be used for algorithms that assume zero-centric data, If the maximum and minimum values of a given predictor are $x_{j(min)}$ and $x_{j(max)}$ respectively, the variable can then be transformed into a range of values. Every i^{th} recorded observation in the j^{th} numeric variable, x_{ij}, is transformed, where $\overline{x_\jmath}$ is the mean of the j^{th} predictor and $\acute{x}_{ij}$ is the new normalised observation. Thus:

$$\boldsymbol{\acute{x}_{ij}} = \frac{\boldsymbol{x_{ij} - \overline{x_\jmath}}}{\boldsymbol{x_{j(max)} - x_{j(min)}}}$$

Like the Minimum-Maximum, the Mean normalisation is also sensitive to the presence of outlier observations [119].

E) Rational scaling is based on the rational function. For each value of an attribute, 1 is divided by the attribute value. It is defined as:

$$\acute{x}_{ij} = \frac{1}{x_{ij}}$$

This type of scaling changes the arithmetic significance of outlier data points. This change may not be an issue if such data points are ignorable or removable. Still, it could be a problem in health datasets if the outliers are not ignorable, representing a phenomenon [119].

F) Logarithmic scaling is a normalisation technique that computes a predictor's log of observed values. This conversion compresses a wide range of values into a narrow range. Log scaling is particularly helpful when the variation of data values belongs only to a few instances, and the majority of your samples have little variation in recorded values. For illustration, Figure 2.4 shows a hypothetical distribution of a variable. It is observed that the greatest outcome was a zero-value at the peak of the distribution, and the smaller values make the rest of the variations of the variable’s values at the tail. Log scaling changes the distribution and helps to improve the prediction model performance [122]. This method converts a skewed distribution to a normal or less-skewed distribution. However, if a variable contains negative observations or values below 1, the log transform is undefined and cannot be applied directly. Therefore, to overcome this limitation, such values can be summed with a large known positive value that makes them greater than 1. Then, the scaler can be successfully applied only when the feature conforms to the power law.

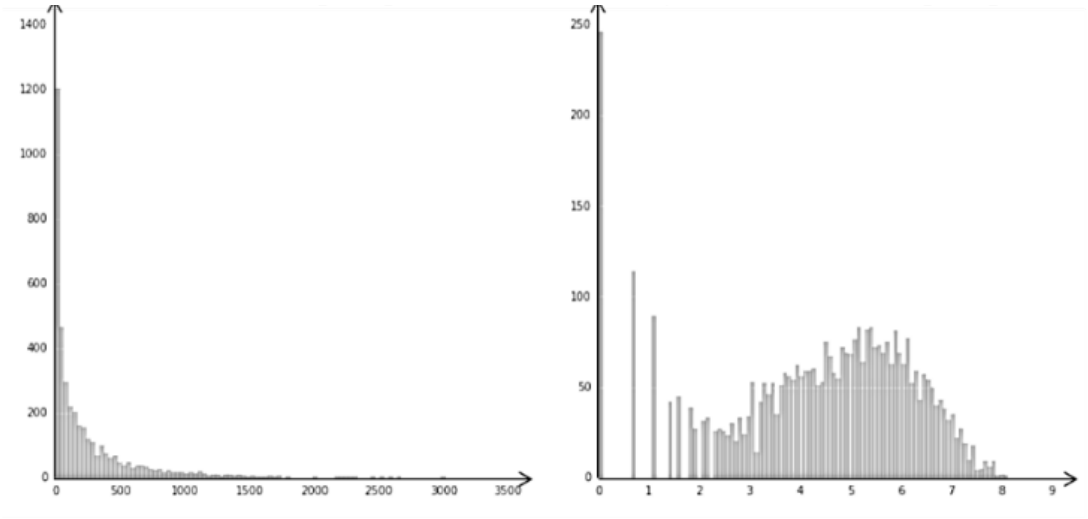


Fig. 2.4 Illustration of a distribution before (left) and after (right) log scaling

G) Maximise Normalisation is a special case of feature scaling where only the maximum value for an attribute is known. It is defined by:

$$\acute{x}_{ij} = \frac{x_{ij}}{x_{j(max)}}$$

Still, if the range is known, then it is more appropriate to normalise within the known range; this is known as **Scaling to a range** [119] [122] and is defined by:

$$\acute{x}_{ij} = \frac{x_{ij} - x_{j(min)}}{x_{j(max)} - x_{j(min)}}$$

Like all other boundary-based scalers, both methods are sensitive to outliers and extreme values.

H) Maximum Absolute Scaling automatically scales the data to a [-1,1] range by dividing every observation by its maximum absolute value. This scaler neither centres the distribution at zero nor shifts the data [123].

$$\acute{x}_{ij} = \frac{x_{ij}}{max(|x_j|)}$$

I) Robust Scalar scales the observations according to the quantile range (IQR: Interquartile Range)[123]. It essentially removes the median and scales the data between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). This method is not affected by outliers and extreme values which usually lay outside the interquartile range, hence not participating in the scaling calculations for data points. However, the outliers themselves are still present in the transformed data; thus, a separate outlier treatment is recommended [123].

$$\acute{x}_{ij} = \frac{\boldsymbol{x_{ij} - Q_1(x_j)}}{\boldsymbol{Q_3(x_j) - Q_1(x_j)}}$$

J) Quantile Transformer is a non-linear feature scaling technique that converts the variable distribution to a normal distribution. It computes a variable's cumulative distribution function (CDF) and maps its recorded observations to a normal distribution using a quantile function. This type of scaling is reported to improve the modelling performance [124]. However, although this technique is resistant to outliers, it will distort associations, including linear relationships such as correlation among variables [124].

K) L-Norm Scaling works differently from all the previous scaling techniques. Unlike the earlier scalers, the transform observations on the individual predictor values, L-Norm transforms the observations on each individual record, hence known as an instance-based scaling technique. Each instance (record) of the dataset with at least one non-zero recorded observation is scaled independently of other instances so that its norm equals 1[125].

There are different Norm indices for each of those; each defines a different way to measure the magnitude of the instance vector, i.e., L1-Norm is also known as the Manhattan Distance, which sums the absolute difference of the components of each instance vector. L2-Norm, also known as the Euclidean norm, where each observation of the sample vector is squared, this will inflate the weight of any present outliers and could skew the results.

L-∞ Norm returns the largest element within the sample vector; therefore, only the largest element within the records has any effect. However, this type of transformation completely changes the information within each variable. Furthermore, it impacts all machine learning techniques since it changes the associations and linear relationships among variables, such as correlation. Therefore, the calculations of any rules used for classification modelling, i.e., purity and entropy in the case of decision trees, for example, also change [125].

From the previous review, when considering the methods described earlier, some scaling techniques are sensitive to outliers and extreme values. Others are more robust. Also, some scalers change the underlying distribution of the data itself, leading to changes in associations among variables in the dataset. Each scaling technique has characteristics which can be leveraged to improve or deteriorate the performance of a prediction model. However, choosing the right scaler is often a trial-and-error process, and no single best scaler works every time.

Generally speaking, a scaler is deemed suitable mainly by the type of data at hand and which machine learning algorithm we apply. One safe way is to try all the scaling methods on training data and pick the best-performing scaled-data model on test data, but this is a time-consuming approach, especially for larger datasets. Additionally, from the previous review of scalers' advantages and limitations, the modelling efficiency when choosing a specific scaler is potentially dependent on multiple elements. Such elements are the distribution of the raw variable, the presence of outliers, the associations among variables, and the particular assumptions or rules made by the modelling algorithm to learn to classify.

Variable scaling has always been unsupervised, not taking the class into account; in that context, some researchers tried to rank the scaling methods based on their effectiveness in improving the accuracy performance of a collection of various machine learning classifiers [124]. But soon, they concluded that despite seeing some improvement in some models, no single scaling approach is observed that can be ranked top performer among all of the scaling techniques [124].

Other researchers took a supervised normalisation approach [126]. In 2006, Yang et al. published a new supervised attribute scaler as a pre-processing step for the KNN classifier based on the features' discriminant power. Each attribute is scaled by multiplying it with a learned weight [126]. The attribute weights are learned by taking the original labelled dataset with *N* instances and creating a new dataset with *N***K* instances, where *K* is the number of neighbours selected. To this end, each instance in the original dataset is paired with its K nearest neighbours, creating K pairs. Then, an instance in the new dataset is created for each pair with the same number of attributes as in the original data. An attribute's value in this new instance is set to the absolute difference between the corresponding attribute values in the pair of original instances. The new instance's label depends on whether the two instances in the pair have the same class label or not, yielding a two-class classification problem. A logistic regression model with non-negative coefficients is learned from this data, and the resulting coefficients are used as weights to rescale the original data. This process assumes that distance in the original space is measured using Manhattan distance because the absolute difference is taken between attribute values. The method can optionally be used to learn weights for a Euclidean distance. In this case, squared differences are taken rather than absolute differences, and the square root of the learned coefficients is used to rescale the attributes in the original data [126].

Still, there remains a limitation in Yang’s approach; all the features, excluding the class, are assumed to be numeric and do not permit the presence of missing values [126].

Researchers in another study tried to solve the problem of selecting the best scaler with a more dynamic approach [127], by which the choice of scaling techniques is automated by applying machine learning algorithms to predict the most appropriate (best-suited) scaling approach. Such a dynamic approach utilises a set of complexity measures extracted from the training datasets considering class-oriented measurements in each set as features in a dynamic classification model [127]. The data complexity measures are metrics used to quantify the difficulty of a classification problem [128][129]. There are various complexity measures for each dataset. These measures are categorised into five main measurement areas: class overlapping, class separability, geometry, topology and density of manifolds [130]. The study trained 14 popular algorithms with complexity features extracted from 48 different datasets comparing the z-score and min-max scalers. The models returned high accuracy, which may not necessarily be the best measure for all classification tasks. Due to the large computational work required to conclude the effectiveness of these models, a significant number of datasets is needed to test these models to draw a conclusion for only two scalers!

A gap in the literature remains and is worth exploring by future researchers, the use of mixed scalers within a single dataset while achieving comparable magnitudes of measurements. There may be an added value in applying multiple scaling techniques to different variables within a dataset. The added value of such local transformation usually serves the overall modelling process and enhances the features by eliminating issues such as skewness and varieties of magnitudes.

Feature scaling is an essential step in data pre-processing. Thus, it is imperative to decide which feature scaling to use. Therefore, many comparison studies of scaling methods for various popular algorithms exist. Unfortunately, this variety of studies only means that feature scaling remains a trial-and-error process without a unified approach to their application.

Finally, in healthcare, where machine learning model interpretability is important [131], scaled features’ interpretability should be mitigated. Scaling techniques alter the original observations to new mapped values losing their original clinical interpretability; therefore, the final model will be subject to interpretability issues. De-normalisation (De-scaling) can always be applied to understand the decision criteria in rule-based models such as decision trees, for example. Still, de-scaling may not be sufficient to evaluate more complex models such as curve-fitting ones. Also, in cases where the data requires error detection and corrections, feature scaling may be necessary before initiating some imputation techniques.

2.6.7 Abnormal records detection

Many machine learning algorithms and data pre-processing techniques are sensitive to the training features’ range and distribution. Outliers are abnormal values that skew the classifier and may deteriorate a model’s performance by potentially producing misleading results. They also prolong the algorithms' training time. Outliers may occur erroneously due to sensors’ faults, systems behavioural irregularities, manual or automatic fraudulent behaviours, including cyber-attacks, human error, instruments malfunction or genuinely through a natural change in population characteristics. Unfortunately, there is no agreed mathematical definition of outliers and extreme values; some may define them as the same. These abnormal observations require detection and a course of action in all cases. Such data points, in fact, depend on how you define their relative presence to the rest of the data points in a dataset. Here we define outliers as data points that appear inconsistent with the rest of the data values, while extreme values are further away, residing at the boundaries of all values for an input feature. Outliers and extreme data points can be illustrated in the scatter plot in Figure 2.5 for illustration purposes only.

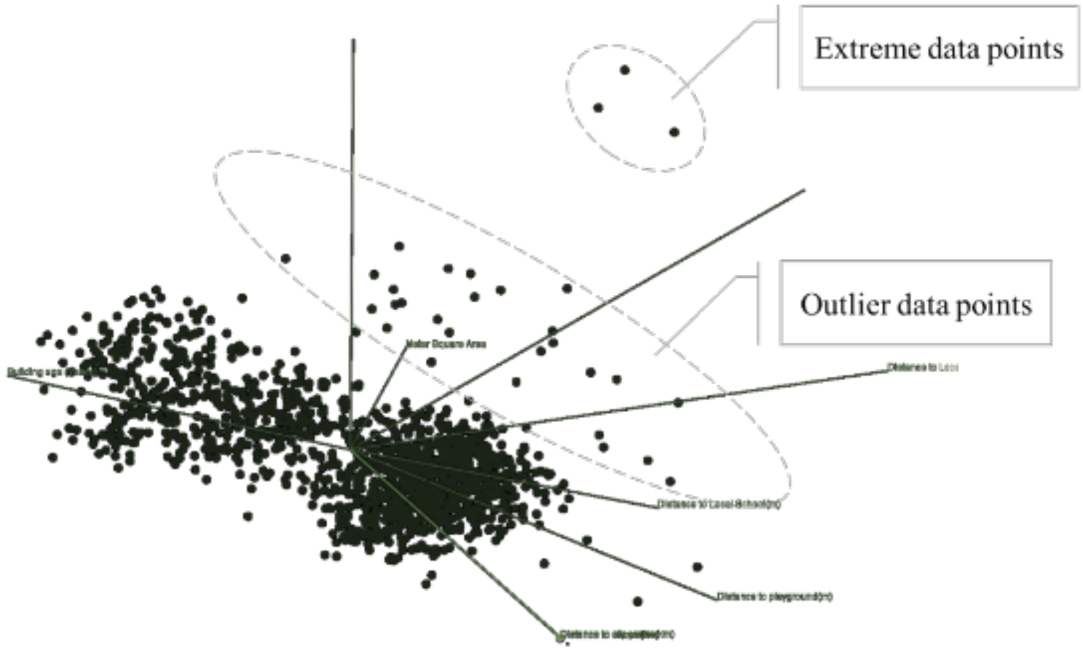


Fig. 2.5 Adaptive projection illustration of outliers and extreme values

There are many methods to detect such abnormal values [132]. However, some are more sophisticated than others. For example, visualisation techniques such as univariate plots, i.e., histograms or multivariate scatter plots, examine each variable's values or groups of related variables. Others perform statistical tests such as Inter Quartile Range (IQR) test [132].

A) Inter Quartile Range or IQR test [132] assumes a distribution and looks for values more than 2 or 3 standard deviations from the mean or 1.5 times from the first or third quartile, as illustrated in Figure 2.6.

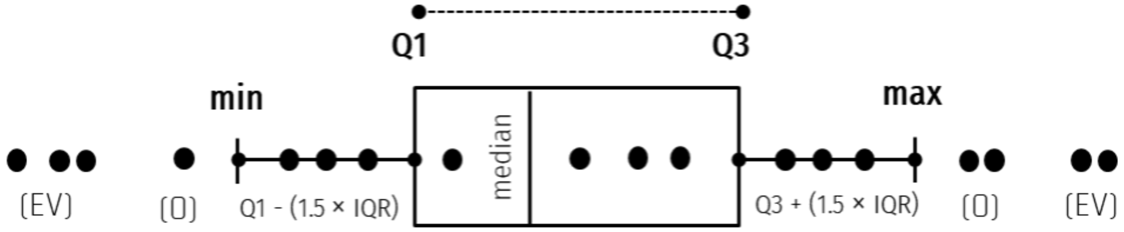


Fig. 2.6 Box plot illustrates an IQR test for detecting outliers (O) and extreme values (EV)

B) Proximity methods are more sophisticated methods that detect abnormal values via clustering methods such as the K-means algorithm [132]. K-means identifies the natural clusters in the dataset by marking the clusters' centroids. Then, by marking a fixed distance (i.e., the Euclidean distance) of the distance from cluster centroids, abnormal data points can be filtered. Feature scaling is maybe an essential preparation step before using such a technique. See Figure 2.7 for an illustration.

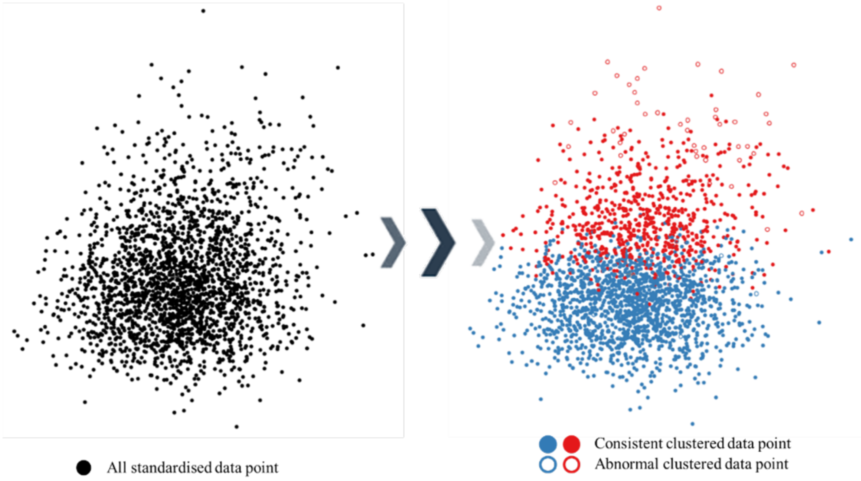


Fig. 2.7 Illustration of detecting abnormal values via K-means clustering

It is a common practice that domain experts examine abnormal data points before embarking on retaining or discarding them. Especially in health research, despite how proportionally small the abnormalities, they may be a phenomenon that needs addressing, i.e., a rare response to a new treatment. If they are errors, they may be corrected or substituted with information stored in other variables since discarding them may bias the analysis if they were proportionally large [132].

If the abnormal observations are simply errors or not a phenomenon, deleting the outliers is the most straightforward approach. If the deletion option is the researcher's choice, a similar ratio to complete-case analysis in missing data can be advised by following the 5% ratio to avoid statistical bias. Otherwise, retaining the outliers is recommended in their original form. If they are larger than 5% of all observations, advanced projection methods, such as the spatial sign technique [133], can be used to suppress their effect.

C) Advanced projection techniques project the variables' values into a multidimensional sphere [133]. In this approach, all instances have the same distance to the sphere's centre. To achieve this, all the variables must be standardised then all instances are normalised with the L-Norm function (i.e., Euclidean Norm). Figure 2.8 shows the transformation of abnormal data points.

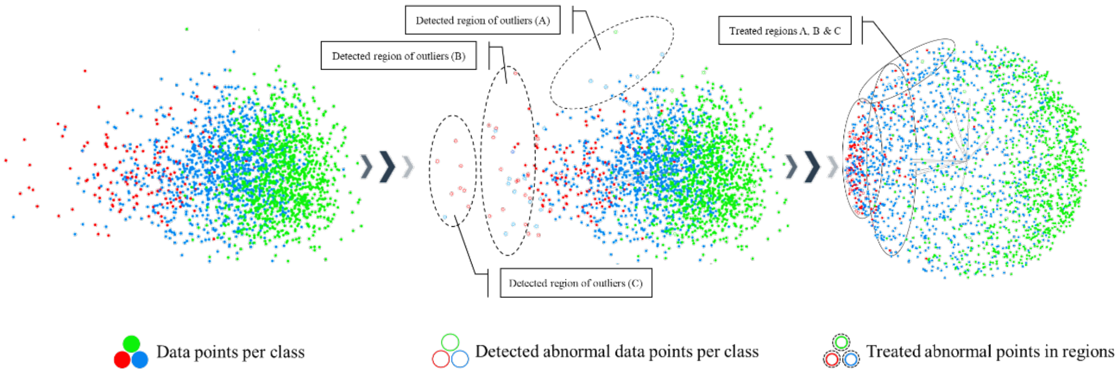

Detected region of outliers (A)
Detected region of outliers (B)
Treated regions A, B & C
Detected region of outliers (C)
Data points per class
Detected abnormal data points per class
Treated abnormal points in regions

Fig. 2.8 Illustration of advanced projection transformation of abnormalities in linear data

A limitation of applying such a technique is due to applying L-Norm which distorts the natural associations among input features. An alternative strategy is to model the data with robust learners to outliers such as Decision Trees.

2.6.8 Feature rounding

When dealing with continuous numeric variables such as fractions, ratios or proportions, the high precision in these raw values may not be needed. Hence, they could be rounded from high-precision percentages into numeric integers. These newly converted integers can be directly used as raw values or as categorical (discrete) features to train the machine learning algorithms.

However, certain types of rounding may result in a bias towards or away from a certain value, which could affect the overall model error and disturb the original data distribution. Rounding can also increase or decrease entropy, affecting the modelling output. One may think rounding is a form of discretisation (binning or bucketing) but on a larger scale. It was reported to enhance the efficiency of the machine learning algorithms training [134].

Feature rounding converts a number represented using the float or a higher precision fixed-point format into a lower precision fixed-point representation. This conversion is a matter of some importance to consider while performing modelling on fixed-point numeric attribute values. There are many feature rounding techniques [135], such as Directed rounding to an integer, Rounding to the nearest integer and Randomised rounding to an integer.

A) Directed rounding to an integer has four methods where the disposition from the original values is all directed towards or away from the same limiting value +∞, 0 or −∞. Directed rounding is used in interval arithmetic. The different types of directed rounding are rounding down (take the floor), rounding up (take the ceiling), rounding towards zero (truncating) and rounding away from zero (rounding towards infinity) [135]. Directed rounding to integer methods results in accumulating rounding errors toward individual values. Methods which round the tiebreaker towards or away from zero treat positive and negative values symmetrically; hence they are free from overall positive/negative bias if the original numbers are positive or negative with equal probability. But they still would have a bias towards zero or infinity, respectively.

B) Rounding to the nearest integer rounds a value to its nearest integer. It requires some tie-breaking criterion for cases when the observation value is precisely located midway between two integers (the fraction part of the value is exactly 0.5). By rounding a large set of fixed-point numbers with uniformly distributed fractional parts, the rounding errors with all values will statistically compensate each other, eliminating those with 0.5 fractional parts. Therefore, it is expected that the mean value of the rounded numbers is equal to the mean value of the original numbers when the numbers with fractional part 0.5 are removed from the set. There are six types of rounding to the nearest integer: round half up, round half down, round half towards zero, round half away from zero, round half to even and round half to odd. And from their names, they are centred on the tiebreak value of 0.5. [135]. Rounding half to even-odd values eliminates the tie-breaking rule bias towards positive/negative bias and bias towards or away from zero. The downside to rounding half to even-odd is that it distorts the distribution by increasing the probability of evens relative to odds or vice versa.

C) Randomised rounding rounds the fractional part 0.5 up or down without remembering the rule for rounding using a random seed. Randomised rounding has two types, Random tie-breaking and Stochastic rounding [134]. In Random tie-breaking, if the fractional part of $\boldsymbol{x}$ is 0.5, choose y randomly among $\boldsymbol{x}+\mathbf{0.5}$ and $\boldsymbol{x}-\mathbf{0.5}$, with equal probability. All other values are rounded to the closest integer. This rule eliminates overall bias. In Stochastic rounding, rounding follows one of the closest straddling integers with a probability dependent on the proximity. The probability of rounding $\boldsymbol{x}$ to $\lfloor \boldsymbol{x} \rfloor$ is proportional to the proximity of $\boldsymbol{x}$ to $\lfloor \boldsymbol{x} \rfloor$. Machine learning training can benefit from randomised rounding to an integer; it is effectively bias-free among all values despite their scaler property. Stochastic rounding is an unbiased rounding scheme with the desirable property that the expected rounding error is zero [134].

Finally, substituting floating-point observations with fixed-point arithmetic points comes with significant gains, efficiency and computational throughput while potentially may impact the machine learning model's performance.

2.6.9 Feature discretisation

This feature conversion, aka binning, transforms numeric data into nominal data by putting the numeric values into distinct groups whose length is fixed. Multiple studies reported that various methods of discretisation impact Naïve Bayes (NB) and Bayesian Network ML models' performance [136][137]. For example, Yang and Web compared the performance of the NB classifier on a dataset with nine different discretisation methods [138].

In their study [138], Naïve Bayes classifiers showed a lower classification error and an improved classifier overall performance on data pre-processed by some discretisation methods. In the past, it was well established in the literature that Naïve Bayes models show a significant performance improvement with discretisation [139]. Naïve Bayes models were always known for their efficiency, optimality and accuracy. Such an improved performance led to widespread implementations of NB learners incorporating built-in discretisation methods. Other algorithms that benefit from discretisation are Decision Trees (DT). However, the discretisation methods enhancing the NB classifier's performance may not yield the same improvement on DT [138]. The learning context of the algorithm is a core driver to witnessing such an improvement [139]. NB classifier is probabilistic and selects the class with the highest probability for a given instance. It is plausible that it is less important to form intervals dominated by a single class for NB classifier than for DT driven by decision rules based on entropy and purity. Thus, discretisation methods that pursue pure intervals might not enhance NB classifiers [138]. Besides, NB classier assumes attributes' independence. Therefore, in the NB case, there is no need to calculate the joint probabilities of multiple attribute values. Thus, discretisation methods that capture inter-dependencies among attributes might be less applicable to NB [138].

In a classification problem, discretisation methods that do not utilise the class labels in the discretisation process are considered unsupervised techniques. In contrast, discretisation methods that use the class labels are known as supervised discretisation methods. In other words, unsupervised discretisation methods are class-independent, while supervised methods are class-dependent [140]. Common examples of unsupervised attribute discretisation techniques are Equal-width Binning and Equal-Frequency Binning [141]. On the other hand, method 1R and Entropy-Based Binning are examples of supervised attribute discretisation.

A) Equal-width binning discretisation method divides the range of observations within an attribute into K intervals of equal size (range) without using the response (class) information.

B) Equal-Frequency Binning is also known as Equal-Depth (height) binning. This discretisation method divides the data into K groups, each containing approximately the same number of values (samples).

There are two key issues associated with unsupervised discretisation, first is how to select the number of bins and how to decide on their width. For the previous unsupervised methods, one way of determining K is by looking at the histogram and trying different intervals or groups.

Yang and Web developed an automatic approach, Proportional k-Interval Discretization (PKID) [142], and later developed Weighted proportional k-Interval Discretization (WPKID) [143]. Both methods were developed mainly for Naïve Bayes to discretise numeric attributes using equal frequency binning by forcing the number of bins to equal the square root of the number of numeric attribute's values. Both methods tune discretisation bias by adjusting discretised interval size and number proportional to the number of training instances. WPKID improves PKID when used for smaller datasets [142] [143].

C) Method 1R is a supervised discretisation algorithm developed by Holte [144] in 1993. It is proven effective on the standard datasets commonly used for evaluation. First, 1R sorts the continuous data values, and then the range of the continuous variable is divided into several disjoint intervals. Next, the boundaries of those intervals are adjusted based on the response class labels associated with the values of the feature (continuous variable). Each interval contains a predefined minimum number of instances, usually six instances by default based on Holte's empirical experiments, except for the last bin. The boundary adjustment continues until the next bin of values belonging to a class differs from the majority class in the adjacent interval [144].

D) Entropy-Based Binning uses a greedy split approach. In 1993, Fayyad and Irani proposed a multi-interval discretisation of continuous attributes for classification learning (MDL) [145]. It uses the information theory to minimise the entropy within a variable based on the class label. Intuitively, it finds the best split so that the bins are as pure as possible, where the majority of the values in a bin map to the same class label, which means finding the best degree of attribute separability influence on the class labels. The best discretisation is characterised by finding the split with the maximal information gain, which entails maximising the entropy function over all possible boundaries. The process of calculating the entropy is repeated recursively until the best split is selected as a binary discretisation (Stopping criterion δ is met). MDL was empirically regarded as beneficial to classification learning algorithms using information entropy minimisation to select cut-off points [145].

Generally, we acknowledge that unsupervised discretisation methods aim to maximise the interdependence between the variable observations and their corresponding class labels for a given classification problem. Also, they aim to minimise information loss while transforming from continuous to discrete values.

However, this discretisation can be disadvantageous if it reduces a classifier's ability to distinguish between the class labels. This reduction is attributed to grouping instances with different class labels into the same interval. In real-world data, it is unusual to encounter uniformly distributed data endpoints in all variables in a single dataset. Usually, there is a mixture of data points from several classes in each interval. Therefore, supervised discretisation can be performed at best to enhance the dataset in the modelling process and improve the overall model's interpretability.

2.6.10 Data-level adjustments for imbalanced-learning

Data level adjustment to improve imbalanced learning is a supervised data pre-processing strategy, aka data resampling, governed by the response labels in a classification problem. It primarily deals with real-world data problems, which often are imbalanced from various domains, including the healthcare domain. The scenario of data imbalance exists when a class of interest is not uniformly distributed among the categories. This problem occurs in two ways: a natural imbalance or a rarity of examples. Thus, the imbalance could be related to the lack of occurrences in nature for a specific phenomenon or possibly the large cost or time to collect enough data about a problem. Many practitioners plagued by the issue are working in isolation and in large research communities actively, looking into ways to alleviate the class imbalance problem [146]. Class imbalance is associated with classification modelling problems known as the accuracy paradox [147]. In this problem, the distribution of examples across the classes is skewed.

There is no consensus in the research community on what threshold must be met for a given dataset to suffer from the imbalance problem, either for binary-class or multi-class situations. The distribution can vary from a slight bias to a severe imbalance where there is one example in the minority class for hundreds, thousands, or millions of samples in the majority classes. As a result of modelling such distributions, the classifiers produce a highly accurate predictive performance on the data as a whole but very poorly for the minority class (typically the class of interest). In 2001, Weiss and Provost argued whether the class distribution of the training data should match the "natural" distribution of the data [52]. Both analysed the relationship between training class distribution and classifier performance on twenty-five data sets and concluded that the natural distribution usually is not the best distribution for learning. When the data set size is limited, a different class distribution should be selected. They quoted their empirical observation [52],

> *"When learning from a balanced class distribution, the classifiers generally come up with fewer but more accurate classification rules for the minority class than for the majority class."*
>
> **Weiss and Provost (2001)**

This observed behaviour exists because the minority class often comprises a more homogeneous set of entities, while the majority class usually corresponds to "everything else." Their results interpreted the classifier behaviour when examining the learning curves for each class label. The learning curve for the minority group was always above, the overall model learning curve positioned in the middle, while the majority-class learning curve was placed at the bottom. Thus, the test examples of the minority class always have a higher error rate than those of the majority class [52].

The proof produced from Weis and Provost's work regarded the natural distribution as often not the best choice for learning. This proof resulted in extensive follow-up research in the data science community to develop approaches to find the optimal training class distribution for classification learning by adjusting (resampling) the training dataset. The strategies followed their work were split between developing a progressive, adaptive sampling strategy that incrementally requests new examples based on the improvement in classifier performance or selecting training instances based on the current error rate for each classifier so that more data is provided to the learner has higher error rates [52]. Based on these two strategies, several approaches emerged with various techniques [148].

These approaches are random over and under-sampling, informed under-sampling, synthetic sampling with data generation, adaptive synthetic sampling, sampling with data cleaning, cluster-based methods and boosted sampling [148]. And depending on the choice of the instances of interest, these methods can be described as either selective, non-selective or combined.

A) Random over-sampling (ROS) was proposed by (Ling & Li, 1998) to solve a problem with classification in marketing data [149]. Their way to tackle class imbalance in classification is to generate new samples in the under-represented classes and append data to the original dataset. The most straightforward strategy is to create new samples by randomly sampling (with replacement) the currently available samples. This technique is a non-selective method that balances the class distribution by multiplying examples of minority class (a process of duplication). Focussed Over Sampling (FOS) is the selective variations of ROS where the replication of examples happens only at the border between two classes. Although this method makes the classes functionally equivalent, it has dire consequences. Since it appends duplicated data to the original training set, the multiple copies of examples overpower their original pattern, leading to overfitting.

B) Random under sampling (RUS) is a simple method for adjusting the balance of the original dataset. It reduces the majority class by randomly omitting instances to match the number of examples in the minority class or a closer ratio [148] [150]. Removing instances from the majority class may cause the classifier to miss important patterns within the majority class. However, RUS was reported to be effective in producing competitive clinical results against other methods in the medical domain [151].

C) Synthetic minority oversampling technique (SMOTE) is an advanced over-sampling technique introduced by Chawla [152]; it aims to enrich the minority class by generating artificial examples in the minority class instead of duplicating the existing instances to avoid the problem of overfitting. By doing so, the examples in the minority class become less rare and more general. The artificial data is generated based on the feature space similarities between existing minority records using K-Nearest Neighbour (KNN). SMOTE showed cases of success in various applications [153]. The synthetic samples help reduce the “overpowering” effect introduced in ROS. But over-generalisation was reported as a major drawback of SMOTE [154]. Overgeneralisation occurs as SMOTE inflates the distribution of the original data. The trained model then shows a sufficient small learning error and test error. These small errors give the perception that a SMOTE-trained model can apply to all other external sets just to find the opposite since the SMOTE-training data no longer represents the population of interest. SMOTE only generated further research questions on using mixed types of values and how to find the optimum data points and regions in the data to populate artificial samples.

The original SMOTE method works on continuous variables, and different variations emerged, such as SMOTE-Nominal (SMOTE-N) and SMOTE-Nominal-Continuous (SMOTE-NC) for mixed data types [155]. In 2005, Han proposed controlled versions of SMOTE’s generation of artificial samples [156], which produced further variations of the technique. Examples of which are borderline-SMOTE1 and borderline-SMOTE2, focus on examples near the classes near borderlines are oversampled [156].

Deterministic SMOTE (SMOTE-D) is a version that eliminates the random component of the original SMOTE [157] by estimating the portions of standard deviation within the data where the population of artificial samples is not required. More methods were developed by incorporating classifiers in SMOTE to find the applicable partitions of the data where the population of synthetic data is required by using a hyper-plane classifier, Support Vector Machine, to make SVM-SMOTE [158]. In 2015, Santos et al. proposed a clustering-based approach with K-means to generate new samples in minority class clusters creating CB-SMOTE [159]. Their development came after realising that the previous controlled versions of SMOTE were found later to amplify noise in datasets. Another hierarchal clustering SMOTE combination approach was soon proposed in 2017 by Ma & Fan, called Clustering Using Representatives SMOTE (CURE-SMOTE) [160], which outperformed some of the previous versions on different datasets.

D) Adaptive synthetic sampling (ADASYN) is another method that tries to solve problems introduced by SMOTE, generating the same number of artificial samples and increasing the overlapping between classes. ADASYN uses a systematic approach to adaptively create varying amounts of artificial data according to their distributions. The key idea of the ADASYN algorithm is to use a density distribution function as a criterion to automatically decide on the number of synthetic samples to be generated for each minority example by adaptively changing the weights of various minority data points to compensate for skewed distributions [161].

E) Cluster centroid under-sampling (CCUS) is a clustering-based prototype generation under-sampling algorithm. CCUS generates prototype examples from the majority class to reduce its size using K-means clustering to reach the minority sample size. The majority class is synthesised with the centroids of the K-means method instead of the original samples [162]. One downside of this method is that it can reduce the sparsity of the dataset, which could impact classifiers' decisions, such as decision trees.

F) Cluster-based oversampling (CBO) algorithm tackles the “within-class” imbalance problem. It uses the K-means clustering technique for both classes. Each training example is then assigned to the cluster that exhibits the smallest distance vector magnitude. Once all examples are assigned to clusters with each class, the CBO algorithm inflates all class clusters other than the largest by oversampling so that all clusters are the same size as the largest. This technique was also integrated with different oversampling strategies such as SMOTE [163].

G) Tomek link (T-link) removal is regarded as a guided under-sampling technique, also known as a guided under-sampling technique developed by Ivan Tomek in 1976 [164], to clean the overlapping introduced from sampling methods such as synthetic generation techniques. Therefore, overlapping samples from the majority and minority classes can be removed. A T-Link exists if the two examples from the opposite class are nearest neighbours of each other. Therefore, only boundary instances and noisy instances will have nearest neighbours. By removing overlapping examples, one can establish well-defined class clusters in the training set, which may lead to better-defined classification rules improving the classification performance. This combined method was applied to the *arterial blood pressure* data and the *Ecoli2* data set. It was reported to have enhanced the classification performance when using T-Link combined with other sampling techniques such as RUS, SMOTE, or ROS [150]. Although T-Link removes noisy examples from the dataset, many examples may be removed if the decision border is unclear between classes.

H) One-sided selection (OSS) is a hybrid under-sampling technique by Kubat and Matwin in 1997 [165]. Kubat and Matwin classify the instances in any dataset as belonging to one of four categories. The first is instances susceptible to class-label noise that will suffer due to one class-label leakage into the other class. The second type of instances is borderline instances, considered unreliable since they easily send one instance from one class to the wrong side of the classifier boundary. The third type is redundant instances which can take over by other examples in the dataset, and such instances only increase the classification cost; these can be obtained by removing all borderline and noise examples from both classes. The last type is safe instances that are easy to predict by classifiers and are worth keeping for future classification tasks [165].

With the T-link approach, OSS detects the borderline and the examples suffering from the class-label noise. Then it removes the majority class instances that participate in forming T-Links and the redundant instances using the KNN (K=1) rule to form a new training set.

The power of this technique lies in not losing any information about the minority group who are usually a condition, disease or risk-positive patients. However, a drawback to this method is that it requires a significant execution time and high consumption of processing resources [166].

I) Condensed nearest neighbour (CNN) is a method for guided under-sampling, also known as Closest Nearest Neighbour. Hart developed this method in 1968 [167]. CNN primarily aims to identify redundant observations using the KNN (K=1) rule to decide whether an example is kept or removed. The result is a subset of a collection of examples that yields no loss in model performance, referred to as a minimal consistent set. Hart quoted:

> *"…the notion of a consistent subset of a sample set. This is a subset which, when used as a stored reference set for the NN rule, correctly classifies all of the remaining points in the sample set".*
>
> **The CNN Rule (Hart Corresp.), 1968**

CNN also gives the benefit of reducing the memory requirements for the K-NN classifier. Ivan Tomek (1976) suggested two possible modifications [164] for CNN. First, since the CNN method selects samples randomly, it results in a retention of unnecessary samples and occasional retention of internal rather than boundary samples [164]. The modifications use the T-Links procedure to locate all cross-class nearest neighbours (NNs). Suppose the instances in the minority class are held constant (not subject to sampling); in that case, the procedure can be used to find all those instances in the majority class closest to the minority class and then remove them. These excluded instances are regarded as ambiguous instances [164].

J) Edited nearest neighbour (ENN) is an under-sampling technique that can be applied to each instance in the majority class, allowing those misclassified instances belonging to the minority class to be removed and those correctly classified to remain. It is also known as the cleaning under-sampling technique, developed by Wilson in 1972 [168].

ENN applies a K-NN algorithm and reduces (edits) the dataset by removing instances that do not agree enough with their neighbourhood. First, ENN identifies the three nearest neighbours (K=3) for each observation; then, it removes all observations whose class labels differ from the 2/3 nearest neighbour. Thus, for each instance in the training dataset, the three nearest neighbours are computed; if x instance belonging to the majority class is misclassified by its three nearest neighbours, then x is removed from the dataset. Otherwise, x is a minority class instance and is misclassified by its three nearest neighbours; therefore, the majority class instances among x neighbours are removed. Ivan Tomek developed an extended version of ENN called Repeated-ENN [169] that repeats the algorithm's execution multiple times, which removes more data points since the number of neighbours of the internal nearest neighbours' algorithm increases at each iteration. Tomek called this approach Unlimited ENN because, after a certain number of iterations, the training set becomes immune to further reduction. The ENN algorithm avoids the issue of over-fitting; however, like other under-sampling techniques, there is a risk of omitting useful examples from the models.

K) NearMiss under-sampling technique, proposed by Zhang & Mani in 2003, randomly selects examples from the original dataset from the majority class to reach the minority class sample size or a closer ratio [170]. Unlike RUS, Near Miss contains a heuristic for sampling. Heuristic rules are based on the nearest neighbours' algorithm, which computes the average distance to the neighbours and pre-selects the examples of interest. There are three Near-Miss under-sampling types: NearMiss-1, NearMiss-2 and NearMiss-3.

NearMiss-1 selects instances from the majority class with the smallest average distance to the three closest instances from the minority class. NearMiss-2 selects examples from the majority class with the smallest average distance to the three furthest instances from the minority class. Finally, NearMiss-3 involves selecting a given number of majority class instances for each instance in the minority class that is closest.

However, there are some reservations about the Near Miss approach. The presence of noise and outliers can alter nearMiss-1. This implies that examples of the targeted class are selected around noise examples; therefore, examples next to the boundaries are selected. NearMiss-2 does not have this effect since it does not focus on the nearest examples but on the farthest ones. NearMiss-3 is probably the version less affected by noise due to the first step in its sampling selection [170].

L) Neighbourhood cleaning rule (NCR) is an under-sampling technique developed by Jorma Laurikkala in 2001. It combines the Condensed Nearest Neighbour (CNN) Rule to remove redundant examples, and the Edited Nearest Neighbours (ENN) Rule to remove noisy or ambiguous examples. CNN is applied in a one-step manner, and then the misclassified instances by a KNN classifier are removed, as per the ENN rule. However, unlike OSS, fewer redundant instances are removed, and more focus is given to "cleaning" the retained examples; by doing so, the objective becomes less focused on improving the class distribution balance but more on the quality of instances retained to improve the classification [171].

M) Instance hardness threshold refers to instances in the datasets that are hard to classify correctly. There are hardness measures (data set complexity measures) [172] to understand why some instances are harder to classify correctly than others—the same set of measures mentioned earlier being used in investigating the efficiency of normalisation techniques.

This data sampling method looks at one complexity measure, class overlapping, that is considered a principal contributor to instance hardness. Also, the misclassification of an instance is dependent on the learning algorithm used in modelling the task it belongs to and its relationship to other instances in the training dataset. Generalisation beyond a single machine learning algorithm performance can be achieved by aggregating the results from multiple machine learning algorithms. Therefore, the instance hardness definition is based on the behaviour of a set of classification learning algorithms selected due to their diversity, utility, and wide practical applicability. A hardness property for each instance in a dataset indicates the likelihood that an instance will be misclassified. For example, outliers and mislabelled instances are expected to have high instance hardness since a machine learning classification algorithm is forced to overfit such data points to classify them correctly. Instance hardness aims to find the probability that an instance in a particular dataset will be misclassified. The notion of instance hardness is found using Bayes' theorem [173]. One reservation here is that instance hardness is classifier-dependent; this dependency bounds the probability outputs. Therefore, it is not always possible to produce a single dataset with a specific number of examples for modelling across all classification algorithms.

Sampling methods are the dominant approach in the imbalanced learning community since they are easy to implement and tackle imbalanced learning in a straightforward manner. However, it is worth mentioning that class resampling is one of many tactics to mitigate the class imbalance challenge known as the data-level adjustment strategy. Other adjustments exist at the classifier level, such as Cost-Sensitive Learning, Ensemble Methods, Active Learning and Anomaly Detection with One-Class Learning. In addition, there are algorithm-specific modifications for SVM and Kernel-based methods [146]. In the data science community, it is also believed that it is an evaluation-level problem due to having less adequate performance metrics that can be wrapped by classifiers.

In practice, we cannot favour one method over another when dealing with a class imbalance issue, especially from the literature. Various studies were conducted on different datasets and produced ambiguous results; Anand et al. conducted an empirical study and concluded that RUS data-level approach outperformed the algorithmic tactic of weighted SVM [174]. Another study by Li et al. in molecular biology used Granular Support Vector Machines (GSVMs) to under-sample the training data (GSVM-RU) and produced a high-class discrimination G-mean and accuracy of ~ 90% and concluded good performances on imbalanced data [175]. On the contrary, Liu et al. conducted an empirical study and concluded that cost-sensitive classification performance is best for unequal classes [176]. McCarthy et al. [177] compared the sampling strategy and the cost-sensitive approach to understand under which circumstances either strategy applies but later concluded that there is no consistent winner for maximising classifier performance. However highlighted, on large datasets with more than 10,000 examples, it appears that cost-sensitive learning often outperforms the resampling strategies, although it does not happen in every case. And surprisingly, the same study found that ROS beats RUS, although the behaviour differs widely for each data set [177]. Other researchers performed their own algorithmic medication to outperform cost-sensitive methods [178]. And before all, Quinlan also published a study favouring ensemble methods such as bagging and boosting to improve decision tree (C4.5) classification on imbalanced data [179].

Several research issues remain open in using sampling strategies for imbalanced learning. One concerns the acceptable level for dataset elimination, duplication, or generation. In some cases, these samples could partially represent a phenomenon, mainly in health and life sciences, such as susceptibility to an infection or a rare response to a treatment associated with influencing factors that cannot be exhaustively gathered.

Another issue relates to the use of augmented data. In recent years, there has been increased scrutiny over using synthetic data in machine learning in the health domain. There have been concerns that algorithms trained with biases in sample selection fail when deployed in health settings, and their results are sufficiently different from those acquired from the training data [180]. A recent empirical study by Vandewiele et al. investigated the use of oversampling techniques, including hybrid sampling approaches on multiple health datasets. It concluded that such resampling techniques, including SMOTE (and its variations), ADASYN, CBO and others provide overly optimistic predictions when tested on unseen data due to class label leakage [181].

Despite the above concerns, however, the medical domain seems not to dismiss such methods and accepts using synthetic data generation as a tool for scientific discovery, although adopting the built models on the artificial data for diagnoses is improper [180].

2.6.11Feature selection

After scoping all applicable variables for analysis with the domain experts' help, feature selection selects a group from the variables that may improve the modelling performance.

This pre-processing technique is a supervised strategy governed by response labels. It primarily assesses the relevance of the input feature presented to the machine learning model to enhance decision-making. Having more features may result in more discriminative power; however, the opposite behaviour was empirically reported true when Decision Tree (C4.5) and Instance-based models (KNN) showed deterioration with additional irrelevant features [182 -184]. Other algorithms, such as Naïve Bayes, are more robust when presented with irrelevant attributes [185]. Therefore, the attribute selection methods consider how the algorithm and the training set interact to achieve the best possible performance by interpreting the concept of “relevance”.

Unfortunately, there is no consensus on interpreting the meaning of feature relevance [186 – 188]. For example, historically, the term feature relevance was interpreted as:

i. A feature X_i is said to be relevant to a concept C if X_i appears in every Boolean formula that represents C and is irrelevant otherwise.

ii. Relevant features as those whose "values vary systematically with category membership.

iii. X_i is relevant if the probability of the label (given all features) can change when knowledge about the value of X_i is eliminated.

The third definition was formulated by Kohei and John in 1997 [188]. Both argued the first two definitions and demonstrated that each of those definitions leads to unexpected results and may result in the misperception of relevant features as irrelevant. They also claimed two degrees of relevance, "Strong" and "Weak". Their definition of feature relevance was driven in light of describing the Bayes algorithm as an optimal classifier. Therefore, a feature is strongly relevant if its removal alone will lead to performance degradation of an optimal Bayes classifier. This means strong relevance implies that the feature is indispensable in the sense that it cannot be removed without loss of prediction accuracy. However, a weak feature relevance suggests that the feature can sometimes contribute to prediction accuracy. Therefore, a Bayes classifier must use all strongly relevant features and possibly some weakly relevant features. Relevant features do not mean optimality, an optimal subset of features in a hypothesis space is the subset that yields the highest accuracy and adding another feature will only deteriorate the model's accuracy [188]. Also, optimality does not imply relevance. Hence, according to Kohei and John's definition [188], relevance does not imply membership in the optimal feature subset. And irrelevance does not imply that a feature cannot be in the optimal feature subset. We define feature selection as the process of finding a subset of the original attributes that yields an optimal model performance.

It is common to precede machine learning modelling with an attribute selection stage that strives to eliminate all but the most relevant attributes. This process could be helpful due to existing high throughput technologies and their recent advancements resulting in high dimensional data and the potential adverse effect of irrelevant attributes on machine learning schemes. In addition, feature selection methods reduce the dimensionality of the data by discarding irrelevant attributes to improve the performance of learning algorithms and speed up the model training time. However, depending on the feature selection method, some may outweigh the computation involved in model building. Therefore, feature selection is treated as handy but sometimes mandatory in the modelling process. Feature selection also addresses two additional challenges, models' interpretability and stability [189].

The best way to select relevant attributes is manually, based on a deep understanding of the learning problem and what the attributes actually mean [190]. However, automatic methods can also be useful. Reducing the dimensionality of the data by deleting unsuitable attributes improves the performance of learning algorithms.

There is no one way to sort feature selection methods. Azuaje, Whitten and Frank [182] classify feature selection techniques into two main types, scheme-independent and scheme-specific. Guyon and Eliddeef's survey describes three groups of feature selection techniques filters, wrappers and embedded methods [191]. We proceed with the earlier grouping, which seems more general and includes the latter.

A) Scheme-independent feature selection algorithms make an independent evaluation of features' strength or relevance based on general characteristics of the data; these techniques assess the input variables without accounting for the intended machine learning algorithm for modelling. Chandrashekar and Sahin's survey describes examples of such methods [192], including *Mutual Information techniques (MI)* such as Information Gain and Gain Ratio and *Correlation Evaluation*. Typically, these scheme-independent methods use Ranking Methods to order the attributes with a score to reflect their usefulness for the classification problem. Alongside the variables' ranking criterion (score), a threshold can be applied to remove variables below a certain calculated level of usefulness. Thus, a useful attribute in a classification problem contains useful information about the different classes in the data. Therefore, the feature that does not influence the class labels is considered useless, hence, discarded.

There are multiple caveats to scheme-independent methods [182]. First, the scheme-independent techniques do not account for interactions with other input variables; their evaluation is class label-focussed. Second, their selection criteria are driven by the assumption that a feature can be independent of the input data but cannot be independent of the class labels. Feature interactions play a major part in some machine-learning models; therefore, the subset of features conceived by these methods may not be optimal, and a similar accuracy may be achieved with a different subset of features for the same classifier. Last, their role is centric on finding the features that lead to the best accuracy. However, model accuracy may not be the most reliable performance metric for evaluating a classifier's performance in imbalanced learning.

The correlation evaluation criteria [192] are based on the Pearson correlation coefficient R. The input data $\left[x_{ij}, y_k\right]$ consists of N samples $i = 1\ to\ N$ with D variables $j = 1\ to\ D$, x_i is the ith sample and y_k is the class label $k = 1\ to\ Y$. where $cov(\,)$ is the covariance and $var(\,)$ the variance. The coefficient R is given by:

$$\boldsymbol{R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}}}$$

A limitation of the correlation ranking of variables is that it can only detect linear dependencies between the input variable and target output [192].

Mutual Information (MI) criteria are based on information theory from Shannon's definition of entropy [192]. The information contained in the output (entropy) *Y* is given by:

$$\boldsymbol{H(Y) = -\sum_{y} p(y)\log\big(p(y)\big)}$$

Once a variable X is observed in the presence of Y indicates that the uncertainty in Y is reduced, the conditional entropy is expressed by:

$$\boldsymbol{H(Y|H) = -\sum_{x}\sum_{y} p(x, y)\log\big(p(y|x)\big)}$$

If X reduces the uncertainty of Y, then the reduction of uncertainty in Y is given by: $\boldsymbol{I(Y, X) = H(Y) - H(Y|X)}$

Therefore, if X and Y are independent, they will have no mutual information, equalling zero. A variable can be considered dependent on the target variable if the mutual information is greater than zero. But since not all variables are discrete, the Kullback–Leibler divergence measure, K [193], is calculated between the two variable distributions densities by:

$$\boldsymbol{K(f, g) = \int f(y) log\left(\frac{f(y)}{g(y)}\right)}$$

There are multiple implementations of feature selection based on the mutual information concept, such as the *Information Gain (IG),* proposed by Quinlan, the *Gain Ratio* [194][195], which is an extension to the Information Gain, and the *Mean Decrease Impurity (Gini Information Function)* [196]. When these methods evaluate continuous variables, they utilise a discretisation technique such as MDL.

Research by Raileanu and Stoffel [197] showed that, theoretically and practically, there is no way to determine the best attributes' filter, as reflected in many other empirical case studies [197]. Moreover, Based on Chandrashekar and Sahin's survey [192], the modelling results of the mutual information feature selection can be poor due to the absence of inter-feature mutual information. The MI and Correlation criteria feature selection techniques are also known as filter methods since a ranker always accompanies these methods with a filtering threshold set to select $d < D$ features.

The advantage of such techniques is that they are computationally light and are reported by Guyon and Elisseeff to avoid overfitting the training data [191]. However, based on their survey, filter methods have some drawbacks too. For example, these methods may not result in choosing the optimal set of features, although redundant features might be obtained; hence, better class separation can be obtained by adding presumably redundant variables.

In terms of correlation, a perfect correlation among features indicates redundancy and adds no additional information to the classification when using them. However, a very high correlation does not imply the absence of feature complementarity. Variables complementarity can be described as a phenomenon known as information synergy, in which two features together provide more information about the class variable than the sum of their individual information [192].

Following up from feature complementarity, with MI techniques and correlation techniques, it is reported on multiple datasets that a variable deemed useless (worthless) on its own with the filter methods can provide a significant performance improvement when modelled with other predictors. The results imply that multiple worthless variables by themselves can be useful together [191].

Historically, accounting for the inter-correlation and feature redundancy in feature selection was addressed by scientists. Hall, in his thesis in 1998, proposed the Correlation-based Feature Subset (CFS) selection method [198]. CFS is an iterative filter approach that assesses the predictive ability of each variable individually in addition to the degree of redundancy among them, favouring sets of attributes that are highly correlated with the class but with low intercorrelation. Before Hall, in 1996, Liu and Setiono Brassard (based on Bartley Las Vegas Algorithm, 1996) developed a probabilistic Las Vegas Algorithm filter (LVF), known as a Consistency Filter [199] to bypass the issue of interactions between predictors by relying on the consistency of the class values when a subset of the training instances is projected on the set of filtered attributes. This method relies on randomness driven by multiple rounds of execution. Each round generates a random subset of features S from the predictors set. S has a criterion known by the inconsistency rate, which is compared to the inconsistency rate of the best subset. If both subsets are consistent, then S is regarded as the new best subset. The inconsistency rate is calculated by counting the number of instances occurring in the group with the most common class values subtracted from the number of instances occurring in the group (the inconsistency count). Finally, an overall inconsistency is calculated by summing up the inconsistency counts in all groups of a matching instance divided by the total number of instances. Lui and Setiono's tests on multiple datasets of different sizes showed that the data size could be significantly reduced by more than half when using their filter [199]. CFS and LVF approaches are usually used in conjunction with a *search engine* that seeks to find the smallest subset of attributes to fulfil the filtering criteria. In the case of the consistency filter, typically, the longer the filter is allowed to run, the better the result, as per Lui and Setiono's results.

Search Engines define how to traverse the predictors' space of all possible variable subsets to find a subset that fulfils the search evaluation criteria. Many search strategies were implemented in the past, such as *Greedy Hill Climbing*, *Best-first*, *Greedy Stepwise* and *Ranking Search*. Greedy Hill Climbing assess the local changes to the current subset of attributes. The local changes affect either the removal or insertion of a predictor into the subset. The mechanism that decides the initiation of removal is known as the direction of search, which can be forward (algorithm starts with attribute addition), backward (algorithm starts with attribute elimination) or bi-directional [191]. Best-First performs greedy hill climbing but with backtracking; backtracking allows for specifying the number of consecutive non-improving nodes to be encountered, so when the search is less promising, the algorithm backtracks to a more-promising attributes' subset [200]. Greedy Stepwise searches the attributes' space and may start with no/all features or from an arbitrary point in the feature space, just like Best-First, but without backtracking; instead, it terminates as soon as the addition or removal of the best remaining attribute degrades the evaluation criteria. It also can rank the attributes by traversing the space from one side to the other in the order that these attributes are selected [200].

Attributes selection with Machine Learning algorithms as a filter by Whitten, Frank and Hall [182] describes how a machine learning algorithm can be used for attribute selection. Machine learning algorithms can be used for attribute selection. For example, a decision tree algorithm is applied first to the whole dataset, and only the attributes that constructed the tree are selected as features. This selection is known to affect different learning algorithms. In particular, a decision tree filter could improve the performance of the K-nearest neighbour (KNN) algorithm. They reported that the new KNN model outperformed the decision tree model built with the same filtered feature. Another example is using a Support Vector Machine (SVM) for feature selection to build linear models that rank the attributes based on the size of the coefficients. Of course, one must ensure all features are scaled beforehand to ensure coefficients are comparable.

B) Scheme-specific selection algorithms. The performance of an attribute subset with scheme-specific selection is measured in terms of the learning scheme's classification performance using just those attributes [182].

Wrapper methods are a type of scheme-specific predictor selection technique. The use of wrapper methods for feature selection was implemented by John and Kohavi [201]. It is named wrapper because the learning algorithm is wrapped into the selection procedure. In other words, they use the input features as a black box and their performance as the objective function to evaluate the variable subset. Suboptimal subsets are determined by using search algorithms which find a subset heuristically. Several search algorithms can be used to find a subset of variables which maximises the objective function (classification performance). However, the search would grow exponentially for a higher number of attributes. Thus, exhaustive search methods can become computationally intensive for larger datasets. Depending on the search algorithms, the Wrapper methods can be divided into Sequential Selection Algorithms and Heuristic Search Algorithms.

Sequential selection algorithms begin with an empty set (complete set) and add features (remove features) until the maximum objective function is satisfied. To speed up the selection, a criterion is chosen that incrementally increases the objective function until the maximum is reached with the minimum number of features. Heuristic search algorithms, such as Genetic algorithms (GA) by Vafaie and DeJong [202], are applied to the wrapper methods to evaluate different subsets to optimise the objective function. Different subsets are generated by searching around in a search space or generating solutions to the optimisation problem.

Embedded Methods aim to reduce the computation time of wrapper methods for reclassifying different subsets. The primary approach is to incorporate feature selection into the training process. For example, Peng and Long proposed the max-relevancy, min-redundancy (MRMR) feature selection [203], a method based on MI. A two-stage approach is implemented; the first is to select the optimal number of features, k, which gives the lowest cross-validation classification error. In the second stage, wrapper methods are used to evaluate different subsets of size k, or direct evaluations are done on different subsets to find the subset which consistently results in the lowest classification error. In 2010, Mundra and Rajapakse provided a framework for combining MI and Wrapper approach for feature selection in Gene Ontology Analyses [204]. They reported that their method improved the identification of cancer tissues from benign tissues on several benchmark datasets.

Historically, feature selection has been investigated in data mining for decades since the early 1960s. Also, in 1978 Kitter published an early survey for feature selection algorithms [200]. Best-First search and genetic algorithms are standard artificial intelligence techniques (Goldberg, 1989; Winston, 1992) [205][206]. In 1994, John et al. experiments showed the performance of decision tree learners deteriorating when new attributes were added [207]. Almuallin and Dietterich (1991) produced the idea of finding the smallest attribute set that carved up the instances uniquely [208] and was further developed by Liu and Setiono (1996) with a probabilistic approach to select features [190]. Kibler and Aha (1987) [209] and Cardie (1993) [210] both investigated the use of decision tree algorithms to identify features for nearest-neighbour learning; Holmes and Nevill-Manning (1995) used OneR to order features for selection [211]. Kira and Rendell (1992) used instance-based methods to select features, leading to a scheme called Relief for Recursive Elimination of Features [212]. In 2004, Gilad-Bachrach et al. showed how this scheme can be modified to work better with redundant attributes [213]. Prior to that, in 2000, Hall implemented the correlation-based feature-selection method [214]. Wrapper methods for feature selection originated in 1994 from John et al. [207]. Later in 1997, Kohavi and John developed a framework for the wrapper approach [188]. Guyon et al. in 2002 presented and evaluated the recursive feature-elimination scheme in conjunction with support vector machines [215]. In 2009, Gütlein et al. investigated how to speed up scheme-specific selection for datasets with many attributes using simple ranking-based methods [216]. In 2017, Lui et al. studied meta-feature selection evaluators [2017], one of which is cost-sensitive feature evaluation by weighting or resampling the training data according to a supplied cost matrix or by making the base attribute evaluator cost-sensitive to demonstrate the significance of cost-sensitive feature selection for the real-world imbalanced data [217].

In our view, whichever feature selection method is applied, there is a possible limitation in many predictive modelling applications, where the structure of the problem being studied evolves over time or differs by population, i.e., "non-stationary" or “heterogenous”. Both of these can introduce systematic differences between the training and test sets rendering some models useless. Nevertheless, models also may be incorrectly specified and vary by modeller biases and/or arbitrary choices of features. When this occurs, there may be an illusion that the changes in the external samples negatively impact the predictions made, whereas the actual reason could be that the model has missed a critical predictor and/or included a confounded predictor. Perhaps, including all logically available applicable features in the model, regardless of their calculated relevance, is more appropriate so any future issues of data structure changes would surface.

2.6.12Data Errors detection and handling

Earlier in this literature review, we established outliers and extreme values detection methods are subjectively effective in spotting inconsistent data points. The detected data points via these methods may or may not be erroneous. But as a result, follow-up actions taken are to handle these inconsistencies before the modelling process.

Here we focus on validating consistent data points that hold erroneous values but hide in the data distribution from abnormality detection techniques. In such scenarios, conventional statistical methods may fail to spot these errors for correction. They can be troublesome to identify without domain experts' help and a clear strategy to detect such points for correction or removal.

While domain advice is specific within every case study, the applied strategy to discover these errors can be the same for all studies. Therefore, we adopted two techniques, Equivalence Class Partitioning (ECP) and Boundary value analysis (BVA), from the Test Engineering discipline to identify system defects. These are adapted to spot erroneous data points empirically [218].

Applying both methods to datasets requires prior knowledge of the data collection and survey design. The same knowledge was described previously to perform logical rule imputation and form valid assumptions about the data for information-based imputation.

In 2008, Beer and Mohacsi used both methods to efficiently generate test data by covering all semantic dependencies plus all (n-dimensional) boundaries with a minimum set of test data [218]. In 2015, Baht and Qadri conducted an analytical and empirical investigation and produced a framework to compare these two testing techniques effectively [219]. ECP and BVA methods exploit semantic relationships and dependencies within $n \times m$ dimensional dataset. However, as described by Arnicane in 2009, applying both methods is highly complex [220].

We are the first to explore using both methods in data science research. Reasons for its non-existence in data science literature may include the time-consuming manual work involved. Nevertheless, its adoption might be useful in clinical data science research projects where data was manually collected while considering the criticality of the prediction task of the health-related outcome. We must emphasise that applying such a strategy will risk prolonging the data preparation cycle. Therefore, if the predictive task is classed as critical, non-critical modelling tasks can be done without them.

A) Equivalence class partitioning (ECP) is generally applied with boundary value analysis (BVA). In this technique, we divide the records in the dataset for a lead variable into partitions of equivalent data (classes) that can be considered the same [220]. This technique tries to define test cases (test observation) that uncover partitions containing errors in other related variables. Suppose that P is a dataset with N number of variables X_i, where $1 \leq i \leq N$. for each observation value domain D_i is partitioned into M_i equivalence classes with extreme points as boundary values. The meaning of boundary value testing is to examine the related variables' observations values when the lead variable's observation values assume extreme values for each equivalence class (maximal, minimal), just above (for some small value ε) or just below the extreme values; and when value is nominal – inside the equivalence class in the distance from extreme values that are considerably bigger than ε [220].

B) Boundary value analysis (BVA) reserves those input values at the extreme edges of the input domain that are subject to more errors in a system. More data errors are discoverable at the boundaries of the input domain. BVA tests the data observations to identify errors at border values (valid or invalid zones) rather than finding those that exist at the centre of partitions [220]. This way, we could discover values out of range as well as errors in data values propagation to other variables. Let a set of test records be $X_1, \ldots, X_n$ and assume that there is an ordering relationship defined over them as $\leq$. Let C_1 and $C_2$to be two equivalent classes, assume $X_1 \in C_1$

and $X_2 \in C_2$. If $X_1 \leq X_2$or $X_2 \leq X_1$, then the classes $C_1$and $C_2$are in the same neighbourhood, and the values $X_1, X_2$are boundary values which means that values on the minimum and maximum edges of an equivalence partition are tested since these boundaries are common locations for errors that result in data propagation inconsistencies. Arnicane described the previous explanation graphically described [220] in Figure 2.9.

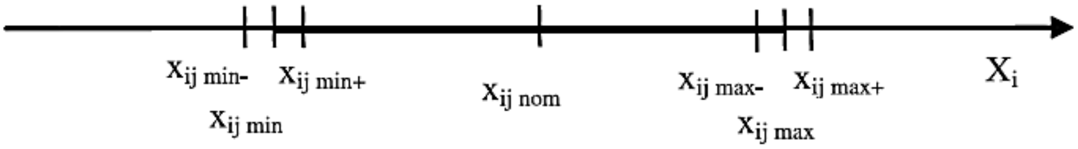


Fig.2.9 Equivalence Class and Boundary Value Analysis
Equivalence class d_{ij}: $[x_{ij\,min}, x_{ij\,max}]$, its boundary values $x_{ij\,min}, x_{ij\,max}$, inner OFF

For the corresponding domain D_i of each variable's X_i, each equivalence class d_{ij} of the ordered elements can be graphically represented, as shown in Figure 2.9. The minimal boundary value of the class is $x_{ij\,min}$, the maximal boundary value is $x_{ij\,max}$, where $1 \leq j \leq M_i$. Nominal value of the class is $x_{ij\,nom}$. Values $x_{ij\,min-}$, $x_{ij\,max-}$ are a little smaller than appropriate boundary values, but $x_{ij\,min+}$, $x_{ij\,max+}$ are a little bit bigger. For simplification, we use $min-$, min, $min+$, nom, $max-$, max, $max+$ instead of $x_{ij\,min-}$, $x_{ij\,min}$, $x_{ij\,min+}$, $x_{ij\,nom}$, $x_{ij\,max-}$, $x_{ij\,max}$, $x_{ij\,max+}$

Boundary values $x_{ij\,min}$ and $x_{ij\,max}$ may belong to an equivalence class, but they can also be tested from it. Nevertheless, they are boundary values for this class [220]. The following inequalities hold for each i, j when $1 \leq i \leq N$ and $1 \leq j \leq M_i$.

$$x_{ij\,min} - x_{ij\,min-} \leq \varepsilon_{ij}$$

$$x_{ij\,min+} - x_{ij\,min} \leq \varepsilon_{ij}$$

$$x_{ij\,nom} - x_{ij\,min} \geq \varepsilon_{ij}$$

$x_{ij\,max+} - x_{ij\,max} \leq \varepsilon_{ij}$

$x_{ij\,max} - x_{ij\,max-} \leq \varepsilon_{ij}$

$x_{ij\,max} - x_{ij\,nom} \geq \varepsilon_{ij}$

Values that are just above the minimal value and just below the maximal value are described as inner OFF points (they are inside the equivalence class), and values which are just below the minimal value and just above the maximal value can be called outer OFF points [220].

When performing these BVA, we form two assumptions [220]:

i. For each variable X_i, the conjunction of all equivalence classes is the main D_i.

$$D_i = \bigcup_{j=1}^{M_i} d_i \;\; \forall\, i,j \;\; where \;\; 1 \leq i \leq N \; and \; 1 \leq j \leq M_i$$

ii. There are no shared values between equivalence classes $\forall\, i,j,k$ where

$$1 \leq i \leq N, 1 \leq j \leq M_i, 1 \leq k \leq M_i \; and \; j \neq k \quad d_{ij} \bigcap d_{ik} = \emptyset$$

Under the data collection rules provided by the data source, we assume a valid propagation path of data values for groups of records to other related variables. Then we examine for any violations of data values, presence, or missingness for those particular records. This test justifies the systematic existence of values, but it also detects the presence of impossible combinations of values due to human manual input errors and data misalignment. Invalid data here does not necessarily mean that the data is incorrect; it means that it may have leaked outside a specific partition [220].

BVA and ECP assist in finding and correcting errors in the raw data before preprocessing; their assumptions help minimise missingness with information-based and logical rule imputations by substituting missing observations with true likeness. They also help discover qualitative and quantitative errors in the raw data. The discovered errors may fall under two causes: Pattern *Violations,* which include value-type, encoding, data linking, conversion and data formatting errors, and *Rule Violations*, which are usually caused by the propagation of conflicting values among variables in the dataset.

2.6.13Data worthfulness

The dimensionality of data could indicate the data's worthfulness for building a machine learning classification model. This section reviews important characteristics of data dimensionality and their perceived effect on classification modelling. There are four fundamental dimensionality measurements in a classification dataset: the number of overall records, the number of attributes, the count of labelled instances, and the count of labelled positive examples in a classification problem.

In short, we are looking at how much data is enough for reliable modelling results. Our term modelling-worthfulness refers to having a dataset whose dimensions keep the risk of overfitting low. Researchers used these four dimensionality measurements to form data characteristics to help pre-assess modelling outputs' reliability. These characteristics include Events per Variable Ratio (EPV), Samples per Feature Ratio (SFR), sufficient events and other considerations.

A) Events per variable (EPV) ratio. For many years, one common rule mentioned in the literature [221] is based on the Events per Variable (EVP) ratio. The rule of 1:10 has been a well-known rule of thumb for modelling with Logistic Regression. Depending on the count of the positive instances (positive class), the 1:10 rule states that for every ten instances in your positive class, one variable can be used for modelling [221]. Therefore, if a training set of 2000 patients are to be modelled and 200 patients in the positive class (so that 1800 patients are in the negative class), the one in ten rule implies that twenty pre-specified features can reliably be fitted to the training data. There is also a mention in the literature for even stricter limits, including the 1:20 and 1:50 rules, while keeping in mind that the 1: 10 rule is the minimum for reliable modelling. When the 1:10 rule is violated, the number of features is considered too large for the training data set and will result in overfitting [221]. In 2007, however, Vittinghoff and McCulloch relaxed the one-in-ten rule. They recommended that one feature per five to nine instances is sufficient [222]. Depending on the research question and their large simulation study, they found a range of other factors, such as variables type (binary vs continuous predictor), that were as influential as or more influential than EPV. Nevertheless, they concluded that larger training sets with more positive instances are almost always preferable, while keeping in mind that situations may arise where confounding cannot be persuasively addressed without violating the EPV rule of thumb. Thus, the results should be interpreted cautiously and compared with those from other models [222].

In 2016, a study published by Smeden et al. described the EPV rule for binary logistic regression as weak, given their simulations, especially where the outcome prediction was perfectly separable [223]. Again in 2019, Smeden et al. challenged the EPV in another study [224], describing how the total training set size (positives and negatives instances) and the positives proportion (positives to total dataset size ratio) can be used to calculate the expected prediction error on (a test set) of the model that is to be developed. Thus, an estimation can be made to the required dataset size to achieve an expected prediction performance with a lower error rate than a predetermined (estimated) prediction error [224]. Their simulations with multiple modelling approaches showed that the EPV 1:10 rule could be either too lenient or too strict, depending on the adhered modelling approach [224]. Their study also found that EPV had only a weak relationship with outcomes of prediction error and a mediocre relationship with the model's capability of discrimination. However, despite their estimations, they also advised that the performances of their modelling simulations were not externally validated. Hence, their approximations may not be efficient outside the scope of their simulations' settings, especially in cases where the positive instances count is very low; their methods of estimations may yield very poor binary discrimination performances [224].

B) Samples per feature ratio (SFR). On the contrary to using the EPV ratio, Foley's sample and feature size considerations [225] were generally accepted by the machine learning community. It recommended that a robust classifier requires a Sample per Feature Ratio (SFR) of at least 5 – 10.

C) Data dimensionality practical considerations. Other data science practitioners relate the answer to how much data is needed for the type of machine learning problem at hand [226] based on the machine learning problem at hand. Nevertheless, they state there is no golden rule for every problem. Other unique cases were argued by Somorjai et al. in gene microarray classification studies [227]. In such studies, the EPV ratio is often violated. Still, promising model performances were reported in the literature, showing near-perfect classification accuracy on both training and independent validation (test) sets [227]. However, the number of modelled instances per class is ~10 – 30, whereas the original count (M) of attributes is in the thousands. Even after deploying feature selection techniques, the number M of 'optimal' attributes found is ~50–100 for microarray [227]. An argument often used is that if perfect or near-perfect classification performance is achievable on both the training data and an independent validation test set, then the results must be reliable, indicating a robust classifier [227]. This argument could be plausible, but in the context of typical microarray datasets, such an argument is generally unwarranted, and the reliability of the results may be considered an illusion. The culprits responsible for the reliability challenge in the context of microarray dataset modelling are based on Bellman's curse of dimensionality (too many attributes) and the curse of dataset sparsity (too few instances) [227].

Microarray datasets are a special case, typically, each instance is attributed to several thousand features, yet only a few instances are available for analysis [227]. Therefore, although good classification performances are still achievable, the large number of features leads to interpretability problems [227]. Somorjai et al. then refer to the choice of the number of features in a model to satisfy the two key objectives for creating a classifier in the first place, and these are high generalization power, where unknown data points are classified correctly and meeting the required medical/biological interpretability. Interpretability aids disease management and treatment and would benefit from having only a few biologically relevant (applicable) features. However, in most cases, producing robust classifiers comes at the cost of achieving this second goal [227], where the data sparsity is an issue, preventing isolating a test set for evaluation. Additionally, this issue may render out-of-sample testing such as 10-fold Cross Validation and a Leave-One-Out (LOO) Cross-Validation unfeasible to conclude a reliable classifier, but rather a reasonable classifier performance estimate [227].

Finally, we believe that one should mention the empirical choice of the learning algorithm as an important ingredient to the mix in addition to the type of machine learning problem, sparsity, generalisation, and interpretability constraints in determining the number of features.

For example, a Naive Bayes classifier can have many more features than samples, yielding a reasonable performance [188]. Logistic Regression limitations could be mitigated by incorporating L-1 regularisation, such as Ridge Logistic Regression, with many more input features while avoiding overfitting. On the other hand, the choice of deep Neural Networks could easily overfit the data with fewer features and many more examples. SVM does not require feature space reduction, although feature selection for SVM is beneficial [227]. While capping the number of features plays a hero role in a model's fitting and interpretability, the use of regularisation overcomes the fitting problem. It also shifts the focus to interpretability complexity in the context of the number of parameters. In addition, when fitting an ensembled tree algorithm such as random forest, and due to their non-parametric nature, the more relevant consideration to pay attention to is the tree size than the number of features. The complexity of interpretability of such models cannot solely be capped by the number of features but more on the number of parameters the model has, which grows with increased instances.

D) Sample size and sufficient event labels is a new under-development area of research. Researchers utilised information theory and Cohen's D size effect to predict the number of required positive labels (events) in a dataset to observe an effect size, give a statistical power, significance level and a fixed degree of freedom [228]. Thus, they created tables to guide practitioners to ensure that they mean the minimum count of positive labels within the data to observe desired mutual information, especially in cases where labelling the data is expensive. This also touches on semi-supervised learning, notably learning from Positive unlabelled data (PU) or the PU constraint [228]. With variable statistical significance (p-value), one can determine if there is an effect but cannot tell how large the effect is. Cohen's D, or standardised mean difference, is one of the most common ways to measure effect size [228]. As a rule of thumb, if the analysis is statistically significant, then a Small Effect = 0.2, Medium Effect = 0.5 and a Large Effect = 0.8. In 2014, Sechidis et al. produced a framework to generate tables [228] to provide guidance to practitioners on the required sample size and positive label size to observe an effect size. In their conclusion, they state that the sample size is application dependent. Thus, their framework encourages practitioners to collect a larger number of examples as a common practice to increase the likelihood of capturing the prevalence of a disease. And only if the collection of instances/labels is a costly matter can one take a riskier approach, but with an informed decision, using fewer instances/labels [228].

The framework developed by Sechidis et al. could be helpful if applied during the data collection phase to ensure enough prevalence is captured, to define the cut-off point of data collection. However, this does not seem beneficial in cases where the data has already been collected for data mining analysis. Still, their study might infer further consideration to dealing with collected unlabelled data.

2.7 Chapter summary

This chapter introduced a variety of preparation methods and considerations that reside before data modelling. The chapter briefly describes the methods' working mechanisms. It presents in the available literature and practice the limitations and implications of each presented method in every section and subsection when applied.

After presenting the general data-driven modelling workflow, the chapter starts with modelling conceptualisation. Without doing so, it would have been impractical to pool applicable methods for this review. The chapter then continued with reviewing and critiquing the essential areas that are required by most data-driven modelling projects, including Data Quality Assessment (DQA), Scales of Measurements, Data Visualisation, Data Preparation and Feature Engineering.

The theoretical and practical recommendations, critique and limitations provided in each subsection highlight gaps in knowledge which the author will consider when building a methodology later on for predictive modelling in preventive medicine.

When building and applying this thesis methodology in later chapters, we will weigh the recommendations made here. Also, we intend to pay extra attention to the methods that are often criticised in medical research, such as missing data mitigation, interpretability and sampling methods for imbalanced learning.

Chapter 3

Modelling & Evaluation Literature Review

This chapter covers and reviews

- *The scope of machine learning tasks*
- *Classification learning workflow*
- *Training, validation and test sets split*
- *Out-of-Sample testing*
- *Machine learning algorithms selection*
- *Processing ordinal class labels*
- *Imbalanced learning classifier adjustment*
- *Classifiers fitting, overfitting and underfitting*

The development of Machine learning has accelerated over the past four years and is poised to keep growing. Although data-driven research, and more specifically with machine learning, already has a long history in biology and chemistry, it only rose to prominence recently in health and medical sciences. In this chapter, we review the literature for the machine learning modelling tasks, principles and common algorithms used in modelling. It identifies the impact of various parameter settings on the presented machine learning algorithms where required. The chapters will also present various recommendations or reservations made by researchers to enhance modelling performances and identify any gaps in knowledge.

Class-Imbalance is one of the challenges impacting machine learning classification algorithms' performances in data mining, where establishing a fit-for-purpose model assessment measure is almost the primary research issue [146]. Many studies demonstrated that skewed class distribution often leads to biased classification and evaluation misjudgement [147]. To address this critical matter, this chapter also describes a series of fundamental machine-learning classification-performance metrics, which are later ingested into combined and Graphical performance evaluation metrics. Also, we discuss the significance of the interpretations of the performance metrics' calculated output. This chapter also sheds some light on the application of these performance measures and their limitations when evaluating the fulfilment of the classification success criteria.

3.1 The scope of machine learning tasks

In data science and machine learning, there are four types of learning, Supervised, Unsupervised, Semi-supervised and Reinforcement learning. The main difference between these types is the level of availability of ground truth data, which is prior knowledge of the model's output for a given input. Given the characteristics of the given datasets in our research, in the previous chapter we conceptualised that patients' susceptibility to adverse events falls under the supervised learning type in this thesis, namely classification. In classification, and depending on the class labels, there are four common problems: binary-class, multi-class, multi-label and imbalanced classification. Unlike binary and multi-class classification, where a single class label is predicted for each example, multi-label classification predicts one or more class labels for each example. From Chapter 2, and based on Parkin and Balbus [16], the scope of predicting patients' susceptibility to adverse health events falls under two states, a binary-class classification task with two labels, Resistant vs Susceptible or a multi-class classification task with three labels, Resistant, Susceptible and Hyper-susceptible [16].

Imbalanced classification, a common problem in classifying health-related outcomes and specifically in our research, exists when the number of examples in each label is unevenly distributed [146][147]. Specialised techniques for imbalanced classification exist for different phases of the classification task [178], including specialist resampling techniques in the data preparation phase, described in the previous chapter. In addition, there are specialised algorithmic methods [178] and specialised evaluation metrics [178] that will be described and discussed in this chapter.

3.2 Classification learning workflow

The workflow for supervised ML is split into two main workflows: Training and Predicting (Testing). Both workflows are linked together and can also be very iterative in nature. Figure 3.1 shows an illustration of the workflow of supervised ML. From the workflow, it is apparent that an additional dataset (test set) is required to evaluate the classification models. In addition, not in Figure 3.1, there is also a recommended requirement of having a validation set also. The definition of validation and test always varies in scientific communities. There seems to be a common misuse of the terms “test set” and “validation set” in applied machine learning. In 2007, Ripley provided a specific definition of training, validation, and test sets [229] in the field of pattern recognition as follows:

- A **training set** is a set of examples (cases or instances) used for learning to fit the learnable parameters of the machine learning algorithm (Learner) to create a model (classifier) [229].
- A **validation set** is a set of examples that tunes the parameters of a classifier, such as choosing the number of hidden layers and nodes in an artificial neural network (ANN) [229].

- A **test set** is a set of instances used only to assess the performance of a fully specified classifier [229].

Brownlee's definition [230] also agrees with Ripley's. The validation set gives an unbiased evaluation of a classifier fit on the training set. But he adds that the evaluation becomes more biased as the classifier becomes skilled on the validation set incorporated into the model configuration [230]. The test set provides an unbiased performance evaluation of the final model fit on the training set. The final model can be fit on the aggregation of the training and validation sets. One should not dismiss that some other domain experts may refer to a validation set as a test set [230].

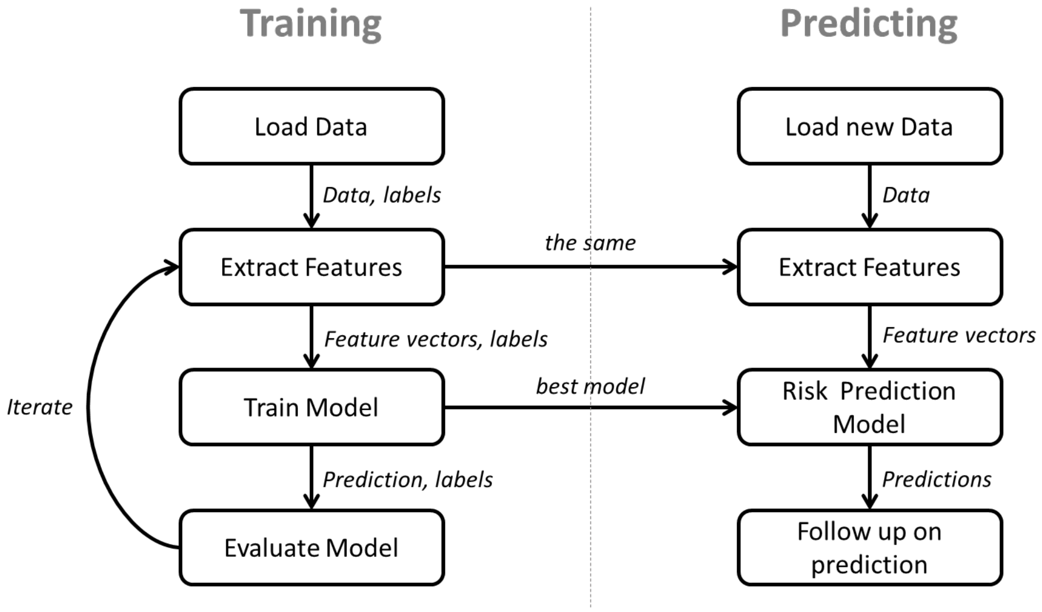


Fig. 3.1 Illustration of supervised learning workflow

3.3 Training, validation and test sets split

Following the illustration and the value of having these training, validation and test sets, it is common that these sets are resampled from one big dataset. In some cases, the entire big set is made available; in other cases, the evaluation test set is extracted, produced or collected at a later stage. There are a couple of machine learning approaches to building models. One uses training, validation and test sets, and the other uses a sharp training and test split [230]. Either approach debates the ratio of split for such sets. There seem to be different recommended split ratios, but in each justification of these approaches, the relevance to the size of the dataset at hand remains.

In 1997, Guyon addressed the problem away from the relevance of instances count in each class [231], relying on solid dependence on the complexity of the learning process. Guyon states that the fraction of patterns reserved for the validation set should be inversely proportional to the square root of the number of free adjustable parameters [231]. In other words, it depends on the capacity of a classification model, which is related to how complicated the model can be. Based on Guyon, for any dataset, if there are 32 adjustable parameters in the learning process, the inverted square root of 32, which is ~ 0.177 or 17.7%, is the fraction of data that should be reserved for validation and 82.3% for training [231]. In Guyon's conclusion, it is stated that her framework is not perfect but simplifies the trade-off relationship between the training set and validation set [231].

We interpret Guyon's framework that different machine learners will require different training-validation sets ratios. Therefore, if we follow a comparative machine learning approach for the same dataset, the presence of the same external test set is imperative to compare multiple models' performance. The comparison may become more tedious if incorporated by resampling methods for imbalance learning.

Some researchers [232] consider the Pareto principle [233], also known as the 80/20 rule or the principle of factor sparsity, a rule of thumb that states that for many outcomes, roughly 80% of consequences come from 20% of causes. This principle was developed in 2008 by Joseph M. Juran in the context of quality control naming it after Italian economist Vilfredo Pareto, who noted the 80/20 connection while at the University of Lausanne in 1896 [232].

The popularity of the 80/20 rule comes from its description by a power law distribution (AKA Pareto distribution) for a particular set of parameters. Many natural phenomena exhibit such a distribution in physics, biology, earth and planetary sciences, economics and finance, computer science, demography and the social sciences [233].

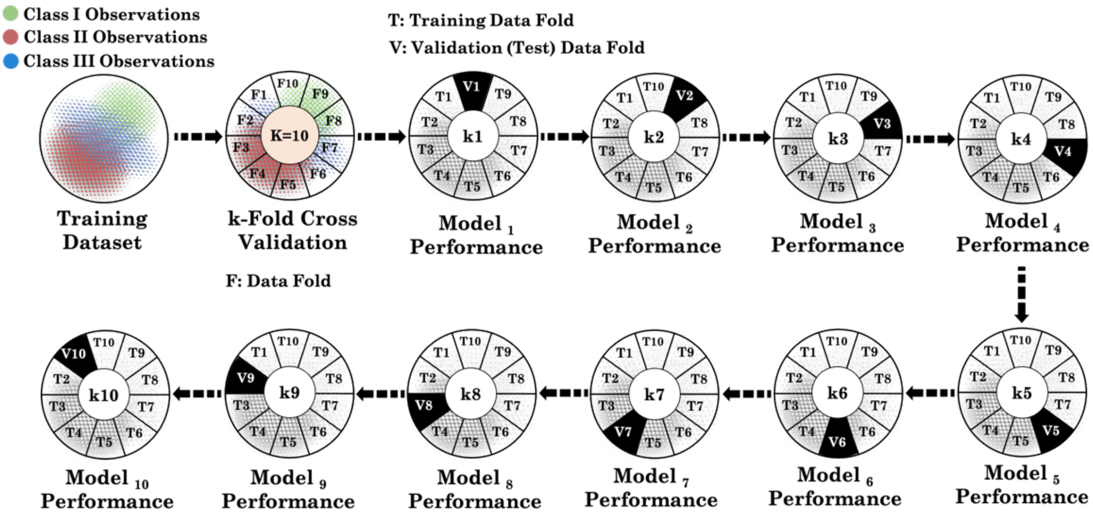
Others claimed good model accuracy results by reserving a third of the dataset for test and two-thirds for training [234]. However, accuracy is not the best metric to evaluate the classification of imbalanced learning. Nevertheless, their results showed that the training sample size influences the accuracy of the classification [234]. On the importance of the Training–Test split ratio, a study by Pawluszek-Filipiak and Borkowski in 2020 [235] empirically concluded that the accuracy measure alone should not be used to evaluate the classification results, especially while accounting for imbalanced learning with other metrics. Further, they recommended that the training–test ratio be around 1 (the training area should be as large as the testing area or very close) [235].

Finally, when splitting the training and test data, there are two competing concerns; with fewer training examples, the parameter estimates may have greater variance, and with fewer test records, the model's performance evaluations may have greater variance. From this point of view, we assume that attention should be paid when dividing the data so that neither variance is too high. Also, the split should account for the instances count in each class label rather than the split ratio only. The model variance-bias trade-off is explained later in this chapter.

3.4 Out-of-sample testing

This data resampling technique is used to validate a supervised machine learning model and estimate how well it will perform (generalise) on unseen data (test data). In addition, it may provide some assurances on how accurately a predictive model will perform in practice. Multiple strategies exist to do this, such as Cross-validation, a resampling strategy that uses different dataset segments to train and test a model on different iterations. The simplest is two rounds of cross-validation involves partitioning a sample of data into subsets, performing the training on one subset, and validating the model on the other subset and vice versa. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds to estimate the model's predictive performance. There are two types of cross-validation, exhaustive and non-exhaustive cross-validation.

A) K-fold cross-validation is a non-exhaustive cross-validation method. The original dataset is randomly segmented into k equal-sized subsamples (non-overlapping folds). Of the k subsamples, a single fold is reserved as a validation set for testing the model, while the remaining $k - 1$ folds are used in training the learner. The cross-validation process is then repeated k times, with each k subset used only once as a validation set. The k results are averaged to calculate a single estimation (See Figure. 3.2). In general, k remains an unfixed parameter. Typical values are k =3, k =5, and k =10; the most used value in applied machine learning is k =10. The popular choice of k =10 is due to various published studies that found it to provide a good trade-off of low computational cost and low bias in estimating a model's performance [236 – 238]. In 2021, Marcot and Hanea [239] conducted experiments to obtain the optimal k with datasets of various counts of samples $n = \{50, 500, 5000\}$, then they assessed the classification success and with seven levels of folds $k = \{2, 5, 10, 20, (n – 5), (n – 2), (n - 1)\}$. Their work supported the commonly used $k = 10$ in the literature although, in some cases, $k = 5$ would suffice with large samples (n = 5000) [239]. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times computation time to make an evaluation.



🟩 Class I Observations
🟥 Class II Observations
🔵 Class III Observations
T: Training Data Fold
V: Validation (Test) Data Fold
Training Dataset
K=10
F1 F2 F3 F4 F5 F6 F7 F8 F9 F10
k-Fold Cross Validation
F: Data Fold
k1 V1 T2 T3 T4 T5 T6 T7 T8 T9 T10
Model 1 Performance
k2 T1 V2 T3 T4 T5 T6 T7 T8 T9 T10
Model 2 Performance
k3 T1 T2 V3 T4 T5 T6 T7 T8 T9 T10
Model 3 Performance
k4 T1 T2 T3 V4 T5 T6 T7 T8 T9 T10
Model 4 Performance
k5 T1 T2 T3 T4 V5 T6 T7 T8 T9 T10
Model 5 Performance
k6 T1 T2 T3 T4 T5 V6 T7 T8 T9 T10
Model 6 Performance
k7 T1 T2 T3 T4 T5 T6 V7 T8 T9 T10
Model 7 Performance
k8 T1 T2 T3 T4 T5 T6 T7 V8 T9 T10
Model 8 Performance
k9 T1 T2 T3 T4 T5 T6 T7 T8 V9 T10
Model 9 Performance
k10 T1 T2 T3 T4 T5 T6 T7 T8 T9 V10
Model 10 Performance

$$Final\ Model\ Performance = \frac{1}{10}\sum_{k=1}^{10} Model_k\ Performance$$

Fig. 3.2 Illustration of k-fold cross-validation

B) Holdout random subsample is the simplest non-exhaustive type of cross-validation, where the dataset is separated into two sets: the training subset and the testing set. First, the learner is fitted using the training set only. Then the algorithm is asked to predict the output values for the data in the test set (isolated unseen data). And finally, the errors it makes are accumulated to give the mean absolute test set error, which is used to evaluate a supervised regression model (the class labels are a range of continuous values). The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. Furthermore, the evaluation may depend heavily on which data points end up in the training set and which end up in the test set. Thus, the evaluation may differ depending on how the division is made [240].

C) Repeated random sub-sampling validation is another non-exhaustive variant of cross-validation that randomly divides the data into a test and training set *k* different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over, and the proportion of the training-validation split does not depend on the number of iterations. The disadvantage of this technique is that validation subsets may overlap; some instances may never be selected in the validation subset, whereas others may be chosen multiple times [240]. As the number of random splits approaches infinity, the result of repeated random sub-sampling validation tends toward that of Leave-One-Out Cross-Validation (LOOCV). A stratified variant of this method is beneficial in imbalanced learning.

D) Leave-one-out cross validation (LOOCV) is a special case of k -fold cross-validation, with k equal to n, the number of data points in the set. LOOCV means that for n separate times, the learner is trained on all the data except for just one data point, and a prediction is made for that point. Although the evaluation given by leave-one-out cross-validation is good, LOOCV is an exhaustive method and very expensive to compute [240].

E) Nested cross-validation is known as double cross-validation, a non-exhaustive technique used in the hyperparameter optimization procedure of a model on a dataset. When the same cross-validation procedure and dataset are used to both tune and select a model, it is likely to lead to an optimistically biased evaluation of the model performance [241].

The challenge in estimating the machine learning algorithm's parameters is that there is no good heuristic for configuring the model hyperparameters for a dataset. Instead, an optimization procedure is used to discover a set of hyperparameters that perform best on the dataset. *Grid-Search* is an example of an optimisation procedure, and each distinct set of a model's hyperparameters is usually evaluated using k-fold cross-validation [241].

The k-fold cross-validation method effectively estimates a model's performance. However, using it multiple times with the same algorithm can lead to overfitting, as described earlier in the classification workflow.

Whenever a model with different model hyperparameters is evaluated on a dataset, it provides information about the dataset. This knowledge about the model on the dataset can be exploited in the model configuration procedure to find the best-performing configuration for the dataset; the result is an overly optimistic estimated model performance that does not generalize to new data. Although k -fold cross-validation reduces this effect, it cannot be eliminated completely. In 2010, the issue of knowledge leakage causing overfitting with k -fold cross validation was presented by Cawley and Talbot [242].

Thus, the k-fold cross-validation for model hyperparameter optimization is nested inside the k-fold cross-validation procedure for model selection (two cross-validation loops in Figure 3.3). Hence, the name "*double cross-validation*". This way, the hyperparameter search does not have an opportunity to overfit the dataset as it is only exposed to a subset of the dataset provided by the outer cross-validation procedure. The whole process should reduce the risk of the search procedure overfitting the original dataset and give a less biased estimate of a tuned model's performance on the dataset [243].

The disadvantage of nested k-fold cross-validation is the increased number of model evaluations performed. If $n \times k$ models are built and evaluated as part of an outer cross-validation hyperparameter search for a given model, then with the inner loop, this is increased to $k \times n \times k$ as the procedure is then performed k more times for each fold in the outer cycle of nested cross-validation. Therefore, it is common to use a larger k for the outer loop and a smaller value of k for the inner loop [243].

Along with the computational cost of the cross-validation procedure, here come some limitations; the validation and training sets must be drawn from the same population. In many predictive modelling applications, the structure of the system being studied evolves over time (non-stationarity). This limitation can introduce systematic differences between the training and validation sets. Furthermore, there is evidence in published studies that cross-validation alone is not very predictive of external validity [244]; therefore, another form of external validation may be required.

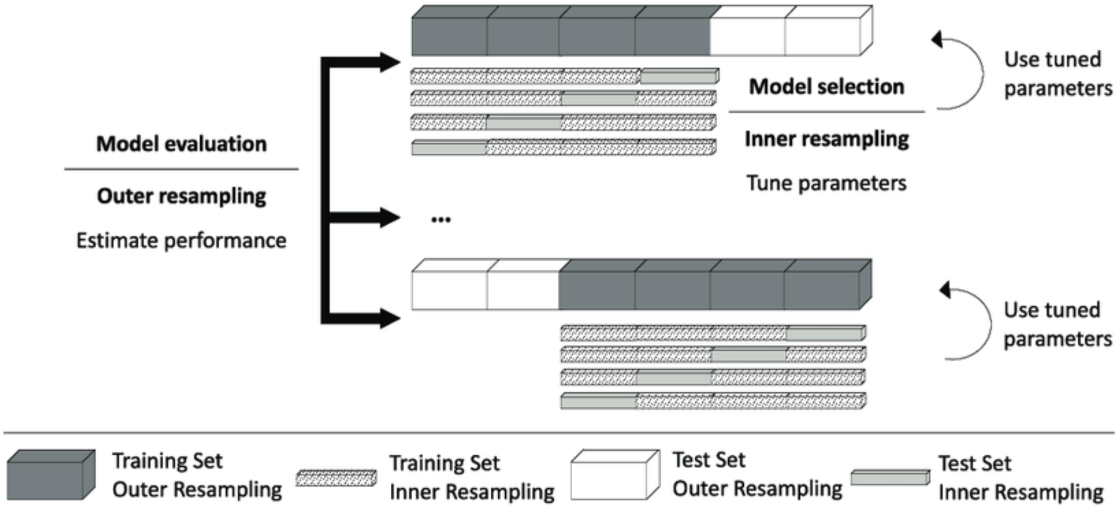


Fig. 3.3 Illustration of nested *k* -fold cross-validation

3.5 Machine learning algorithms selection

Generally, machine-learning classification models can be either parametric or non-parametric. In a parametric model, the number of parameters is fixed with respect to the sample size. In a non-parametric model, the number of parameters can grow with the sample size. When modelling, researchers tend to take a comparative approach to evaluate the performance of each of these types [245]. The impact of using either type in data mining was investigated by Khire et al. in 2021 [246]. It was found that the performance of the algorithms measured in accuracy, recall, precision, and F1-score for non-parametric algorithms, like Random Forest and Decision Trees, outperformed the parametric algorithms, such as Logistic Regression and Naïve Bayes, more specifically under the circumstances of imbalanced classification [246].

In 2019, Brownlee, a data scientist, a well-known writer and blogger in the data science community, categorised machine algorithms based on their learning style, supervised, unsupervised and semi-supervised, but preferred grouping them based on their functional similarity and purpose [247]. Based on Brownlee's categorisation, machine-learning algorithms (learners) can be grouped into Regression, Instance-based, Regularization, Decision Trees, Bayesian Algorithms, Clustering, Association-Rules, Artificial Neural Networks, Deep Learning, Dimensionality Reduction and Ensemble learners. Examples of such algorithms are grouped in Figure 3.4.

Brownlee’s grouping is not exhaustive [247]. For example, other algorithms are used in speciality tasks in the machine learning process, such as data preparation and transformation. Also, the grouping does not contain speciality subfields of machine learning, i.e., Computer Vision, Natural Language Processing (NLP), Recommender Systems, etc.

Modelling with all possible algorithms’ variations on a particular study's dataset is unpragmatic. Therefore, for our research purposes, we examined eight machine-learning algorithms. Some of which are parametric and others non-parametric. Our choice accounts for the variation of machine-learning algorithms based on their functional differences that are well established in the literature [248]. And according to the ‘no free lunch’ theorem [249], there is no optimal classifier that works perfectly for every group of problems, as the classifier's performance depends mainly on the domain problem and the data used.

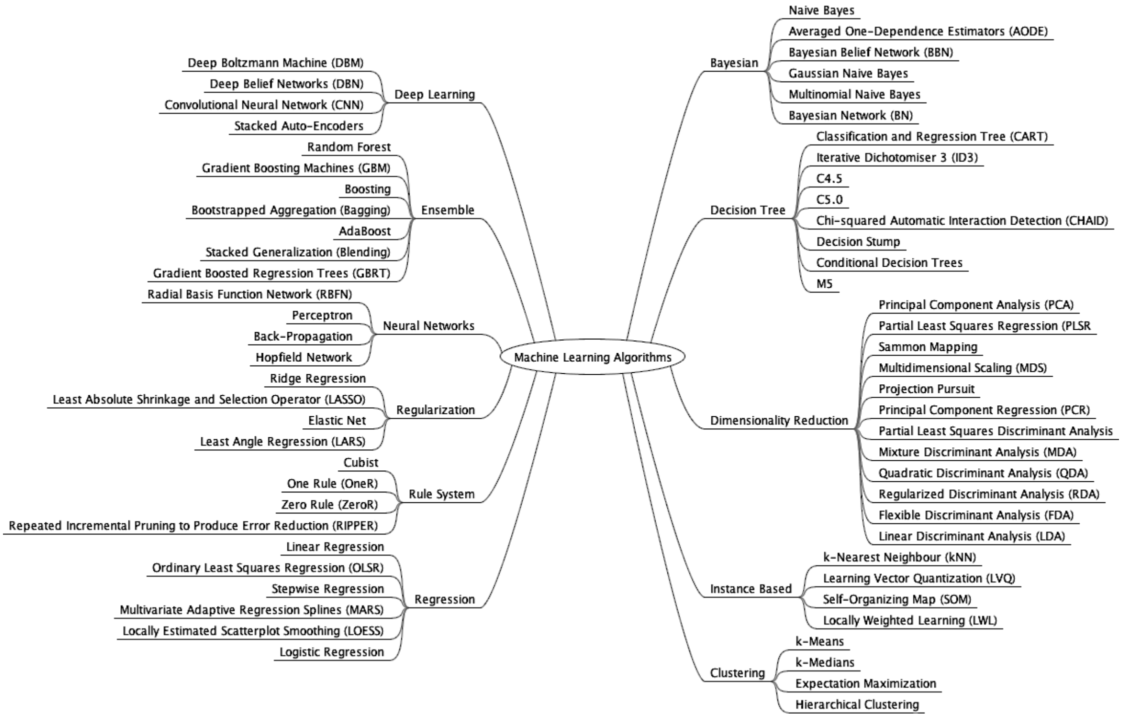

Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

Fig. 3.4 Brownlee's machine-learning algorithms categorisation

Therefore, we review a number of selected classification algorithms, Naïve Bayes (NB), Logistic Regression (LR) with Ridge Estimator, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Decision Tree (C4.5), Multi-Layer Perceptron (MLP), Logistic Model Tree (LMT) and Random Forest (RF). We also present some specific cases of algorithmic modifications as published in the literature for handling ordinal-class classification and imbalanced learning.

A) Naïve Bayes (NB) learner is a parametric and probabilistic modelling algorithm. It is popular in medical diagnosis research [250] [251]. Among all the different approaches used in medical diagnoses, Naïve Bayes is considered one of the most effective and efficient classification algorithms, successfully applied to many medical problems [252]. NB assumes that all predictor variables have an equal effect on the response outcome and that all predictors are independent and have no interactions.

The foundation of the classifier is based on Bayes Theorem, where ***B*** is the evidence and ***A*** is the hypothesis, is given by:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Or

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Where y is the response variable (class variable), and X is a set of features mapped to the labelled response. $X = \{x_1, x_2, x_3, \cdots, x_n\}$, therefore,

$$P(y|x_1, x_2, x_3, \cdots, x_n) = \frac{P(x_1|y)P(x_2|y)P(x_3|y) \dots P(x_n|y)P(y)}{P(x_1)P(x_2)P(x_3) \dots P(x_n)}$$

$$P(y|x_1, x_2, x_3, \cdots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

$$P(y|x_1, x_2, x_3, \cdots, x_n) = \frac{1}{Z} P(y) \prod_{i=1}^{n} P(x_i|y)$$

The above is an independent feature model where Z is the evidence representing the scaling factor dependent on $x_1, x_2, x_3, \cdots, x_n$. The independent feature model is combined with a decision rule to construct a classifier. One of the commonly used rules is the most probability, the Maximum a Posteriori (MAP) decision rule. Therefore, the Bayes classifier is expressed as,

$$\hat{y} = arg\, max_y\, P(y) \prod_{i=1}^{n} P(x_i|y)$$

Despite the demonstrated naïve assumptions of Naïve Bayes classifier in the previous equations, hence the name, these assumptions work quite well in many complex real-world situations. For example, when dealing with medical data, the Naïve Baye learner accounts for evidence from many attributes to make the final prediction and provides transparent explanations of its decisions. Thus, it is considered one of the most useful classifiers to support physicians' decisions [252]. There are multiple implementations of the NB algorithm with embedded pre-processing techniques, thus, from our review of data preparation in the previous chapter, we established that NB could benefit from feature discretisation, i.e., MDL discretisation technique [145].

B) Logistic regression (LR) learner is another popular parametric deterministic algorithm used for prediction analysis in medical research [253]. LR produces a probability-based model that takes into account the probability of an event occurring (the class label) depending on the values within the predictors (categorical or numerical). LR estimates the probability of an event occurrence for a random observation compared to the probability of the non-occurrence of the same event. In other words, it predicts future observations for a categorical class variable. The goal of the logistic regression is to estimate the probability p for a linear combination of independent variables denoted by $\boldsymbol{\hat{p}}$.

Due to the categorical nature of the dependent variable (the class) in logistic regression, its distribution follows the *Bernoulli distribution*; therefore, logistic regression links the predictor variables to the Bernoulli distribution in a process called the Logit, and is given by:

$$\boldsymbol{logit(p) = \; ln(p) - ln(1-p)}$$

The inverse logit function $\boldsymbol{logit^{-1}(\alpha)}$ is also known as the Mean function $\boldsymbol{\mu_{(y|x)}}$ returns the probability of an event occurring in the predictor variable. The mean function is given by:

$$\boldsymbol{\mu_{(y|x)} = logit^{-1}(\alpha) = \; \frac{e^{\alpha}}{1+e^{\alpha}}}$$

Where $\boldsymbol{\alpha}$ represents the linear combination of independent variables and their coefficients. The LR coefficients' calculations are made using an algorithm called the Maximum likelihood Estimator (MLE) [101]. Thereafter:

$$\boldsymbol{logit(p) = log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\chi_1}$$

$$\boldsymbol{\frac{p}{1-p} = e^{(\beta_0+\beta_1\chi_1)}}$$

$$\boldsymbol{\widehat{p} = \frac{e^{(\beta_0+\beta_1\chi_1)}}{1+e^{(\beta_0+\beta_1\chi_1)}}}$$

Where $\boldsymbol{\beta_0}$ is the bias and $\boldsymbol{\beta_1, \beta_2, \beta_3, \dots, \beta_n}$ are the logistic regression coefficients associated with a set of predictors $\boldsymbol{X = \{x_1, x_2, x_3, \cdots, x_n\}}$ expressing the expected change in the log odds of having the labelled outcome per unit change of $\boldsymbol{X}$

Therefore, for a subject $\boldsymbol{i}$, if the estimated probability $\boldsymbol{\widehat{p_\iota}}$ exceeds 0.5, the subject is classified into the occurrence group; otherwise, it is classified into the non-occurrence group as follows:

$$\boldsymbol{\widehat{p_\iota} = \frac{e^{\left(\beta_0+\beta_1\chi_{1_i}+\beta_2\chi_{2_i}+\beta_3\chi_{3_i}+\cdots\beta_n\chi_{n_i}\right)}}{1+e^{(\beta_0+\beta_1\chi_1)}}}$$

Classical logistic regression estimates its parameters using the Least Square method. However, the Least Square Estimation has issues dealing with multicollinearity in data. Multicollinearity can affect a logistic regression model with more than one predictor and occurs when two or more predictor variables overlap so much (highly correlated) in what they measure that their effects become indistinguishable and potentially could lead to data overfitting.

One way to avoid such a problem is to use Regularisation to avoid overfitting by penalising high-valued regression coefficients [254]. In reality, any data set can be fitted to a model, even if that model is ridiculously complex. Therefore, regularisation is used to reduce the parameters and to shrink (simplify) the model. It adds penalties to more complex models and then sorts potential models from least overfit to greatest; The model with the lowest "overfitting" score is usually the best choice for predictive power. Regularisation works by biasing the data towards particular values. The bias is achieved by adding a tuning parameter **λ** to encourage those values. There are L1 or L2 regularisations. L1 regularisation adds a penalty equal to the absolute value of the magnitude of coefficients (limiting the size of the regression coefficients). L2 regularisation, also known as Ridge Estimator [254], adds a penalty equal to the square of the magnitude of coefficients (shrinking the regression coefficients). The tuning parameter **λ** controls the strength of the penalty term. When **λ = 0**, ridge regression equals least squares regression. If **λ** = ∞, all coefficients are shrunk to zero. The ideal penalty is, therefore, somewhere in between 0 and ∞.

In 1992, Cessie and Van Houwelingen embedded the use L2 regularisation method in logistic regression, known as *Logistic Regression with Ridge Estimator***,** to avoid overfitting and produce simpler models. This variation is adhered to reduce the potential effect of *multicollinearity* [254].

C) Artificial neural networks (ANN), a parametric learner, showed exceptional ability to fit observed data, especially with high dimensional datasets. Hence capable of modelling the extensive amounts of information currently available to clinicians, ranging from details of clinical symptoms to various biochemical data and outputs of imaging devices.

ANNs are popular in medical research due to their adaptive learning approach and ability to handle diverse medical data and integrate them into categorised outputs [255]. In 1958, Frank Rosenblatt designed the fundamental element of a neural network, the perceptron, which simulates the human brain [256]. An ANN is made of elementary autonomous computational units. Each has inputs, a bias and an output known as *neurons*, illustrated in Figure 3.5.

When values are processed as inputs to the neurons, each input is multiplied by a weight value which gets adjusted (optimised) all the time during the training phase via the Gradient Descent optimisation algorithm.

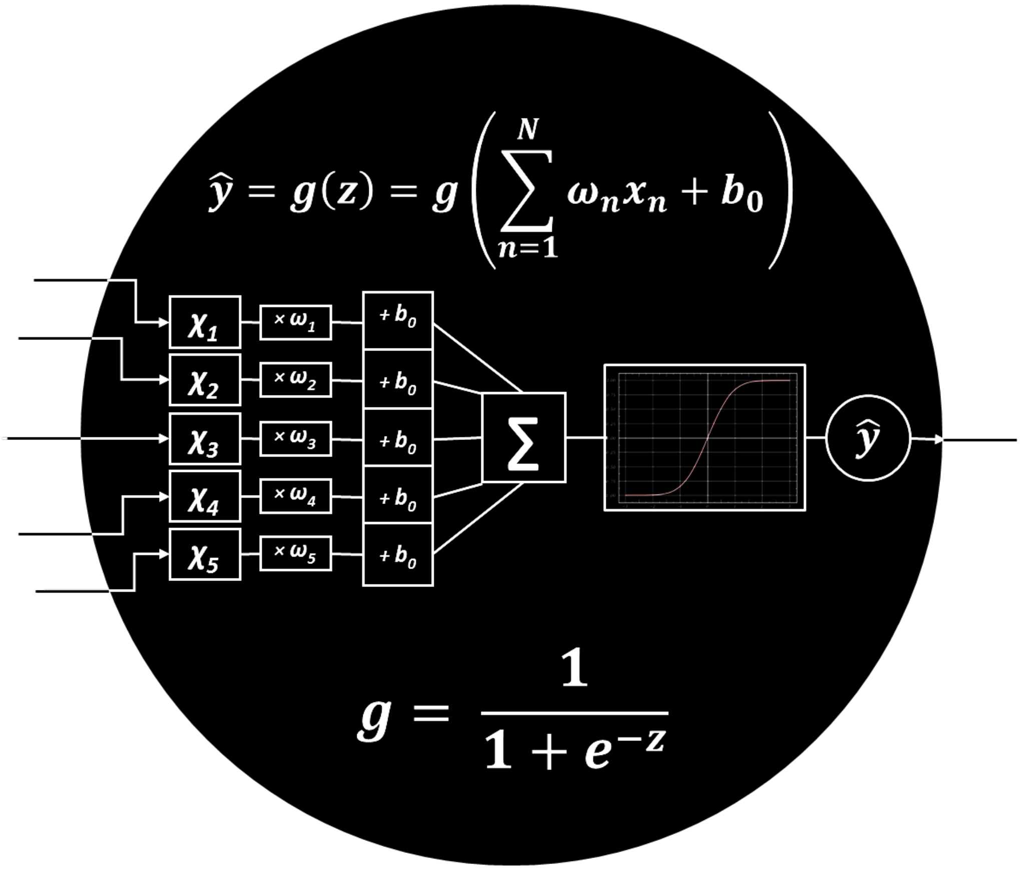


Fig. 3.5 Illustration of the perceptron with a sigmoid activation function

The neurons are inter-connected via weighted connections given by:

$$z = \sum(\boldsymbol{weight} * \boldsymbol{input}) + \boldsymbol{bias}$$

ANNs are usually organized in layers forming various structures, known as architectures. Depending on the processed data type and the application, the number of layers, the direction of the flow of computation and the type of connections differ. Examples of ANN architectures range from the perceptron, the most basic element, and more advanced ones, such as the Multi-Layer Perceptron (MLP), Long-Short-Term Memory (LSTM) and Deep Convolutional Networks (DCNs). See Figure 3.6 for an illustration of the MLP network. In our research, the Multi-Layer Perceptron (MLP) ANN architecture is used for classification, being the most commonly used architecture for data-driven prediction in data mining.

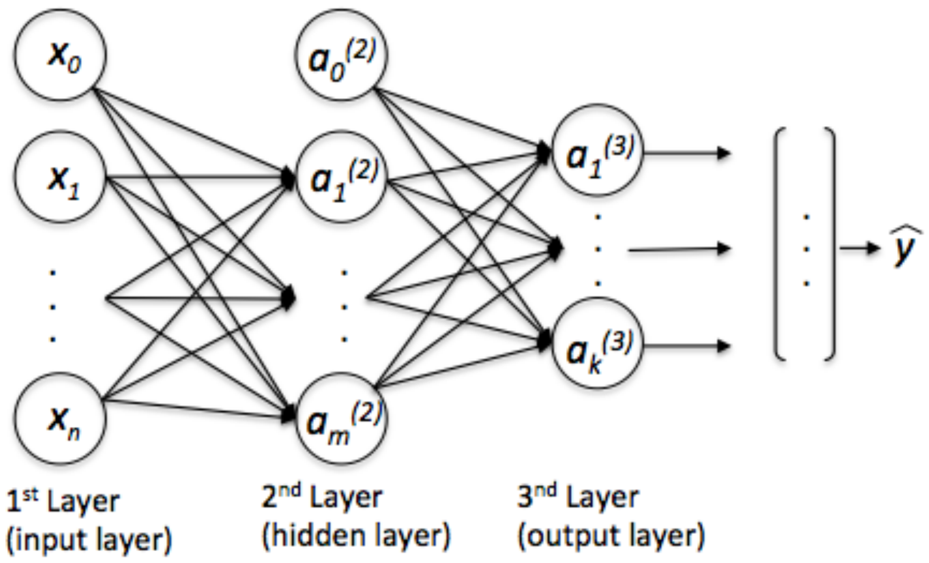


Fig. 3.6 Illustration of the multi-layer perceptron (MLP) ANN

From Figure 3.6, the ANN basic layout consists of an input layer which passes the input values to the next layer without any operations applied, hidden layers with neurons which transform the different input data, and finally, the output layer gives the desired number of values and in the desired range.

The primary aim of training an ANN is to update this weights value to decrease the error. The neurons are interconnected from one layer to another via weighted-synoptic connections. The bias (offset) represents an input for each neuron with an always value of 1. In case of all inputs are missing (no inputs), this ensures that there will be an activation function to the neuron. The activation function, also known as the transfer function, introduces non-linearity to the neural network. The sigmoid activation function is used throughout our research and given by:

$$\boldsymbol{g = \frac{1}{1+e^{-z}}}$$

$$\boldsymbol{\hat{y} = g(z) = g\left(\sum_{n=1}^{N} \omega_n x_n + b_0\right)}$$

The activation function can be either linear or non-linear and control the outputs of neural networks across different problem domains from object recognition and classification. However, various activation functions may be preferred in the data science community, such as Reflected Linear Unit (ReLU), Scaled Exponential Linear Unit (SeLU) and Hyperbolic Tangent (TanH) [257].

A 2018 survey by Nwankpa et al. compared most of the activation functions used in Machine Learning and Deep Learning [257]. The survey outlined the current trends in the applications and usage of these functions in practical deployments against the state-of-the-art research results. Their comparison aimed to assist in making effective decisions when selecting the activation function for any given application. They established that the sigmoid function was recommended for predicting probability-based output since it has been applied successfully in binary classification problems, modelling logistic regression tasks, and other neural network domains. They also highlight the main advantages of the sigmoid functions as being easy to understand and used mostly in shallow networks [257]. Hence, our practical projects adopted the sigmoid function for our MLP classification networks.

D) K-nearest neighbour (KNN), a non-parametric learner, was created in 1967 by Marcello Pelillo and has been used in many data mining applications, including compensating for missing data, as mentioned briefly in the previous chapter. In this chapter, we review its workings.

The NN algorithm is built upon considering examples as vectors in a high-dimensional space, allowing us to apply geometric concepts to machine learning. Therefore, distance computations are one of the most basic calculations that can be computed in a vector space. There are multiple techniques to measure such distances in D-dimensional space, Minkowski Distance, Manhattan Distance, Cosine Distance, Jaccard Distance, Hamming Distance and Euclidean distance [258]

In general, the Euclidean distance is the most widely used distance metric. It is the default metric in machine learning tools and libraries for K-Nearest Neighbour implementation [258]. It is a measure of the true straight-line distance between two points.

In 2019, Alfeilat et al. produced a state-of-the-art review to determine the best distance metric that can be used for a KNN learner [258]. Their study attempted to answer this question by assessing the KNN models' performances (measured by accuracy, precision, and recall) built with a large number of distance measures (more the ten different distance measures) and tested on several real-world data sets, with and without adding different levels of noise. Their empirical results showed that the performance of the KNN classifier strongly depends on the distance metric of choice, and the results showed varied performances among different distances. Their study artificially generated multiple noise levels reaching 90% on 28 different datasets in a noisy data setting while controlling K values to {1, 3, √n}. One important limitation of their study is that the generated noise does not simulate the kind of noise occurring naturally in real-world data. In their conclusion, no optimal distance metric can be used for all types of data sets, and their results showed that each data set favours a specific distance metric, which complies with the no-free-lunch theorem [249].

An ongoing research question has been there for many years regarding the best choice of K. Choosing a very small or very large K could result in an unstable model. For example, with $\boldsymbol{K = 1}$ lies a risk of overfitting. In contrast, with a large K, for instance, $\boldsymbol{K = N}$, the KNN classifier could always predict the majority class label causing the issue of underfitting. The problems of overfitting and underfitting are introduced and discussed later in this chapter. Therefore, a good choice of the hyperparameter K in KNN allows for to trade-off between overfitting and underfitting.

In 2014, Hassanat et al. extensively reviewed various methods by researchers to obtain an optimal K parameter [259], comparing many attempts to solve the K parameter problem in the KNN classifier, which were proposed in different studies [260-264].

Their paper proposes an ensemble learning approach using the same Nearest Neighbour rule [259]. Basically, the traditional KNN learner is used each time with a different K, starting from k=1 to k equal to the square root of the training set, and each classifier votes for a specific class. Then a multi-classifiers system uses the majority rule to identify the class, i.e., the class with the highest number of votes (by 1-NN, 3-NN, 5-NN… √n-NN) is chosen.

Their approach [259] was applied to multiple datasets using the accuracy metric to compare their results with their predecessors' attempts and, more specifically, with Jirina and Jirina's approach [264], the inverted indexes of neighbours' classifier (IINC). From Hassanat's analyses, using k = √n did not yield excellent results compared to other methods, so using k = √n as a rule of thumb is not a good choice for the KNN classifier [259]. In addition, using a large number of neighbours, such as k= 30, 45 and 60, does not help increase the accuracy of the KNN classifier. Their experiments confirmed that choosing the optimal K is almost impossible for various reasons. The variation in KNN performance is due to the change in K and the distance metric used. Furthermore, determining K becomes more difficult when the examples are not uniformly distributed [259].

Nevertheless, from our point of view, we emphasise that the use of accuracy in their comparison could be misleading as there is no declaration of the state of class distribution imbalance within their used datasets. Usually, the K parameter in the KNN classifier is chosen empirically as an odd number to have a tiebreaker. Then, depending on each problem, different numbers of nearest neighbours are tried, and the K parameter with the best performance, as defined per the classification problem at hand, is chosen to define the classifier.

However, from our analysis of Hassanat's et al. model's accuracy tables [259] with $K = \{1,3,5,7,9\}$, in comparison with the rest of the models with K determined by various methods in their paper [259], we note that a difference of a small maximum margin of ±2% in the accuracy metric between the best performing model (for $K = \{1,3,5,7,9\}$) and all other models. From their study, it is also noticed that building KNN models with $K = \{1,3,5,7,9\}$ conquered the best accuracy performance in 18 different datasets. In comparison, their ensemble approach redeemed the best accuracy score for only seven datasets, and the competing IINC harvested a higher accuracy performance on nine datasets. The √n -NN models dominated the best performance for six datasets while choosing a large K achieved the highest accuracy performance on just five different datasets.

E) Support vector machines (SVM) were developed in the 1990s by Cortes and Vapnik [265]. It is challenging to explain SVM in an interdisciplinary context that suits various readers from multiple domains since most of the description of SVMs is either computer or data-sciences-centric. However, we recommend a look at Daumé's explanation which seems to be phased to readers from multiple disciplines [266].

SVMs can be explained by a Maximal-Margin classifier that sets up an optimisation problem that attempts to find a separating hyperplane with as large a margin as possible between class labels. The distance between the hyperplane and the closest data points is called the margin. The optimal hyperplane that can separate the two classes has the largest margin. This is called the Maximal-Margin hyperplane. The margin is calculated as the perpendicular distance from the line to only the closest points. Only these points are relevant in defining the line and constructing the classifier. These points are called support vectors. They support or determine the placement of the hyperplane [266].

Therefore, the hyperplane is learned from training data using an optimisation procedure that maximises the margin. The best choice of hyperplane is most likely to be the furthest away from the closest training points (large margin). The desire for hyperplanes with large margins is another example of an inductive bias. Normally, there could be various numbers of hyperplanes, and the data does not indicate which hyperplane is best. Thus, the choice is made using some other source of information.

Therefore, SVM sets up this problem as a constrained optimisation problem [266]:

$$\min_{\boldsymbol{\omega},\boldsymbol{b}} \frac{\mathbf{1}}{\boldsymbol{\gamma}(\boldsymbol{\omega},\boldsymbol{b})}$$

$$\boldsymbol{subj.to}\ \boldsymbol{y_n}(\boldsymbol{\omega}\,.\,\boldsymbol{x_n}+\boldsymbol{b}) \geq \mathbf{1} \qquad (\forall \boldsymbol{n})$$

To solve this optimisation problem that consists of weights (ω) and a bias (b), SVMs seek the parameters that maximise the margin, denoted $\boldsymbol{\gamma}$, subject to the constraint that all the training data points are correctly classified [266]. The optimisation problem remains unchanged, even when replacing "1" with any positive constant value. "1" is interpreted to ensure a non-trivial margin between the class labels. The main issue here is that the data is not linearly separable. This issue yields having no set of parameters ω, b that can simultaneously satisfy all the constraints. In other words, there is not a set of any possible feasible solutions to this optimisation problem (empty feasible region). Here SVM is trying to enforce a hard constraint known as *hard-margin SVM.* In order to handle inseparable data, the optimisation problem needs further modification, which is done by introducing a *slack parameter*. In the case where a point is not placed correctly across the hyperplane (misclassified), a penalty is paid to move the point to the correct side with a considerable amount of movement denoted by $\xi(x_i)$. By introducing one slack parameter per training example, and a penalty for having to use slack, the objective function represents a *soft-margin SVM* in the following [266], therefore:

$$\min_{\boldsymbol{\omega},\boldsymbol{b}} \frac{\mathbf{1}}{\boldsymbol{\gamma}(\boldsymbol{\omega},\boldsymbol{b})} + \boldsymbol{C}\sum_{\boldsymbol{n}} \boldsymbol{\xi}_{\boldsymbol{n}}$$

$$\boldsymbol{subj.to}\ \boldsymbol{y}_{\boldsymbol{n}}(\boldsymbol{\omega}\,.\boldsymbol{x}_{\boldsymbol{n}} + \boldsymbol{b}) \geq \mathbf{1} - \boldsymbol{\xi}_{\boldsymbol{n}} \qquad (\forall \boldsymbol{n})$$

$$\boldsymbol{\xi}_{\boldsymbol{n}} \geq \mathbf{0} \qquad (\forall \boldsymbol{n})$$

The above objective function aims to ensure that all points are correctly classified. But if a point n cannot be correctly classified, then the slack $\boldsymbol{x}_{\boldsymbol{n}}$ can be set to a value greater than zero to "move" it in the correct direction. However, for all non-zero slacks, a penalty is paid in the objective function proportional to the amount of slack. The hyperparameter $\boldsymbol{C} > \mathbf{0}$ controls overfitting versus underfitting. The smaller the value of C, the more sensitive the algorithm is to the training data and vice-versa. $\boldsymbol{C} = \mathbf{0}$ indicates no slack (violation of the hyperplane) using the inflexible Maximal-Margin Classifier described above. The larger the value of $\boldsymbol{C}$**,** the more violations of the hyperplane are permitted. As $\boldsymbol{C}$ affects the number of instances allowed to fall within the margin, $\boldsymbol{C}$ influences the number of support vectors used by the model [266].

The advantage of the soft-margin SVM over the hard-margin SVM is evident, where the feasible region is never empty. Therefore, there will always be some solution, regardless of whether or not the training data is linearly separable. There are multiple methods to solve the optimisation problem published by Daumé in 2017 [266].

The SVM model needs to be solved using an optimisation procedure. There are specialised optimisation procedures that re-formulate the optimisation problem to be a Quadratic Programming problem. The most popular method for fitting SVM is the Sequential Minimal Optimisation (SMO) method developed by Platt in 1998 [267] and is proven to be very efficient. Further modifications to the SMO algorithm were developed in 2001 by Keerthi et al. to perform significantly faster than the original SMO on all benchmark data sets tried [268].

The SVM algorithm is implemented in practice using a kernel. The kernel defines the similarity measure between new data and the support vectors. There are various kernel types, such as linear, polynomial and radial (radial basis function RBF) kernels [266]. The kernel selection is very important in the classification problem and depends on the data. The more complex the kernel, the better it is to separate the classes that are curved or even more complex, but the longer time it takes for training. RBF is a popular kernel in the machine learning community, while the Polynomial kernel is a non-stationary (non-monotonic) kernel well suited for problems where all the training data is normalised.

F) C4.5 decision tree (DT), a non-parametric learner developed by Ross Quinlan in 1993 [196]. The C4.5 algorithm is a Decision Tree Classifier. C4.5 decision tree algorithm has been known as a strong competitor in comparative machine learning studies [269] [270].

The two main aspects which distinguish tree models are the division criteria and the method used to reduce the dimensional growth of the tree, known as pruning. The algorithm J48 is an improved implementation of the C4.5 algorithm [271] and comes with automatic pruning via a Confidence Factor (CF) parameter to avoid overfitting. The confidence factor determines how aggressive the pruning process will be. On the one hand, the higher this value, the more 'confident' you are that the learnt data is a good representation of all possible events, and therefore less pruning will occur [271].

On the other hand, smaller CF values induce more pruning which noticeably affects classifier learning performance [272]. J48 trees will change in depth with increased CF. J48 performs better with noise-free data. But its learning strategy may overfit the training examples with noisy data [273]. J48 has been extensively experimented with in medical research. For example, in 2012, Stiglic et al. developed a different automatic technique for pruning J48 known as Visual Tuned J48 (VTJ48) [274] to reduce the complexity of its tree model and enhance interpretability for medical experts' use. VTJ48 is a one-button decision tree without the need to tune the parameters and build multiple decision trees [274]. However, their performance evaluation in accuracy and AUC over 40 medical datasets have proven the superiority of J48 in more datasets with automatic pruning via its confidence factor parameter [274]. For the datasets where VTJ48 overtook J48 performance, it was confirmed that there is no statistically significant difference in the predictive performance between the decision trees built by J48 and VTJ48 [274].

G) Random forests (RF) is a generalisation of standard decision trees. In 1990, Kwot and Carter proposed that the average multiple decision trees of varied structures based on randomisation give better classification results than a single tree [275]. In 1994, Breiman proposed another generalisation of standard decision trees [276] based on Bootstrap Aggregation (BAgging) from a single training set or random unpruned decision trees. Bootstrap is a statistical method for estimating a quantity from a data sample. BAgging is a simple ensemble method. Ensembles are methods which combine the predictions models from multiple machine learning algorithms together to make more accurate predictions compared to an individual model. As mentioned previously, decision trees have high variance and are sensitive to the specific training data; if the latter is altered, the model can be quite different. Hence BAgging is applied to such algorithms [266].

When BAgging decision trees, the attention is shifted away from the problem of overfitting; the individual trees are fully grown without pruning and have fewer observations classified at each leaf node of the tree. The trees are sub-models used to combine predictions. Increasing the number of trees increases the number of samples. Therefore, the number of trees is increased at each run until no improvement in classification accuracy is observed [266].

When looking back at the decision trees, decision trees choose a variable to split instances using entropy/purity algorithms such as information gain and gain ratio. Such algorithms are known as greedy algorithms. Although they make an optimal choice of variables for the splitting criteria, their choice does not necessarily result in an optimal solution (classification models). Theoretically, when BAgging decision trees, the decision trees can be of similar structure, resulting in highly correlated (parallel) predictions. However, in Random Forests, the models of the learnt subtrees vary, so the resulting predictions are less correlated. This process is controlled by choosing $\boldsymbol{m}$ predictor variables from all input predictors $\boldsymbol{p}$ for each split in a new bootstrapped sample. [266]

Random Forests algorithm takes three arguments: the $\boldsymbol{m}$ number of features competing in a node, a desired depth of the decision trees $\boldsymbol{d}$, and a number $\boldsymbol{K}$ of total decision trees to build. The algorithm generates each of the $\boldsymbol{K}$ trees independently, which makes it very easy to generalise. For each tree, it constructs a full binary tree of depth $\boldsymbol{d}$. The features used at the branches of this tree are selected randomly, typically with replacement, meaning that the same feature can appear multiple times, even in one branch. The leaves of this tree, where predictions are made, are filled in based on the training data. This last step is the only point at which the training data is used. The resulting classifier is then just a vote of the $\boldsymbol{K}$ many random trees [266]. Biau and Scornet's survey in 2016 found that literature on tuning the Random Forest parameters seems rare and an ongoing research topic [277]. They highlighted that tuning the Random Forest parameters becomes a computational burden, particularly for large data sets with hundreds and thousands of samples and variables [277].

To circumvent Random Forests' computational intensity issue in larger datasets, Schwarz et al. implemented a fast version of the original random forest algorithm, Random Jungle [278]. In terms of impact on performance, Bierman argues that increasing the parameter $\boldsymbol{K}$ would increase the Random Forests model's prediction performance and won't result in an overfitting scenario [276]. Therefore, the computational cost for inducing a forest increases with parameter ***K***. Thus, a trade-off is required between the computational complexity and the model's accuracy when selecting ***K***.

On the contrary, Díaz-Uriarte and De Andres in 2006 [279] argued that the value of ***K*** is irrelevant (large enough) through their experiments with a prediction problem involving microarray data sets to classify patients according to their genetic profiles [279]. Also, both investigated the effect of ***m*** thoroughly to show that this parameter has little impact on the performance, although larger values may be associated with a reduction in the predictive performance [279].

Genuer et al. recommend ***m*** to be as large as possible, limited by the available computing resources [280]. In terms of parameter ***d,*** Random Forests are fully grown in most applications; the impact of tree depth on the Random Forest model's performance is still an open question [277].

Random forests can rank the importance of variables in a model. The ranking is produced via two measures. The first measure was introduced by Breiman in 2001 and called Mean Decrease Accuracy (MDA) [276]. It originates from the idea that if the variable is not important, then rearranging its values should not degrade prediction accuracy. Breiman also proposed the second measure in 2003, Mean Decrease Impurity (MDI), calculated based on the total decrease in node impurity from splitting on the variable, averaged over all trees [281]. However, various empirical studies point out that MDA and MDI behave poorly when correlation increases [282 – 284]. For example, Genuer et al. observed that MDA can detect the most relevant variables less when the number of correlated features increases [280].

Scientists hinted that Random Forests mimic the effect of *Deep Artificial Neural Network* architectures [285] [286]. However, the present empirical results remain insufficient to explain in full generality the behaviour of random forests. Scientists' intuition leans toward the idea that tree aggregation models can estimate more complex patterns than classical ones. These complex patterns, which are beyond the reach of classical methods, are still to be discovered, quantified, and mathematically described [277].

H) Logistic model trees (LMT) is a model proposed by Landwehr et al. in 2005 with a tree structure but with logistic regression (LR) functions at the leaves [287]. One of the Logistic Regression drawbacks is its limited ability to resolve non-linear problems. Thus, the prediction of non-linear relationships could be improved by incorporating a decision tree into a regression model [288]. LMT structure is made of a set of non-terminal nodes and a set of leaves (terminal nodes). In one implementation of LMT models by Szymanski [288], the output of the node's decision tree is transformed into a categorical variable and then deployed into a logistic regression by transforming each of the categories (nodes) into dummy variables.

Additionally, LMT overcomes the problem of the model's interpretability, multicollinearity [289] and Simpson's Paradox [290] in classification tasks. Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations [290]. As the sample size or the number of predictors increases, so does model complexity in prediction models. However, a more complex model is always harder to interpret, while an overly simple model may perform poorly. This is a known concern in collaborative data science and medical research. Therefore, LMT offers a way to simultaneously retain the visual interpretability of simple models and the predictive power of denser ones.

The LMT model has two components, a binary tree structure showing the data partitions and a set of simple linear logistic models, each fitted to each partition. Hence this division of model complexity makes the model easy to interpret [287].

In addition to the advantage of interpretability, combining a tree structure and logistic regression models in a single tree produces class probability estimates and classification labels. Landwehr's team [287] compared the LMT classification accuracy performance against the performance of the decision tree learner C4.5 and the Logistic Regression on 32 UCI datasets [291]. Their experiments showed that the LMT algorithm achieved higher average accuracy than the traditional C4.5 and Logistic Regression models.

3.6 Processing ordinal class labels

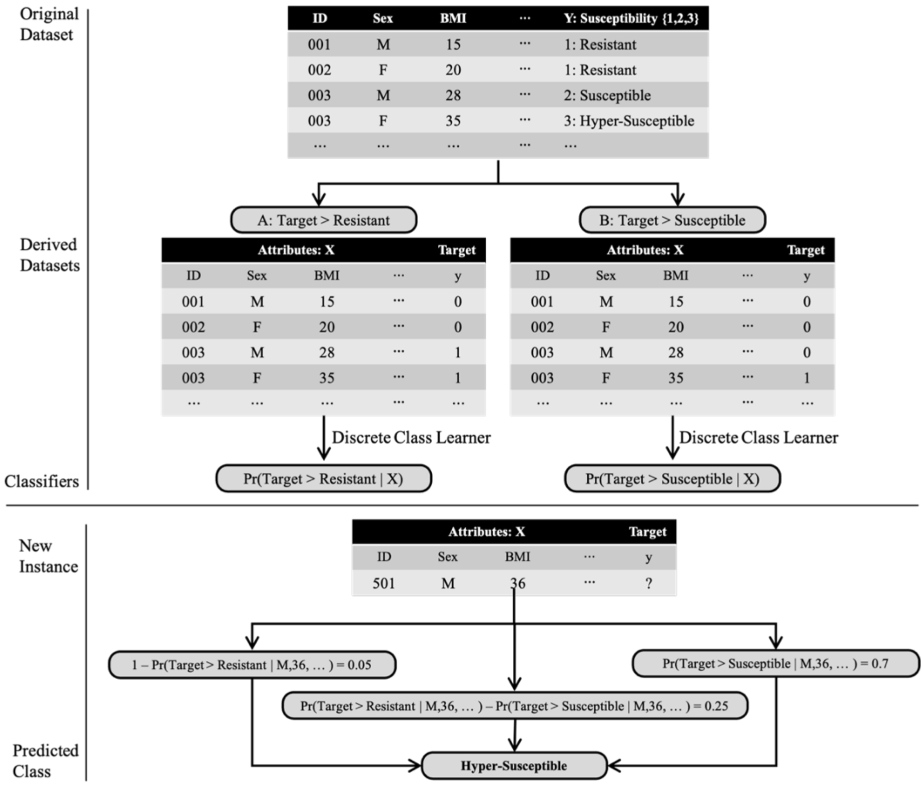
In Chapter 2, we presented Steven's early topology to measurement scales. We also reviewed multiple methods used to classify scales of measurement. We stated some critical points: ordinal labels could be handled as nominal-type measurements. And the classification of measurement scales relies on the used prediction modelling tool (the implementation of the algorithms) and the arithmetic significance of the attributes' values.

Standard classification algorithms implementations cannot use the ordering information in the class variable (ordinal class) because they treat each class label in the class attribute as a set of unordered values [292]; hence the choice of nominal type is preferred. An ordinal class variable's value exists on an arbitrary scale where only the relative ordering between different values is significant. However, there may be a ranking among nominal-class labels driven only by the domain experts' interest in nominal-class labels.

When considering susceptibility to adverse health events labelling, in this thesis, we see how important it is to distinguish patients whose characteristics are deemed to put them at higher risk (hyper-susceptible) of developing a chronic disease. People at moderate risk (susceptible) come second in their importance to be discovered. At the same time, healthy individuals (resistant) are of the slightest worry.

Treating this problem as a classical classification problem loses the ordering preference of each class because a standard classifier treats each label as a set of unordered values. Current suggestions in the field are to treat such a problem as a regression problem [293], knowns as ordinal class regression. This basic approach assumes that the scale of intervals is known. Therefore, it replaces the ordinal labels with their original range as actual values and fits a regression algorithm.

While the ordinal class regression approach seems intuitive, it does not assist in class labels that are not represented by any numeric interval or ratio. In 2001, a more sophisticated method was proposed by Frank and Hall [292]. Their approach relies on decomposing such an ordinal classification problem of $\boldsymbol{k}$ class labels into $\boldsymbol{k-1}$ binary classification problems, as long as the classifier can estimate output class probability, citing the C4.5-ORD algorithm as an example in their paper. However, in the same paper [292], the evaluation of their approach on multiple datasets was assessed on the accuracy metric. It is not apparent how to deal with the presence of a rare group of labels (minority class) in the original non-binarised dataset, whose probability possibly diminished if classified in a binary model on its own against all other labels as one label [293]. Figure 3.7 demonstrates Frank and Hall's [292] handling of ordinal class labels.


Original Dataset
Derived Datasets
Classifiers
New Instance
Predicted Class
A: Target > Resistant
B: Target > Susceptible
Discrete Class Learner
Discrete Class Learner
Pr(Target > Resistant | X)
Pr(Target > Susceptible | X)
1 – Pr(Target> Resistant | M,36, …) = 0.05
Pr(Target> Susceptible | M,36, …) = 0.7
Pr(Target> Resistant | M,36, …) – Pr(Target> Susceptible | M,36, …) = 0.25
Hyper-Susceptible

ID	Sex	BMI	···	Y: Susceptibility {1,2,3}
001	M	15	···	1: Resistant
002	F	20	···	1: Resistant
003	M	28	···	2: Susceptible
003	F	35	···	3: Hyper-Susceptible
···	···	···	···	···

Attributes: X				Target
ID	Sex	BMI	···	y
001	M	15	···	0
002	F	20	···	0
003	M	28	···	1
003	F	35	···	1
···	···	···	···	···

Attributes: X				Target
ID	Sex	BMI	···	y
001	M	15	···	0
002	F	20	···	0
003	M	28	···	0
003	F	35	···	1
···	···	···	···	···

Attributes: X				Target
ID	Sex	BMI	···	y
501	M	36	···	?

Fig. 3.7 Illustration of Frank and Hall's ordinal class classification approach

In Figure 3.7, if there are class labels with ordinal value susceptibility from 1 to 3. The ordinal target labels can be transformed into two binary classification problems, so after training the two binary classifiers, the probabilities of the ordinal values can be predicted [293],

$$\boldsymbol{Pr(y=1), Pr(y=2)\ and\ Pr(y=3)}.$$

To predict the class value of an unseen data point, the probabilities of the $\boldsymbol{k}$ original ordinal classes are estimated using our $\boldsymbol{k-1}$ binary classification models. Estimating the probability for the first and last ordinal class value depends on a single classifier. The probability of the first ordinal value (Resistant) is given by [293]:

$$\boldsymbol{Pr(y=1) = 1 - Pr(Target > Resistant)}$$

And the last ordinal value (Hyper-Susceptible) is computed from [293]:

$$\boldsymbol{Pr(y=3) = Pr(Target > Susceptible)}.$$

For class values in the middle of the range, in our case, there is only one (Susceptible label); the probability depends on a pair of classifiers. In this example, it is given by [293]:

$$\boldsymbol{Pr(y = 2) = Pr(Target > Resistant) - Pr(Target > Susceptible)}.$$

When predicting an unknown class for a new instance, the prediction is given by each of the $\boldsymbol{k-1}$ classifiers, and the class with maximum probability is assigned to the new instance. Besides the previous problem of diminished minority label effect, this approach increases the computation time since building a single model means building $\boldsymbol{k-1}$ multiple models [293].

Another approach, known as cost-sensitive classification, can shift the classifier's attention to the class of interest by adjusting the misclassification penalties for these classes [294]. However, in some domains, including health research, where the cost is not monetary, it becomes harder to estimate the misclassification cost.

3.7 Cost-sensitive learning

A highly skewed class distribution often biases the machine learning model produced by training the algorithm to classify the majority class much more accurately than the minority class [146]. This is a consequence that most classifiers maximise accuracy by design [146]. However, in many domains, such as medical screening, the minority class is the class of primary interest. Hence, this classification modelling behaviour is unacceptable nor efficient for making predictions. Cost-sensitive learning is one of the approaches to tackle imbalanced learning in datasets alongside other data resampling approaches mentioned in the previous chapter, Chapter 2.

Cost-Sensitive learning was proposed by Zadrozny and Elkan in 2001 [294]. It is a type of learning in data mining that considers misclassification costs (and possibly other types of costs). The aim of this subtype of ML learning is to minimise the total cost. The key difference between cost-sensitive learning and conventional classification learning is that cost-sensitive learning treats the different misclassifications differently. On the other hand, standard (cost-insensitive) learners do not account for the misclassification costs [294].

Class-imbalanced datasets occur in many real-world applications. Often the minority class is a very small portion of the dataset. Suppose traditional (cost-insensitive) classifiers are applied to most of these datasets. In that case, they are likely to predict most positive labels as negative (where positives are a minority), which is often regarded as a problem in learning [149]. In 2000, Provost listed two assumptions on which classifiers build as the root cause of this learning problem [295]. The first is that the goal of the classifiers is to maximise the accuracy (or minimise the error rate), and the second is that the class distribution of the training and test datasets is the same.

Therefore, under these two assumptions, predicting everything as negative for a highly imbalanced dataset is often the right thing to do. Based on the previous assumption, the problem of class imbalance manifests in the classifier's learning from imbalanced datasets if the cost of different types of error (misclassification, i.e., FP and FN) is not the same or if the class distribution in the test data is different from that of the training data. Therefore, cost-sensitive learning is often used to mitigate modelling performances for datasets with highly imbalanced class distribution [296].

For a binary classification (i.e., positive and negative class), in cost-sensitive learning, the costs of false positive (actual negative but predicted as positive; denoted as FP), false negative (FN), true positive (TP) and true negative (TN) can be given in the following binary cost matrix [294] (Figure 3.8),

		Actual	
		Negative (N or 0)	Positive (P or 1)
Predicted	Negative (N or 0)	$C(0,0), or\ TN$	$C(0,1), or\ FN$
	Positive (P or 1)	$C(1,0), or\ FP$	$C(1,1), or\ TP$

Fig. 3.8 Elkan's cost-sensitive matrix

The notation $\boldsymbol{C(i,j)}$ represents the misclassification cost of classifying an example from its actual class $\boldsymbol{j}$ into the predicted class $\boldsymbol{i}$. The negative and positive instances are encoded with 0 and 1, respectively [294]. Domain experts may be able to give these misclassification cost values. It is usually assumed that such a cost matrix is given and known in cost-sensitive learning. For multiple classes (i.e., a susceptibility matrix), the cost matrix can be easily extended by adding more rows and columns, as per Figure 3.9.

		Actual		
		Resistant (R or 0)	Susceptible (S or 1)	Hyper Susceptible (H or 2)
Predicted	Resistant (R or 0)	$C(0,0), or\ TR$	$C(0,1), or\ FR$	$C(0,2), or\ FR$
	Susceptible (S or 1)	$C(1,0), or\ FS$	$C(1,1), or\ TS$	$C(1,2), or\ FS$
	Hyper Susceptible (H or 2)	$C(2,0), or\ FH$	$C(2,1), or\ FH$	$C(2,2), or\ TH$

Fig. 3.9 Extended multi-class cost-sensitive matrix

Note that $\boldsymbol{C(i,i)}$ (TP and TN) or (TR, TS and TH) are usually regarded as the “benefit” (i.e., negated cost) representing a correctly classified instance. Usually, the minority class is viewed as the positive class, and it is often more expensive to misclassify an actual positive example into a negative than an actual negative example into a positive.

In healthcare research, and depending on the classification problem, on one hand, classifying a person with a disease otherwise may result in treatment exclusion worsening the person’s health, a wider spread of infectious disease, and impact on quality of life (QoL) or death. On the other hand, classifying a Resistant (Healthy) person otherwise may result in unnecessary interventions taken or additional clinical screening tests. In some cases, interventions may be insignificant such as unnecessary adjustments made to lifestyle (i.e., diet and exercise). In other cases, they could become significantly expensive but unharmful such as vaccination, or catastrophic by receiving unsuitable or unnecessary treatment impacting their QoL. The effect of making a classifier cost-sensitive can be illustrated in Figure 3.10; the classification of minority red labels increased due to incorporating a cost matrix to make the algorithms cost-sensitive.

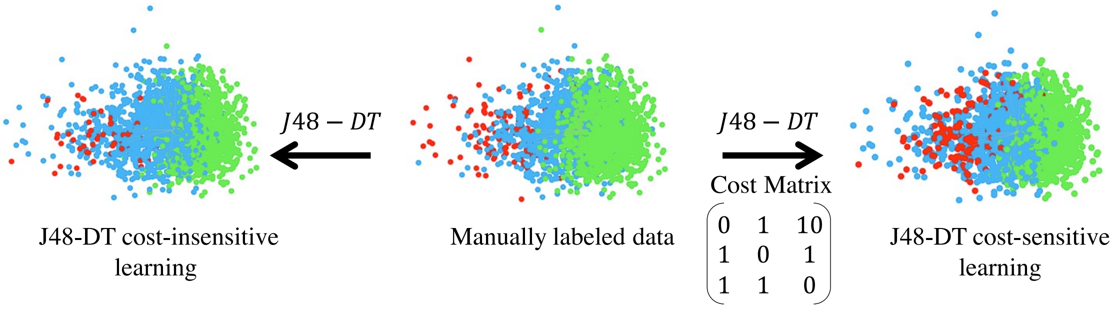


Fig. 3.10 Illustration of cost-sensitive learning effect on classifier behaviour

Given a binary cost matrix (Figure 3.8), an example should be classified into the class with the minimum expected cost. The expected cost $\boldsymbol{R(i|x)}$ of classifying an instance x into class i (by a classifier) can be given by [297]:

$$\boldsymbol{R(i|x) = \sum_{j} P(j|x)C(i,j)}$$

where $\boldsymbol{P(j|x)}$ is the probability estimation of classifying an instance into class $\boldsymbol{j}$. That is, the classifier will classify an instance $\boldsymbol{x}$ into the positive class if and only if [297]:

$$\boldsymbol{P(0|x)C(1,0) \; + \; P(1|x)C(1,1) \; \leq \; P(0|x)C(0,0) \; + \; P(1|x)C(0,1)}$$

This is equivalent to the following [297]:

$$\boldsymbol{P(0|x)(C(1,0) - C(0,0)) \; \leq \; P(1|x)(C(0,1) - C(1,1))}$$

The decision to classify an example as positive will not be changed if a constant is added to a column of the original cost matrix [297].

Therefore, the original cost matrix can always be converted to a simpler one [297] by replacing $\boldsymbol{C(0,0)}$ from the first column and $\boldsymbol{C(1,1)}$ from the second column with **zero**. After such conversion, a simplified cost matrix is shown in Figure 3.11.

		Actual	
		Negative (N or 0)	Positive (P or 1)
Predicted	Negative (N or 0)	0	$C(0{,}1) - C(1{,}1)$
	Positive (P or 1)	$C(1{,}0) - C(0{,}0)$	0

Fig. 3.11 Simplified version of Elkan's cost matrix

Thus, any given cost-matrix can be converted to one with $C(0{,}0) = C(1{,}1) = 0$. Under the assumption that $C(0{,}0) = C(1{,}1) = 0$, the classifier will classify an instance x into positive class if and only if:

$$\boldsymbol{P(0|x)C(1,0) \; \leq \; P(1|x)C(0,1)}$$

As $P(0|x) = 1 - P(1|x)$, a threshold p^* can be obtained for the classifier to classify an instance x into positive if $P(1|x) \geq p^*$, where [297]:

$$\boldsymbol{p^*} = \frac{\boldsymbol{C(1,0)}}{\boldsymbol{C(1,0) + C(0,1)}} = \frac{\boldsymbol{FP}}{\boldsymbol{FP + FN}}$$

Traditional cost-insensitive classifiers are designed to predict the class with a default fixed threshold of $\boldsymbol{p^* = 0.5}$. A classifier can become cost-sensitive by simply choosing the classification threshold $\boldsymbol{p^*}$and classifying any example to be positive whenever $\boldsymbol{P(1|x) \geq p^*}$. This is what several cost-sensitive meta-learning algorithms do in the form of Relabelling. Cost-sensitive meta-learning transforms existing cost-insensitive classifiers into cost-sensitive ones without modifying the base learner. Thus, it can be regarded as a middleware component that pre-processes the training data or post-processes the output from the cost-insensitive learning algorithms [297].

In 2005, McCarthy et al. compared strategies for dealing with data of a skewed class distribution and non-uniform misclassification costs [298]. Their study compared cost-sensitive learning to oversampling and undersampling strategies by modelling 14 different datasets with a decision tree learner. The minority class ratio ranged from 4% to 50%. Model building was validated with testing on a holdout sample with a 75:25 training-test split. Models' performances were compared based on the total cost metric given by [298],

$$\boldsymbol{Total\ Cost} = (\boldsymbol{FN} \times \boldsymbol{C_{FN}}) + (\boldsymbol{FP} \times \boldsymbol{C_{FP}})$$

In their conclusion, there was no clear winner for maximising classifier performance when cost information was known. However, on larger data sets with more than 10,000 total instances, the cost-sensitive learning often outperforms the sampling techniques, although it is still inconsistent for every case [298].

3.8 Model fitting, overfitting and underfitting

Earlier in this chapter, we described some machine learning models' data fitting, such as Naïve Bayes and Logistic Regression. Machine learning algorithms are optimisation methods at their core. They all depend on determining a loss (cost) function to minimise. In essence, model fitting measures how well a model generalises to similar unknown data to that it was fitted (trained). The process of model-fitting data points is a trade-off between the classifier's bias and variance.

A well-fitted model produces more accurate responses. An overfitted model matches the actual training data too closely. In contrast, an under-fitted one doesn't match closely enough. Improper fitting of a model to the training sample results in outcomes that are not accurate enough to be useful for practical applications [299]. Therefore, here the two essential parts of model fitting are explained from the literature:

A) Loss functions optimisation. Classification methods use the structure in the data to predict a label, and optimisation methods assist in using the best structure found (patterns) within the data [300].

The data does not precisely fit a trend in any real-world process, whether natural or artificial. There is always noise or other variables in the relationship we cannot measure. Thus, the classifier uses the input features to build a mathematical function, mapping the input features vectors from a given sample, usually the training set to the response output labels. The classification process (label predictions) is the output of mathematical functions, $\boldsymbol{f}(\boldsymbol{X})$, with $\boldsymbol{X}$ are the input features set. $\boldsymbol{f}(\boldsymbol{X})$ has internal parameters that help map $\boldsymbol{X}$ to the predicted labels. Model fitting is an automatic process that ensures that machine learning models have the individual parameters best suited to solve your specific real-world business problem with high accuracy [300].

There are two types of parameters; learnable parameters, which the algorithms estimate (learn) on their own during the training phase for a given dataset, and hyper-parameters, whose specific values are assigned by the data scientists to control the methodology the algorithms learn and also to tune the model's performance.

The optimisation processes engage with the parameters of $\boldsymbol{f}(\boldsymbol{X})$ and try to minimise the difference between the function output, the predicted labels, and the original response labels, given by the simplified representation of loss [300]:

$$(\boldsymbol{Y})\colon \sum(\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{X}))^{2}.$$

The minimisation of loss to improve the model's accuracy can be achieved by searching for the optimal parameter values an ideal value of the loss function is zero. While model fitting is automatic, more complicated techniques may be required to increase models' accuracy, such as hyperparameter tuning, which needs additional time, knowledge and experience [300][301].

B) The variance–bias trade-off. Overfitting and underfitting are represented by how much trade-off (flexibility) the classifier has between bias and variance. The variance refers to the classifier's attention to the data points and how much the model depends on the training data. The classifier's bias relates to the assumptions it makes about the data. An underfit model has a low variance and a high bias. Therefore, an underfit model is of low flexibility and cannot account for the data. When the model makes predictions on unseen samples, the high bias leads it to make inaccurate estimates, attributed to its failure to learn the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ [302].

Learning the training data by increasing the model's flexibility to capture every change in the data is also not the best practice. As mentioned earlier, real-world data contains noise; in this learning scenario, the model also ends up fitting the noise. Such a model is of high flexibility with high variance and low bias since it changes solely depending on the training data. The resulting predictions on the unseen test data are better than the under-fitted model. However, such a model does not learn the relationship; it essentially memorises the data points in the training sample and any embedded noise [302].

Both under-fitted and overfitted models cannot generalize well to testing data. The goal is to develop an optimal model, and there are multiple techniques for creating the optimal model: such as the cross-validation techniques mentioned earlier in this chapter. Overfitting and underfitting are fundamental problems in data-driven modelling. The model's performance evaluation on unseen test samples, discussed in the follow up sections, gives the best indication of a model being fit for purpose in real-world applications [302].

3.9 Fundamental classification evaluation metrics

These represent the building block of all classification metrics. These metrics are purely based on the raw comparison of a classification model's prediction (labels) for a given set of instances to their actual labels. Some fundamental metrics are compared in their raw formats, such as the common Confusion Matrix and the True and False Positives and Negatives. In contrast, others are interpreted arithmetically, like accuracy, or in pairs, i.e., Precision and Recall and False Omission and False Discovery Rates.

A) The confusion matrix is the base evaluator of a machine-learning model [303]. It is a square $n \times n$ matrix consisting of n groups of predicted class (C) labels with corresponding n groups of actual class labels. It is structured as per Figure 3.12.

Depending on the domain problem, there might be two or more classes in the classification problem, and out of these, there may be one class of interest. This class of interest is usually labelled positive. Whilst and the label of the rest of class/es observations are labelled negative**.** Examples of positive classes in the medical have many examples depending on the medical intervention objectives. For example, in a multi-class classification problem, a triple negative cancer tumour requires critical medical attention to be distinguished compared to the other types since it is more aggressive and lethal. Hence, called the positive class label (primary class of interest), while the rest of the classes are secondary (Negatives). But also, there are other cases in a multi-class classification with ordinal importance for each group. For example, in the case of a three-category susceptibility classification, those deemed hyper-susceptible to a disease, or a health condition, reserve the highest priority to be distinguished, followed by those who are susceptible, while predicting the resistant subjects might be the least critical task mortality-wise [304]. Some researchers may condense the multi-class classification labels into a binary classification task, a one-class learning or an ordinal class learning to focus on the class of particular interest (i.e., all types of cancer tumours as the negative class "non-triple negative" vs "triple negative" as the positive class) [305].

		Actual Class Labels			
		C_1	C_2	$C_{\dots}$	C_n
Predicted Class Labels	C_1	*True* C_1	*False* C_1	*False* C_1	*False* C_1
	C_2	*False* C_2	*True* C_2	*False* C_2	*False* C_2
	$C_{\vdots}$	*False* $C_{\dots}$	*False* $C_{\dots}$	*True* $C_{\dots}$	*False* $C_{\dots}$
	C_n	*False* C_n	*False* C_n	*False* C_n	*True* C_n

Fig. 3.12 Multi-class confusion matrix structure

Nevertheless, there are standalone binary classification problems, such as predicting patients' reaction to treatment (Susceptibility to side effects) which could deteriorate the patient's quality of life or even result in death [151]. The confusion matrix in Figure 3.13 illustrates the positives and negatives where the class label C_5 is deemed to be the label of interest.

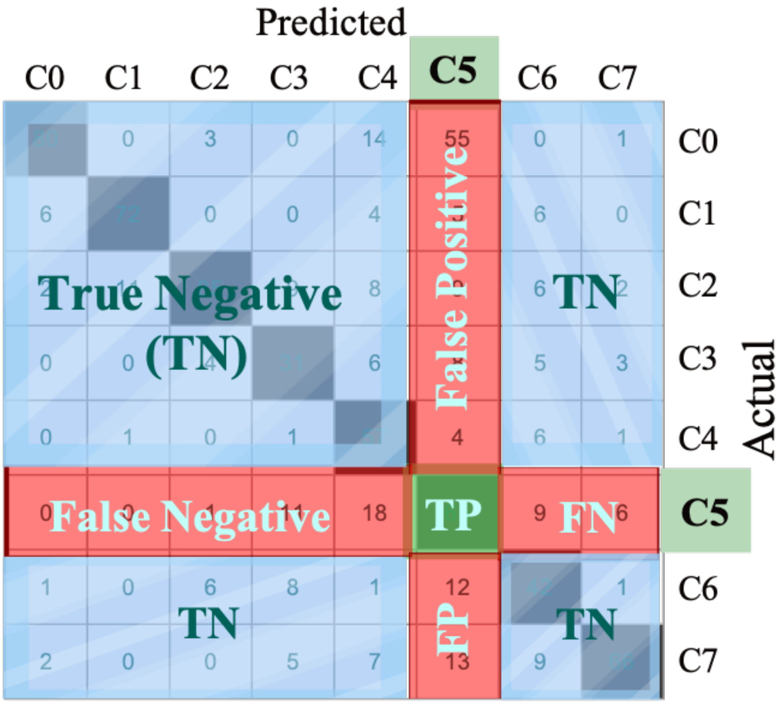


Fig. 3.13 Multi-class confusion matrix with a single class of interest (C5)

The classification confusion matrix analyses how well the model can distinguish various class labels. In addition, the confusion matrix elements TP, TN, FP and FN are used to derive other performance measurements.

B) The classifier's accuracy is the most common evaluation metric for a classification model. It assesses the overall effectiveness of the algorithm by estimating the probability of the true value of the class label [303]. The classifier's accuracy in any confusion matrix is given by:

$$\boldsymbol{Accuracy_{model}} = \frac{\boldsymbol{Correctly\ classified\ instances}}{\boldsymbol{All\ instances}}$$

The Error Rate is an estimation of misclassification probability according to model prediction given by:

$$\boldsymbol{Error\ Rate = 1 - Accuracy}$$

C) Sensitivity (Recall), also known as True Positive Rate (TPR), describes the accuracy of the predictions for the positive class (class of interest) [303].

It is considered a measure of the completeness of positive examples. Sensitivity can be calculated from a binary-class confusion matrix by:

$$\boldsymbol{Recall = TPR = \frac{TP}{TP + FN}}$$

D) Precision, also known as the Positive Predictive Value (PPV), measures how powerful the model is in predicting positive labels correctly out of all positive predictions, including FP [303]. Precision is also considered a measure of the correctness of positive examples predictions. The precision evaluation metric is calculated by:

$$\boldsymbol{Precisoin\ (PPV) = \frac{TP}{TP + FP}}$$

A perfect model should capture all positive instances making its Recall equal to 1, and these correct positive predictions are the only term that empowers the Precision value to become 1. This way, the model is known not to sacrifice accuracy.

E) Specificity, also known as the True Negative Rate (TNR), indicates the prediction rate within the secondary class/es outside the positives [303]. Specificity approximates the probability of the negative label/s being true; in other words, both Sensitivity and Specificity assess the algorithm's effectiveness on a single class, positive and negative, respectively.

F) False Discovery Rate (FDR) is the expected ratio of false positive classifications (false discoveries) to the total positive classifications [303]. Therefore, the addition of Precision and the FDR equals 1. Thus:

$$\boldsymbol{FDR\ =\ 1 - Precision = FP\ /\ (FP\ +\ TP)}$$

G) False Omission Rate (FOR) is the expected ratio of the number of false negative classifications (False negation or omission) to the total count of negative classifications given by:

$$\boldsymbol{FOR = 1 - NPV = FN/(FN + TN)}$$

H) Negative Predictive Value (NPV) measures how powerful the model is in predicting negative labels correctly out of all negative predictions, including FN. The sum of NPV and FOR equals 1

Accuracy places more weight on the majority labels than on minority labels, which makes it challenging for a classifier to perform well on smaller classes. Hence Accuracy becomes a misleading metric. Thus, observing a variety of the class fundamental metrics becomes essential but tedious. Therefore, researchers developed additional metric by combining some of the fundamental metrics.

3.10 Combined classification evaluation metrics

Combined metrics try to compensate for the weighting placed by accuracy on the majority labels to obtain an appropriate assessment of the classifier's performance on the minority labels too, especially in the case of imbalanced learning. Examples of combined metrics are: Balanced accuracy, F-measure, G-mean, Youden's index and others.

A) Balanced Accuracy averages the Sensitivity and Specificity, which can also be called the average accuracy obtained in either class [303]. Thus, it is given by:

$$\boldsymbol{Balanced\ Accuracy = \frac{1}{2}(Sensitivity + Specificity)}$$

If the classifier performs equally well on the minority class as well as the majority, this term implements a reduction to the typical accuracy measure. Thus, if the typical accuracy is high only because the classifier is taking advantage of the correct predictions of the majority labels, the balanced accuracy suffers a decayed performance [151].

C) The Geometric Mean (G-Mean) is another way to account for each class's accuracy within the classification model. Its origin is traced back to 1997, by Kubat and Matwin, when addressing learning from imbalanced classes [305].

The G-mean is the product of the prediction accuracies, Sensitivity, the accuracy on the positive labels, and Specificity, the accuracy on the negative instances. This metric offers a balanced assessment of the model's performances on both majority and minority groups. A poor performance in predicting either the positives or negatives results in a low G-mean value. The G-Mean is calculated from [303]:

$$\boldsymbol{G - Mean = \sqrt{Sensitivity \times Specificity}}$$

The G-mean is popular, and researchers often use it for assessing classification tasks over imbalanced data [307] [308] [151].

D) The likelihood ratios are two types, the positive likelihood ratio and the negative likelihood ratio. The positive likelihood ratio $\boldsymbol{\rho^+}$ represents the ratio between the probability of correctly predicting an instance as positive (TPR) and the False Positive Rate (FPR), the probability of positive prediction incorrectly. Therefore, the positive likelihood ratio is given by:

$$\boldsymbol{\rho^+ = \frac{TPR}{FPR} = \frac{TP \div (TP + FN)}{FP \div (FP + TN)} = \frac{Sensitivity}{1 - Specificity}}$$

The negative likelihood ratio $\boldsymbol{\rho^-}$ is the ratio between the probability of incorrectly predicting an example as negative (False Negative Rate or FNR) when it is actually, and the probability of making correct negative instances prediction (True Negative Rate or TNR). Thus, the negative likelihood ratio is given by:

$$\boldsymbol{\rho^- = \frac{FNR}{TNR} = \frac{FN \div (TP + FN)}{TN \div (FP + TN)} = \frac{1 - Sensitivity}{Specificity}}$$

The likelihood ratios are commonly used in medical diagnosis predictions learning from pharmaceutical datasets, as described by Bekkar et al. [303]. When interpreting both ratios, a greater positive likelihood ratio and a lower negative likelihood ratio indicate better performance on positive and negative classes, respectively. Also, there are thresholds for $\boldsymbol{\rho^+}$ to express how well the model is performing. A positive likelihood ratio of 1 indicates that the model contribution is negligible, but for $\mathbf{1}< \boldsymbol{\rho}^{+} < \mathbf{5}$, the model is considered a poor-performing model. The model is fair when the ratio increases to $\mathbf{5}\leq \boldsymbol{\rho}^{+} \leq \mathbf{10}$. A model with a positive likelihood ratio $\boldsymbol{\rho}^{+} > \mathbf{10}$, is regarded as a good-performing model [303] [309]. When comparing a pair of classifiers, A and B, based on likelihood rations, they are interpreted as follows [303] [310]:

$\boldsymbol{\rho_A^{+}} > \boldsymbol{\rho_B^{+}}$ ***and*** $\boldsymbol{\rho_A^{-}} < \boldsymbol{\rho_B^{-}}$ implies that model A is superior overall;

$\boldsymbol{\rho_A^{+}} < \boldsymbol{\rho_B^{+}}$ ***and*** $\boldsymbol{\rho_A^{-}} < \boldsymbol{\rho_B^{-}}$ indicates that model A is superior at the confirmation of negatives;

$\boldsymbol{\rho_A^{+}} > \boldsymbol{\rho_B^{+}}$ ***and*** $\boldsymbol{\rho_A^{-}} > \boldsymbol{\rho_B^{-}}$ regards model A as superior for confirming positive cases;

$\boldsymbol{\rho_A^{+}} < \boldsymbol{\rho_B^{+}}$ ***and*** $\boldsymbol{\rho_A^{-}} > \boldsymbol{\rho_B^{-}}$ Concludes that classifier A is inferior overall.

E) F-Measure is also known as the harmonic mean of Precision and Recall. A high F1 score indicates a similar precision and recall, and it is given by [310]:

$$\boldsymbol{F} = \frac{\mathbf{2}}{\frac{\mathbf{1}}{\boldsymbol{Precision}} + \frac{\mathbf{1}}{\boldsymbol{Recall}}} = \frac{\mathbf{2} \times \boldsymbol{Recall} \times \boldsymbol{Precision}}{\boldsymbol{Recall} + \boldsymbol{Precision}}$$

Precision and Recall are not always similar, and in some applications, such as medical diagnoses, higher precision is preferred over recall. This preference empowers the model in detecting more infections as much as possible to minimise cases going unnoticed, hence untreated. On the contrary, in investment banking, analysts may prefer a model for lending eligibility to have a higher precision to maximise business opportunities [310].

Unfortunately, a classifier can't always be tuned to achieve almost equally high precision and recall rates in a test. Increasing precision reduces recall and vice versa; this is known as the precision/recall trade-off [311].

The general F-measure provides insight into the inner functionality of a model than the accuracy measure. Various studies use it to evaluate classification models with imbalanced data in bioinformatics [312]. The F-measure puts classifiers with higher sensitivity under the spotlight while challenging algorithms with higher specificity [303].

However, the relative importance of Precision or Recall can be adjusted and accounted for via the coefficient $\boldsymbol{\beta}$ to produce a **Varied F-Measure** given by:

$$\boldsymbol{F_{\beta} = \frac{1+\beta^2}{\frac{1}{Precision}+\frac{\beta^2}{Recall}} = \frac{(1+\beta^2)\times Recall \times Precision}{(\beta^2 \times Recall) + Precision}}$$

$\boldsymbol{\beta}$ is a coefficient to adjust the relative importance of precision versus recall. Decreasing $\boldsymbol{\beta}$ leads to a reduction of Precision importance, while increasing $\boldsymbol{\beta}$ leads to increased importance of Recall. In the general case, $\boldsymbol{\beta}$ equals 1, where Precision and Recall are equally important. In 2008, Chawla et al. suggested exploiting $\boldsymbol{\beta}$ in the context of imbalanced learning [313] by adapting the classical F-measure to the cost-sensitive learning methods developed by Zadrozny and Elkan [294]. From Figure 3.14, in the binary classification cost matrix, 1 indicates a positive class of interest and 0 otherwise. Thus, $\boldsymbol{\beta}$ can be defined based on the binary cost matrix as follows [303]:

$$\beta = \frac{C(0,1)}{C(1,0)}$$

		Actual	
		-ve (N or 0)	+ve (P or 1)
Predicted	-ve (N or 0)	$C(0,0), or\ TN$	$C(0,1), or\ FN$
	+ve (P or 1)	$C(1,0), or\ FP$	$C(1,1), or\ TP$

Fig. 3.14 β evaluation in Elkan's cost-sensitive binary-classification cost matrix

Where $C(1,0)$ is the cost associated with predicting a False Negative sample, $C(0,1)$ is the cost associated with predicting a False Positive instance. In the imbalanced-learning context, the rationale behind the $\boldsymbol{\beta}$-varied F-measure is that the misclassification cost within minorities is often higher than that of majority instances. Consequently, improving the recall profoundly affects the F-measure more than the precision. Therefore, the β-varied F-Measure is more appropriate in the context of imbalanced learning, and more specifically, it is difficult to implement outside of cost-sensitive learning. Hence, the general version of F-measure ($\beta = 1$) is often used [303]. Another difficulty of using β-varied F-Measure emerges in cost-sensitive learning, where classifiers' prediction performances of different algorithms tend to respond differently to the same cost matrix. This leads to the belief that the cost of misclassification cannot be a unique single value but may vary depending on the used algorithm.

F) Discriminant Power (DP) is another metric used in the context of imbalanced learning proposed by Blakeley & Oddone in 1995. The DP measure summarises the True Positive and True Negative rates (TPR & TNR) given by:

$$\boldsymbol{DP = \frac{\sqrt{3}}{\pi}(logX + logY) = \frac{\sqrt{3}}{\pi}(logXY) = \frac{\sqrt{3}}{\pi}log\left(\frac{TPR \times TNR}{FNR \times FPR}\right)}$$

$$\boldsymbol{DP = \frac{\sqrt{3}}{\pi}(logTPR + logTNR - logFNR - logFPR)}$$

$$\textbf{Where}\ \begin{cases} \boldsymbol{X = Sensitivity/(1 - Sensitivity) = TPR/FNR} \\ \boldsymbol{Y = Specificity/(1 - Specificity) = TNR/FPR} \end{cases}$$

From the above expression, the DP measure through the positive correlation with the product of sensitivity (TPR) and specificity (TNR) reinforces correctly classified instances in both classes and imposes a loss of the overall model's performance for the misclassifications in both groups. The DP metric does not have lower or upper limits. Nevertheless, it indicates how well a classifier distinguishes between positive and negative instances following a range of values. The performance of a classification model with a DP below

1 is considered “Poor”. The performance is considered “Limited” if the DP metric calculation is between 1 and 2. For a range of DP values between 2 and 3, the model performance is regarded as “Fair”, and any DP value above 3 indicates a good classification performance [303].

G) Youden Index was proposed by Youden in 1950. It is a measure to evaluate the classifier’s avoidance of failure by equally weighting the model’s performance on positive and negative examples [314]. It is derived from the fundamental two measures, sensitivity and specificity, but also linearly relates to the Balanced Accuracy matric [303]. Youden’s index $\boldsymbol{\gamma}$ is given by:

$$\boldsymbol{\gamma = Sensitivity - (1 - Specificity) = (2 \times Balanced\ Accuracy) - 1}$$

The higher the value of γ, the better the classifier’s ability to avoid failure. Youden’s index was used to evaluate diagnostic models [151] [309]. Youden Index can also be interpreted in relation to the likelihood ratio by:

$$\boldsymbol{\gamma = \frac{(\rho_-^{-1} - 1)(\rho_+ - 1)}{(\rho_+ \times \rho_-^{-1}) - 1}}$$

The above expression shows that $\boldsymbol{\gamma}$ favours classifiers with higher $\boldsymbol{\rho}^+$and lower $\boldsymbol{\rho}^-$ which explains its power to account for the model’s performance in the context of diagnosing positive cases in health research.

H) Matthew’s Correlation Coefficient (MCC), also knowns as Phi (Φ) coefficient, was invented by Karl Pearson in 1912 and introduced as a performance measure for binary classification in machine learning by Brian Matthews in 1975. The MCC is the one metric considered to offer a less biased performance and is not affected by the problem of imbalanced learning. The MCC considers accuracies and error rates on both classes mutually and involves all values in a confusion matrix [303]. MCC is given by:

$$\boldsymbol{MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}}$$

The MCC values range from +1 for a perfect prediction to –1 for the worst possible prediction. An MCC value close to 0 indicates a model that performs its predictions randomly. Also, it can be simplified further, showing that it incorporates all the fundamental metrics. Therefore, the MCC can be denoted by:

$$\boldsymbol{MCC = \sqrt{PPV \times TPR \times TNR \times NPV} - \sqrt{FDR \times FNR \times FPR \times FOR}}$$

When we look at the above MCC previous formula, the acclaimed balanced nature of performance evaluation of the MCC may be allotted to combining symmetric fundamental evaluation metrics components and the difference between two entirely opposite terms. The first term combines all the measurements that express how powerful the model is in making positive and negative predictions. The second term resembles a loss or a penalty by combining the terms representing the model's power in making inverted predictions. Therefore, some researchers consider the MCC the best single-balanced evaluation metric for classification [315].

3.11 Graphical classification evaluation metrics

These metrics are useful in illustrating specific pairs of performance measures for classifiers. Multiple graphical metrics can be derived depending on pairs of performance metrics, such as the Receiver Operating Characteristic (ROC) Curve and the Precision-Recall Curve (PRC). A single computation for performance evaluation can then be extracted from the graphical representation of the classifier’s performance to summarise the modelling performance.

A) Receiver Operating Characteristic (ROC) Curve is a plot that gives the True Positive Rate (TPR=Sensitivity) as a function of the False Positive Rate (FPR=1–Specificity) for the same group. Provost and Fawcett proposed the ROC curve in 1997 as an alternative performance metric to the accuracy measure to combat misleading accuracy performance in imbalanced learning [316]. Since then, the ROC curves have become widely used in research. A ROC curve plots the classification results from the most positive to the most negative classification. The ROC approach involves representing the sensitivity value as a function of (1-specificity) for all possible threshold values and joining the points with a curve. The more inclined the curve is toward the upper left corner, the better the classifier can discriminate between positive and negative labels [303]. Multiple ROC curves can be represented in the same space to compare the results of different classifiers (Figure 3.15).

Due to the complex intersected curves in Figure 3.15, it is sometimes difficult to identify the top model. To solve the ROC curves' intersection situation, the optimal classifier among many is the one that attains the largest area under the ROC curve for its model.

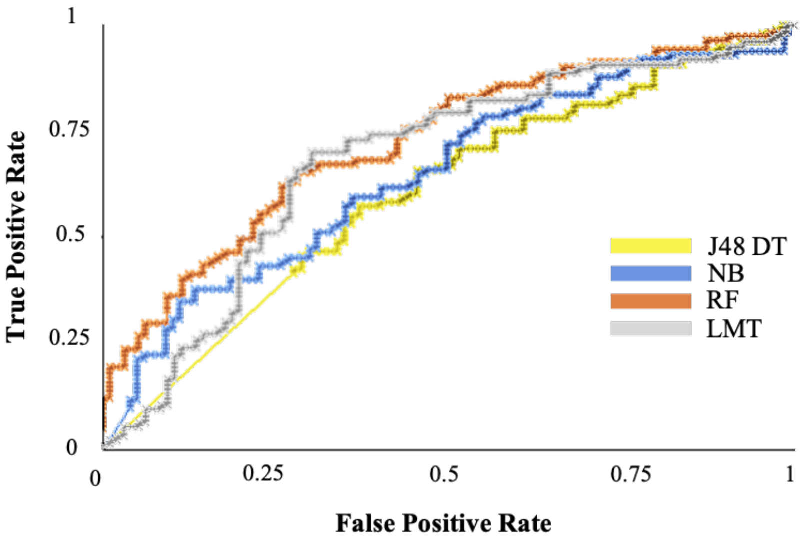


Fig. 3.15 Illustration of four classifiers' intersecting ROC curves

Despite the ROC curve's popularity in research to compare classification models, it was criticised by Drummond and Holte when evaluating models trained with imbalanced datasets and called for an alternative [317].

Furthermore, scientists like Muschelli explicitly stated that it could be potentially a misleading metric in binary classification [318]. Halligan et al. also explained concerns around the ROC metric for imbalanced learning [319] in the context of ROC curves for not considering the prior distribution of the classes and their cost.

The Area Under ROC Curve (AUC) evaluates the ROC curve evaluation. It summarises the performance of a model into a single metric between 0.5 and 1. This metric resulted from the difficulties in comparing multiple ROC curves intersecting on the same plot. The AUC sorts all classifiers by their overall performances. Hence, the reporting of AUC outweighs the ROC when reporting models' evaluation in research papers. The AUC is given by:

$$\boldsymbol{AUC = \int_{x=0}^{1} TPR\left(FPR^{-1}(x)\right) dx}$$

In practice, various models' performances are interpreted differently in relation to the AUC score, although there is no exact consensus across domains on the AUC value interpretations [303]. However, these ranges are often used as indicators of performance:

$\mathbf{0.5 \leq AUC < 0.6}$ implies a poor-performing classifier;

$\mathbf{0.6 \leq AUC < 0.7}$ indicates a fair classifier performance;

$\mathbf{0.7 \leq AUC < 0.8}$ shows a good classification evaluation;

$\mathbf{0.8 \leq AUC < 0.9}$ refers to a very good model's performance;

$\mathbf{0.9 \leq AUC \leq 1}$ demonstrates an excellent predictive model [308].

Variations of the AUC were developed to improve its suitability for imbalanced learning, one of which is Partial AUC (PAUC), proposed by McClich in 1989 [320]. PAUC estimates the AUC on a specific area of the decision threshold; thus, the PAUC can compare different models for the same benchmark of the decision threshold. Figure 3.16 shows an illustration of the ROC curve decision thresholds.

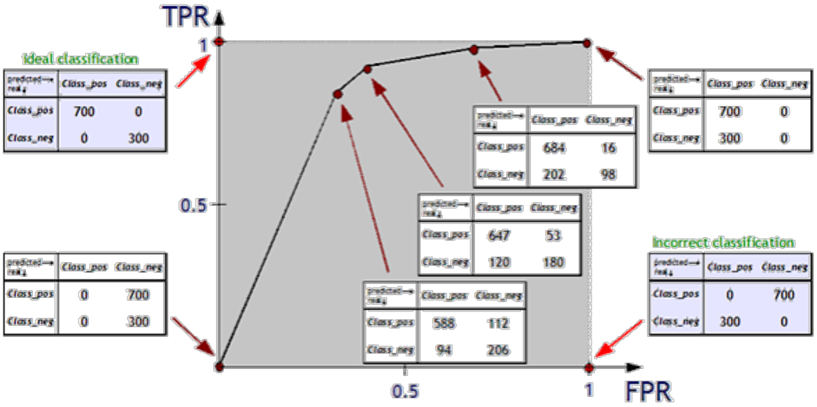


Fig. 3.16 Illustration of ROC curve decision thresholds

Another variant is the *Weighted AUC (WAUC)* was introduced by Weng and Poon in 2008 [321]. WAUC addresses the evaluation of a classifier when learning from imbalanced datasets. From Figure 3.16, the classifier that performs well in the higher TP region is preferred over the ones that do not. So instead of summing up the area under the curve with equal weights, more importance is given to the area near the top of the graph by creating a skew weight vector which distributes more weights towards the top of the ROC curve. The conventional AUC is divided into ***N*** number of areas. Thus, the WAUC is given by [321]:

$$\boldsymbol{WAUC = \sum_{i=0}^{N} area(i) \times W(i)}$$

$\boldsymbol{W(0)}$ is the weight for the bottom of the ROC curve. The new weight of an area is defined as a recursive formula using the weight of the previous area [321]. The new weights are given by [321]:

$$\boldsymbol{W(x)=} \begin{cases} \boldsymbol{\alpha} & \boldsymbol{x = 0} \\ \boldsymbol{W(x-1) \times \alpha + (1-\alpha)} & \boldsymbol{0 < x < N} \\ \boldsymbol{\dfrac{W(x-1) \times \alpha + (1-\alpha)}{1-\alpha}} & \boldsymbol{x = N} \end{cases}$$

Where $\boldsymbol{\alpha}$ is the percentage of weight to transfer to the next area towards the top. α ranges from 0, no weight transfer, to 1, a total weight transfer. When $\boldsymbol{\alpha}$ is 0, the resulting weighted AUC is equal to the conventional AUC; when $\boldsymbol{\alpha}$ is 1, only the area at the top is considered. If the cost is known, we can use the cost ratio between the positive and negative classes to set the weight transfer rate [321], $\boldsymbol{\alpha}$ by:

$$\boldsymbol{\alpha = 1 - cost\,ratio}$$

Despite Weng and Poon's modelling experiments on 20 different datasets with four different classifiers, their evaluation (score) of WAUC had so little difference compared to the conventional AUC.

Therefore, the WAUC still has the same problem as the traditional AUC when the learners have equal weighted AUC values but different ROC curves. Figure 3.17 illustrates the score (Area evaluation) similarity and shape difference of the ROC curve between WAUC and Conventional AUC for three models A, B and C. Model A and Model C have the same weighted AUC. Since the weighted AUC aims to achieve cost bias evaluation, classifiers A and B can be separated. And it is acceptable to have two different learners both perform equally well under the same cost constraint [321].

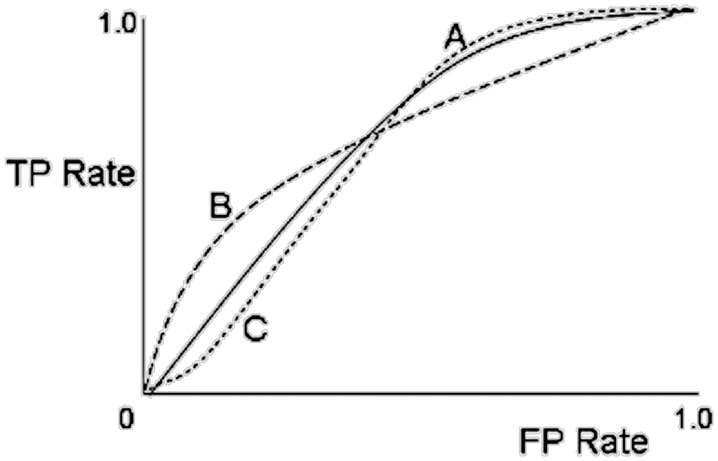


Fig. 3.17 Illustration of Weng and Poon's weighted ROC curves

Finally, when using the AUC score, one must realise that only balanced class distributions give a trustworthy ROC curve and AUC score interpretations.

However, when the effect of class imbalance is unknown whether or not to form a problem in modelling, the AUC evaluation integrity can be supported with another curve, the *Precision-Recall Curve (PRC)* [322].

B) Precision-Recall Curve metric comes in handy when suspecting that the AUC-ROC was inflated by too many correct predictions when summarising the model's performance across all possible thresholds while achieving some sort of trade-off between sensitivity and specificity [322].

The precision-recall (PRC) plot shows precision values for corresponding sensitivity (recall) values. Similar to the ROC plot, the PRC plot provides a model-wide evaluation. The area under the Precision-Recall curve (AUPRC) is a classifier performance metric considered appropriate for imbalanced learning and not dependent on model specificity. Researchers such as Saito and Rehmsmeier [322] acknowledged the AUPRC's usefulness in evaluating imbalanced class models and considered the AUPRC more informative than the ROC plot. They showed ROC plots to be misleading when applied in imbalanced classification scenarios [322]. A PRC plot can provide the viewer with an accurate classifier prediction performance since it evaluates the fraction of true positives (TPR) among all positive predictions (PPV or Precision). Their findings have potential implications for the interpretation of a large number of published studies. The impacted studies used ROC plots on imbalanced datasets for performance evaluation. Saito and Rehmsmeier [322] concluded that PRC plot is more suited for in high-throughput biological experiments that usually produce a number of large-sized datasets and the majority of such datasets are expected to be imbalanced. Saito and Rehmsmeier's experiments also highlighted other performance metrics' inability to capture good or poor classifier performance for balanced and imbalanced data samples, except only three metrics, Precision, MCC, and the ß-Varied F-Measure scores [322]. Their comparison of the PRC to other variations of ROC curve plots demonstrated that the ROC curve variations remained unchanged in contrast with the PRC plots between the balanced and imbalanced datasets; accordingly, AUPRC scores also changed [322].

C) Crossover Error Rate (CER), also known as Equal Error Rate (EER), is one of three performance metrics that are used in the information security technology industry in assessing the performance (accuracy) of biometric devices [323]. The usability of a biometric device is measured by its False Acceptance Rate (FAR), which measures the Permeability of the algorithm to attacks, False Rejection Rate (FRR), which measures the resistance of the algorithm to accept a legitimate user and the CER, which is the point of interception of the FAR curve with the FRR curves. The CER indicates the usability level of the technology. See Figure 3.18 for an illustration of CER.

Magalhães et al. state that as an algorithm gets more demanding, its FAR metric gets lower, and its FRR gets higher [323]. Usually, the system administrator can define a threshold and determine an acceptable average of FAR and FRR for the applied algorithm, according to the need for security. The process depends on the risk evaluation and the value of the protected asset. Also, in theory, the threshold can be defined by an Intrusion Detection System (IDS) [323].

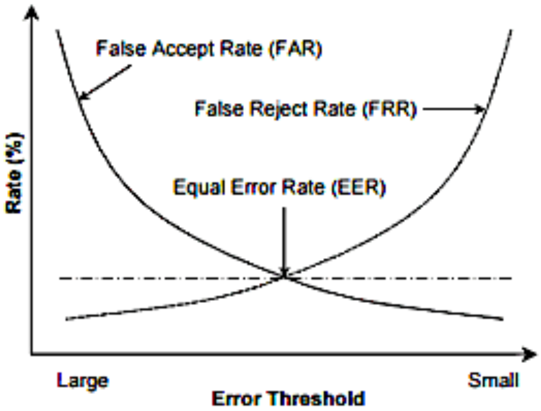


Fig. 3.18 Illustration of Equal Error Rate (EER), FAR and FRR curves

From a criticality point of view, we assume that misclassification in healthcare (misdiagnosis) is as substantial or even more than falsely granted access to a system. The effect of a confidentiality breach in an information system cannot be undone since human memory cannot be reset. Similarly, the consequences of misdiagnosing a health condition could be fatal death or life-changing, and also cannot be undone.

We proceed with this assumption in this thesis, which cannot be validated by literature or justified beyond the logical view of the author. This assumption is made due to the need for a new balanced evaluation metric for susceptibility prediction in healthcare.

3.12 Chapter summary

This chapter discussed a variety of machine learning modelling methods and listed considerations that reside in data modelling cycle. The chapter briefly describes the methods' learning mechanisms of eight algorithms of different learning schemes. It presents in the available literature and practice the limitations and implications of each presented method in every section and subsection when applied.

After presenting the general scope of machine learning tasks, the chapter starts with the classification learning workflow, without doing so, it would have been impractical to pool applicable algorithms for this review. The chapter then continued with reviewing the essential algorithmic settings, including, parameterisation, meta-learning techniques, and model performance estimation methods and metrics which are required to assess the model's generalization capabilities.

The theoretical and practical recommendations, critique and limitations provided in each section highlights gaps in knowledge which the author will consider when building a methodology later on for predictive modelling in preventive medicine. When building and applying this thesis' methodology in later chapters, we will weigh the recommendations made here. Also, we intend to pay extra attention to the methods that enhance model fitting, generalisation and provide most accurate description of performance, specifically those for imbalanced learning.

Chapter 4

Methodology & Framework

This chapter covers

- *Introducing Octopus methodology & framework*
- *The rationale behind Octopus methodology*

Formulating a suitable methodology to develop new tools to predict susceptibility to adverse events in healthcare is one of our primary contributions to the interdisciplinary scientific community. To signify this contribution, we go beyond the experimental research results spread in the literature to gather acceptance from domain experts, including life sciences, clinicians, medical practitioners and data scientists at every step through our approach.

Data-driven modelling faces challenging issues in machine learning when predicting outputs for a given set of inputs, primarily if little or nothing is known about the dataset performance.

The absence of prior knowledge increases the complexity of modelling since there are typically tens, if not hundreds, of data preparation techniques, machine learning algorithms to choose from and performance metrics. Furthermore, the problem elevates when various combinations of techniques produce varied prediction performances. Finally, the challenge reaches its peak with all the different interpretations and evaluations of the prediction models.

The prediction performance is often imperfect. Therefore, reaching a consensus on the best-performing model can also be challenging, requiring critical considerations and trade-offs. In addition, building predictive machine learning models may require using techniques of high computational costs, such as feature search, optimisation problems and imbalanced learning treatments to reach an acceptable bias-variance trade-off. Finally, approving the models is highly iterative in all cases until matching an evaluation meets domain experts' success criteria.

This chapter presents the basis of our new Octopus Methodology. It shows various considerations and recommendation behind the chosen data-driven methods from those presented in the literature review, hoping to effectively tackle the predictive modelling for preventive medicine tasks in obesity [324] and treatment side effects [325].

4.1 Introducing Octopus methodology & framework

In the previous two chapters, Chapter 2 and 3, we established that machine learning classification is the desired route to create new predictive entities (models). If approved, they can potentially be deployed or further validated as tools to screen patients' susceptibility to adverse health events. We also reviewed various methods that can be used in the process. An informed collection of such methods forms the fundamental block of building a robust and efficient methodology for predicting those at risk of adverse health events. This methodology is referred to as OCTOPUS based on its graphical representation.

In the data science community, a common approach to deal with the overwhelm of a new classification project is to use a favourite technique or algorithm. Another approach is to scour the research literature for descriptions of vaguely similar domain problems and attempt to reimplement the described algorithms and configurations. Both strategies may be effective but are proven hit-or-miss and time-consuming.

Unlike these common strategies, our Octopus methodology is not based on previous similar predictive modelling attempts to solve our new domain problems, but purely on our original research into the workings of many techniques, algorithms, and their limitations, evident by their effectiveness in various experiments where possible from multiple domains. The formulation of the Octopus methodology takes a heuristic-systematic strategy to scope a collection of techniques that could solve our new domain problems to guide future research. The Octopus methodology can be considered a new contribution to knowledge for both data science and health applications.

It is common to inherit a choice of methods from similar domain problems. However, we do not have that option in this thesis. To our knowledge, modelling our datasets features was not attempted previously, and our results are leading. Therefore, our research, methodology and models are original, and cannot mimic previous strategies.

Octopus is conceived from our research and reviews in the previous chapters. In our methodology, we consider the data dimensionality, acceptability of specific preparation techniques, the choice of modelling algorithms, the interpretability issues and the computational cost of execution within healthcare settings.

Our methodology also imports other techniques from different domains. The review of methods which led to formulating Octopus also led to the development of a new technique to enhance imbalanced data resampling, AKA Minority Pattern Reconstruction (MPR), and our discovery of a potential new empirical performance ranking metric, XDistance, adapted from the field of cyber security. These will be presented later in this chapter.

4.2 The rationale behind Octopus methodology

A heuristic-systematic rationale is a driver behind the building blocks of Octopus. This rationale proposes two distinct aspects, heuristic and systematic when scoping multiple methods to model our domain problems. They are seen in the detailed presentations, reviews, critiques, and discussions of techniques in the literature review chapters. In the previous chapters, we engaged the reader in the heuristic and systematic processing of available information on a wide variety of methods. And in this chapter, we cover and explain our decisions on using specific methods that govern our approach to model patients' susceptibility to adverse events. The theory behind the heuristic and systematic judgements is described in social psychology [326].

We reviewed a wide range of applicable techniques; each is relevant and may be helpful to our intended modelling journey of patients' susceptibility prediction. However, the process is complex due to so many factors that went under consideration. The construction of the building blocks of the Octopus methodology is illustrated in Figure 4.1. Our methodology, Octopus, has three major parts, *Arms, Head and Mantle.*

4.2.1 Methods at Octopus arms

The outer blocks represent the arms that cover all methods visited in the literature review. The systematic rationale involves attempts to thoroughly examine any available information on a range of investigated methods through careful attention to design and intensive reasoning. Whereas the heuristic aspect involves focusing on salient and easily comprehended clues that lead to well-learned shortcuts through published scientific experiments in the machine learning field. Both aspects formed the building blocks of Octopus.

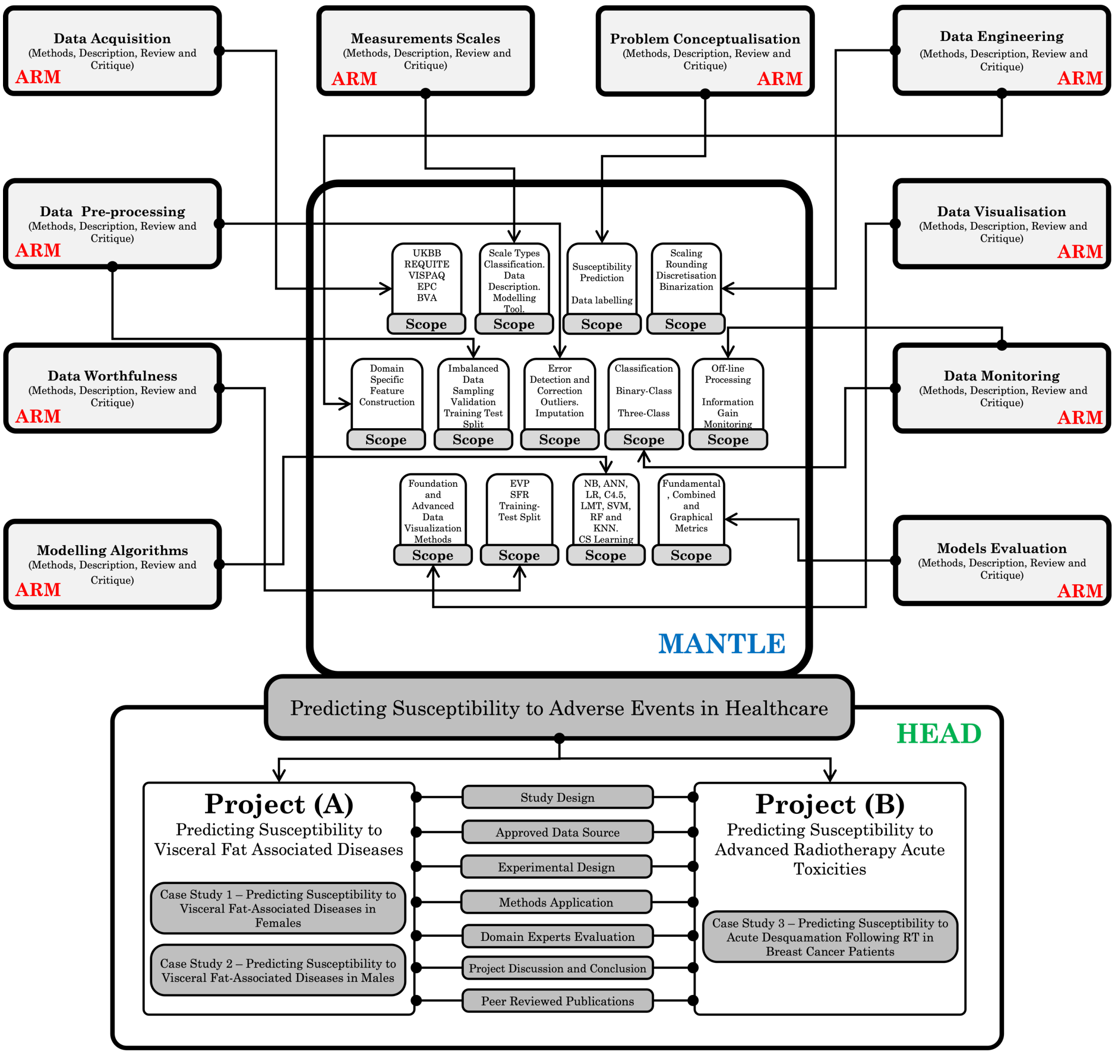

Data Acquisition
(Methods, Description, Review and Critique)
ARM
Measurements Scales
(Methods, Description, Review and Critique)
ARM
Problem Conceptualisation
(Methods, Description, Review and Critique)
ARM
Data Engineering
(Methods, Description, Review and Critique)
ARM
Data Pre-processing
(Methods, Description, Review and Critique)
ARM
Data Visualisation
(Methods, Description, Review and Critique)
ARM
Data Worthfulness
(Methods, Description, Review and Critique)
ARM
Data Monitoring
(Methods, Description, Review and Critique)
ARM
Modelling Algorithms
(Methods, Description, Review and Critique)
ARM
Models Evaluation
(Methods, Description, Review and Critique)
ARM
UKBB REQUITE VISPAQ EPC BVA
Scope
Scale Types Classification. Data Description. Modelling Tool.
Scope
Susceptibility Prediction
Data labelling
Scope
Scaling Rounding Discretisation Binarization
Scope
Domain Specific Feature Construction
Scope
Imbalanced Data Sampling Validation Training Test Split
Scope
Error Detection and Correction Outliers. Imputation
Scope
Classification
Binary-Class
Three-Class
Scope
Off-line Processing
Information Gain Monitoring
Scope
Foundation and Advanced Data Visualization Methods
Scope
EVP SFR Training-Test Split
Scope
NB, ANN, LR, C4.5, LMT, SVM, RF and KNN.
CS Learning
Scope
Fundamental, Combined and Graphical Metrics
Scope
MANTLE
Predicting Susceptibility to Adverse Events in Healthcare
HEAD
Project (A)
Predicting Susceptibility to Visceral Fat Associated Diseases
Case Study 1 – Predicting Susceptibility to Visceral Fat-Associated Diseases in Females
Case Study 2 – Predicting Susceptibility to Visceral Fat-Associated Diseases in Males
Study Design
Approved Data Source
Experimental Design
Methods Application
Domain Experts Evaluation
Project Discussion and Conclusion
Peer Reviewed Publications
Project (B)
Predicting Susceptibility to Advanced Radiotherapy Acute Toxicities
Case Study 3 – Predicting Susceptibility to Acute Desquamation Following RT in Breast Cancer Patients

Fig. 4.1 Building blocks behind the formulation of Octopus methodology

From Figure 4.1, the Octopus blocks are formed using the heuristic-systematic rationale. The arms contain a collection of applicable methods heuristically researched in the literature review. Following a heuristic aspect seems efficient in formulating the arms. It may confer less judgmental confidence but increases the feasibility of scoping from a range of methods in a domain where several hundreds of interweaving methods exist and compete with no specific winner.

The systematic aspect confers more confidence but is relatively effortful and time-consuming; adhering to this aspect plays a major part in formulating the mantle, which contains the collection of methods to be applied. The mantle's scope of methods is driven by the joint acceptability of these methods by the healthcare and data science communities. They are also motivated by the precision of the medical and health tasks in this thesis to gain domain experts' confidence and approval.

4.2.2 Methods at Octopus mantle

The inner blocks in the Octopus methodology are scoped methods for practical applications. In other words, they form the mantle. The mantle has a scope of specific methods whose collection follows a systematic rationale.

Adhering to the systematic aspect confers confidence but is relatively effortful and time-consuming. The collection of methods within the mantle is motivated by the precision of the medical and health tasks in this thesis to gain domain experts' confidence and approval. All scoped recommended methods are based on the previous literature review chapters. The recommended methods scoped in the mantle are:

A) Problem conceptualisation methods: Earlier in this thesis, we defined data-driven susceptibility prediction as a machine learning classification problem.

The proposed definition adopted by this thesis is: "Susceptibility is a capacity characterisable by intrinsic and extrinsic factors that modify the impact of a specific exposure upon risks/severity of outcomes in an individual or population."

We also adhere to three categoric statuses; resistant individuals are those at no risk; in some parts, we refer to them as healthy for having no relative propensity to develop disease; susceptible individuals are those who are a risk of developing a disease, their risk is maybe referred to as moderate in the presence of individuals who are at high risk in relation of their propensity to develop a disease or in another word hyper-susceptible.

The adopted definition lacks statistical concepts, which limits its direct applicability to technical methods. However, it is simple to deploy by an interdisciplinary group of experts in a project. The consensus among the collaborators in this thesis is to predict the susceptibility categories as labels to address the health risk. This agreement led all data-driven modelling in this thesis to follow the machine learning classification approach.

Based on the adopted definition of susceptibility, this domain problem could have two labels; these are resistant and susceptible, or three by adding a hyper-susceptible label. Therefore, the mantle targets two different susceptibility problems in healthcare, these are located at head of the Octopus Framework.

One problem is the prediction of breast cancer patients' susceptibility to a radiation therapy side effect, moist desquamation (Desq). This is a binary classification problem with two labels, resistant and susceptible. The second is the prediction of patients' susceptibility to visceral fat, also known as Visceral Adipose Tissue (VAT) associated diseases for both genders. Again, this is a multi-class classification problem with three groups of patients, those who are deemed healthy or have no risk of developing such diseases, also considered resistant to VAT-associated diseases; a group of moderate-risk patients with a higher likelihood of developing VAT-associated diseases; which are deemed susceptible, the last group is those who are at the highest risk of developing VAT diseases who also can be considered hyper-susceptible.

B) Data acquisition and quality methods: REQUITE provides breast cancer radiotherapy data. REQUITE is an international project that aims to predict which patients are more likely to develop side effects from radiotherapy. It is a multi-centre observational study and the largest study of its kind, collecting data from 5,300 cancer patients (~ 2000 breast cancer patients) and a credible data source to model patients' risk of developing side effects following radiotherapy.

It is a unique resource for studying the relationships between side-effect endpoints and quality-of-life (QoL). This project received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 601826 [327].

The VAT dataset is a dataset provided by the UK-Biobank (UKB). The UK Biobank is a large-scale biomedical database and research resource containing in-depth health information from half a million UK participants. Their data repository is regularly augmented with additional observations and is globally accessible to approved researchers. The UKB data repository is the largest and richest dataset of its kind. It is anonymised and made widely accessible by UK Biobank to researchers worldwide to make new scientific discoveries about common and life-threatening diseases, such as cancer, heart disease, diabetes and stroke, to improve public health [328].

Both sources provide a large variety of data types, including clinical, imaging and genomics data. However, we would like the reader to think that our approach is unique as we strive to develop inexpensive, approved models. Our aim is achieved by processing variables associated with low-cost collection methods.

Therefore, the processed dataset's attributes are numeric or nominal, and the analysis is cross-sectional. The collected data are obtained from low-cost measurement methods such as anthropometry, physical exercises, email surveys, questionnaires, clinical observations and historical familial and medical health records. Our approach is not formulated to process data associated with higher-cost collection methods, such as imaging and genomic sequencing data.

From our literature review, it is acknowledged that there are many frameworks for Datasets Quality Assessment (DQA) [19], but they are objectives-centric and context-dependent. To address this problem in our methodology, we propose new Data integrity checks using our new VISPAQ six-point quality check. VISPAQ is a non-metric operational approach to explore the data quality at the repository to raise our confidence when approaching such a data source to acquire data for a project (see Figure 4.2). We recommend following this approach before making financial or project commitments.

VISPAQ six-point checks stand for Versatility, Informativity, Suitability, Presentation, Accuracy and Quality. The **versatility** of any dataset held by any source is a gate to various feature selection tasks and techniques, offering different ways to analyse the data for better insights. It is about checking the blend of comprehensive granular, detailed and summarised data levels. Such a variety makes deriving new features from the original data beyond limits. This check should be done within the type of data required for a study, i.e., transactional, imaging or sensor data types.

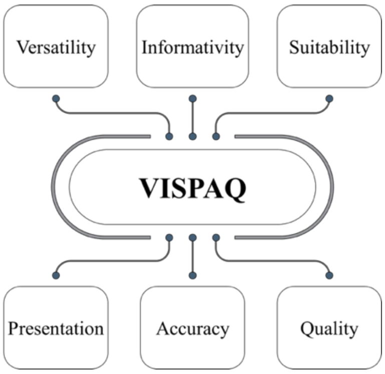


Fig. 4.2 Data source six-point check

Data **informativity** can take multiple shapes that vary from a broad material to a detailed, in-depth material. No matter how predictable data fields might be, resources, definitions, and examples must be comprehensive to tackle any unforeseen variability within the data. In addition, gathering information regarding how the data was collected and should be handled and any future changes is a big plus.

Data **suitability** is to do with declaring some recommendation of possible research domains for researchers, typically based on historical access applications made to the source and previous publications. Data sources should advertise key research areas where subsets of their data are applied. Adding a repository of related research publications forms a major advantage.

Examining the **presentation** of the data characteristics in various graphical, tabular, or statistical formats is crucial. Although such an assessment mainly enhances the data scientists' and analysts' understanding of the domain problem, it can become project-critical if a unique cohort is required for analyses.

It is challenging to judge data **accuracy** without access to real datasets. Nevertheless, some indications could reflect on the accuracy of the data, such as the planning efforts in organising the data collection, the tools and devices which were used for obtaining readings, the presence of validation via repetitive measurements, the existence of data collection protocols for recording the measurements and the ability to customise data before release. Integrating the data source with other credible systems like national health records is also a promising indicator.

Everything we mentioned so far is related to the data quality itself. However, systems **quality** and maintenance at a data source usually show confidence. For example, seeing communications and announcements of new processes and system changes by the data source may indicate that the data quality does not degrade over time and is well maintained. In addition, a data source declaration of a system's industry certification, i.e., ISO certifications, is a strong indicator of data excellence.

VISPAQ checks are advisable but not comprehensive, and there are always other attributes by which data quality and source are attested. Also, we are not against the use of publicly-free available ML datasets [20].

However, using such data comes with many challenges and unanswered questions, especially in results-critical domains like healthcare [21]. For instance, patients' role is discarded in such data due to ethics. But what if follow-up information is required, like death or dropouts? Also, the lack of predictability and limits of measurements is another issue. The natural occurrence of a health issue takes time and cannot be projected; data needs to capture enough incidents, raising the issue of "how this wait is monitored?" Every measurement technique can be used arbitrarily frequently for technical, ethical or compliance reasons. To derive conclusions that matter for the healthcare community, like improved prognosis or improved quality of life (QoL), you need observation periods of ten or more years. At this time, the individuals' characteristics will change and continue to be stored in a distributed public health system; this raises the question of how this data can be accumulated from the distributed systems into one place for analysis. [21].

Nevertheless, there are wider challenges that face the analyses from all sources, such as data heterogeneity and concept drift is another issue that renders the independent testing and verification of published results and their translation into practice very difficult [21]. With all the above in mind, there are now enough considerations to sway this research away from publicly-free available datasets. As a result, and based on the collaborators' recommendations, two health data providers met this thesis's research domain: the UK Biobank [328] and the REQUITE Consortium [327].

The UK Biobank is a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants. REQUITE is a European observational study, the largest of its kind. It is recognised as an exemplar of multi-disciplinary, multinational work that should be carried out in radiotherapy-related research.

Both REQUITE and The UK Biobank resources validate all VISPAQ checkpoints to a certain limit and overcome potential issues with publicly available datasets clarified by visiting their websites. Both sources are subject to legal restrictions such as complying with legislation including consent, confidentiality, licensing, ethics and intellectual property, which prolongs the timeframe for data release between six months to a year, and their data collection is subject to rigorous protocols. There are three extracted datasets from these resources which we utilised by our methodology for our practical modelling of susceptibility in this thesis:

- The female cohort MRI visceral fat dataset is extracted from the UK Biobank resource (4327 subjects).
- The male cohort MRI visceral fat dataset is extracted from the UK Biobank resource (4126 subjects).
- The Breast Cancer Radiotherapy Toxicity prediction models' dataset is extracted from the REQUITE Consortium (2069 subjects).

To examine both data sources, we obtained a data sample from each source to determine the feasibility of completing a successful data analysis. Both data sources, the UK Biobank and REQUITE Consortium are not perfect; however, their protocols and controls provide confidence in their data collection. Both met most of the quality indicators regarding *Versatility*; they had a wide range of data for their cohorts covering various patient characteristics. This coverage is translated into thousands of variables offering different ways to analyse the data to enhance insights. *Informativity* is incredibly unique in the UK Biobank case, and resources are provided in detailed literature and statistical forms on each variable, including data stability status and dictionary. Unlike the UK Biobank, REQUITE Consortium does not offer similar detailed information and resources regarding its data. Nevertheless, their data collection and survey design are precise. Still, the absence of a clear data dictionary and literature resources makes all analyses on their cohorts require significant consultations from physicians.

Both data sources have a good list of publications to guide data suitability, indicating the historical analysis types. The REQUITE Consortium is relatively new compared to the UK Biobank. Thus, the latter holds a more extensive base of publications. The UK Biobank advertises the data scale measurement types per variable. In the REQUITE Consortium, the data measurement scale types can be partially determined from the data collection design but, in all cases, require clarifications from physicians and informed discussions. The UK Biobank advertises high standard *presentation* of data characteristics on their website in graphical, tabular, or statistical formats; however, REQUITE had their basic data stats published in a research article.

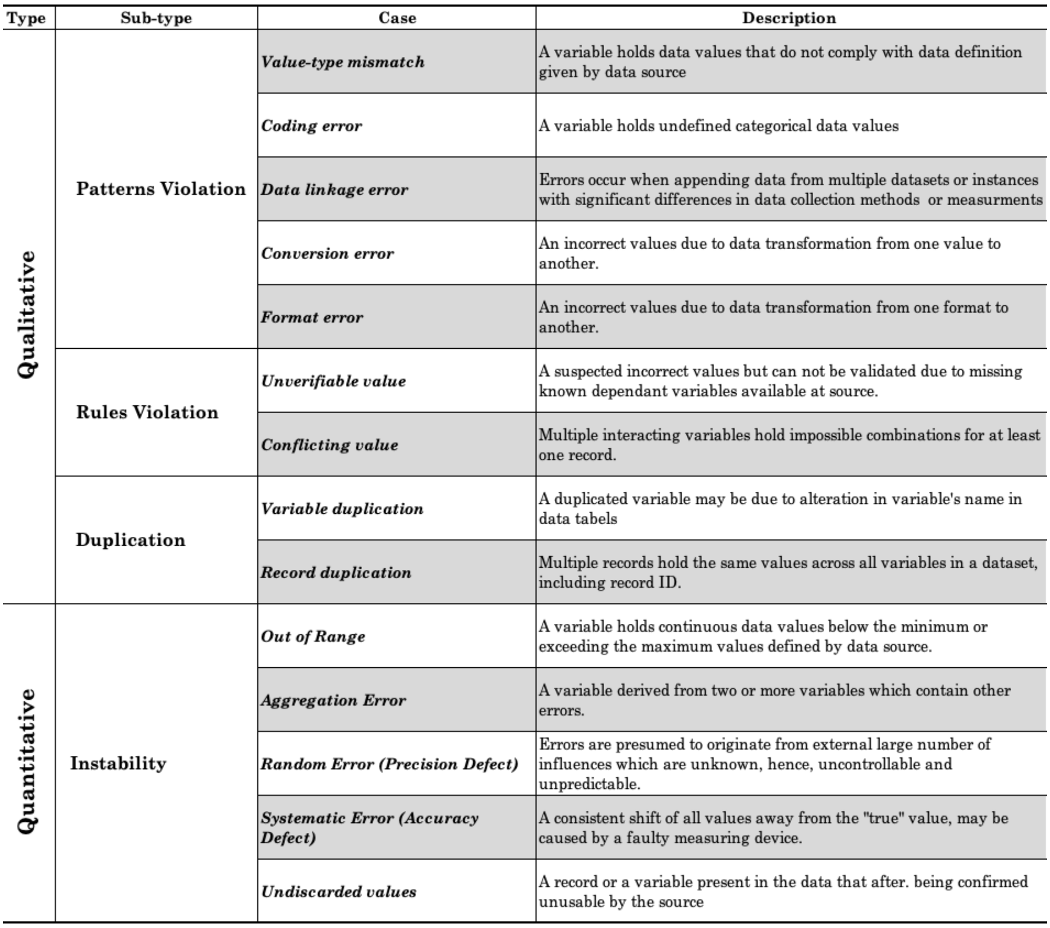
We examined observations at random for their *accuracy* in the samples provided. Unfortunately, both sources had accuracy deficiencies. The REQUITE cohort was more severe than the UK Biobank. Both record samples had qualitative and quantitative issues, prompting the necessity of finding a suitable strategy, such as *Boundary Value Analysis (BVA),* to further investigate the entire extracted data. *Qualitative defects* include pattern violations, rule violations and duplications. *Quantitative problems* arise in the form of inconsistencies and outliers. Such a scenario, if found, could lead to excluding some variables and records from analyses. Finally, *quality* controls at both data sources are declared in their followed protocols and their industrial certifications for collecting and maintaining their data.

Any accuracy defects found in the samples provided by the sources would influence our decision to apply the Equivalence Class Partitioning (ECP) and Boundary value analysis (BVA), described in the literature reviews. Both methods assist in validating observations around specific boundary values within each variable where recorded values outside these boundaries alter the recorded values in other variables. However, this process is highly iterative, time-consuming, requires information from the data source and domain knowledge, and is very lengthy to document.

Often, in interdisciplinary case studies, the data scientist governs the execution of this strategy. Still, the decision on boundary values, errors, corrections and the best course of action lies in the hands of domain specialists. Table 4.1 shows a non-exhaustive list of frequently known defects that can cause concern for healthcare specialists if observed in a dataset.

The discovery of such errors (defects) provides assurances of data correctness before feeding to transformation and modelling algorithms. The severity of errors varies depending on their propagation in the data transactions. Errors could propagate horizontally at a record level and vertically at a variable level.

Table 4.1 A list of common defects found in datasets



Type	Sub-type	Case	Description
Qualitative	**Patterns Violation**	***Value-type mismatch***	A variable holds data values that do not comply with data definition given by data source
		Coding error	A variable holds undefined categorical data values
		Data linkage error	Errors occur when appending data from multiple datasets or instances with significant differences in data collection methods or measurments
		Conversion error	An incorrect values due to data transformation from one value to another.
		Format error	An incorrect values due to data transformation from one format to another.
	Rules Violation	***Unverifiable value***	A suspected incorrect values but can not be validated due to missing known dependant variables available at source.
		Conflicting value	Multiple interacting variables hold impossible combinations for at least one record.
	Duplication	***Variable duplication***	A duplicated variable may be due to alteration in variable's name in data tabels
		Record duplication	Multiple records hold the same values across all variables in a dataset, including record ID.
Quantitative	**Instability**	***Out of Range***	A variable holds continuous data values below the minimum or exceeding the maximum values defined by data source.
		Aggregation Error	A variable derived from two or more variables which contain other errors.
		Random Error (Precision Defect)	Errors are presumed to originate from external large number of influences which are unknown, hence, uncontrollable and unpredictable.
		Systematic Error (Accuracy Defect)	A consistent shift of all values away from the "true" value, may be caused by a faulty measuring device.
		Undiscarded values	A record or a variable present in the data that after. being confirmed unusable by the source

ECP and BVA can use several tests to discover errors or suspicious values in a dataset. Such tests include *Parameters' Verifications* against given specifications from the data source, *Logical Comparisons* that rely on the data collection rules and survey design, *Value-Decompositions* of aggregated values with known origins, *Stability (behavioural) checks* by using indicators confirming the reliability of a record or a variable declared by the data source.

C) Measurements scales methods describe our strategy for handling data scales of measurements before transformation and modelling. We adopt scale types for our practical modelling in line with the capabilities of the modelling tool and algorithms.

Some machine learning algorithms have built-in perspective strategies to handle specific data types. For example, the modelling algorithms used for this thesis, built in the Waikato Environment for Knowledge Analysis (WEKA), can handle categorical value type as "Nominal" and both continuous and discrete values as "Numeric".

Although our selected algorithms have variations that handle ordinal class labels based on class decomposition in Frank and Hall's approach [292], we treated all ordinal class labels as nominals.

As for predictors, we reviewed two approaches; the first passes the ordinal values as numeric type to the classification algorithm. However, we do not recommend such an approach; this may produce problematic results if the classification algorithm is led to consider non-existent orders in between the ordinal objects for decision making, in decision stumps, for example. The second approach encodes the categorical features into multiple binary features using one-hot encoding [329]. But we have reservations when transforming a small number of ordinal categories into new variables as the information levels within the original variable vanish.

In practice, addressing data measurement types in data-driven modelling depends on the capabilities of the tool of choice. Various modelling tools have built-in perspective strategies to handle specific data measurement types before consuming a variable by the modelling algorithm. In our methodology, we classify all ordinal and binary attributes as nominals. Thus, the predictors in our practical modelling will only be processed by the modelling algorithm as numeric or nominal types. The class labels are classified as nominals.

Figure 4.3 shows the classification of measurement scales followed in this thesis to assign to different variables. By accounting for the element of arithmetic significance (applicable arithmetic operations), we simplify giving the attributes to a higher-level scale of measurements consisting of three ubiquitous data types: nominal, numeric, and binary. Variables of categorical values without a known magnitude are nominal. Ratio, counts, and interval type values are considered numeric type. Ordinal attributes with known magnitude can take either form of numeric or nominal data types. Likewise, binary variables can be assigned binary or nominal types. However, there is a limitation. Our topology assumes all numeric features to be continuous and does not account for discrete numeric values such as counts. This assumption may change the natural distribution of parent variables when applying data preparation and pre-processing techniques such as data imputation and augmentations.

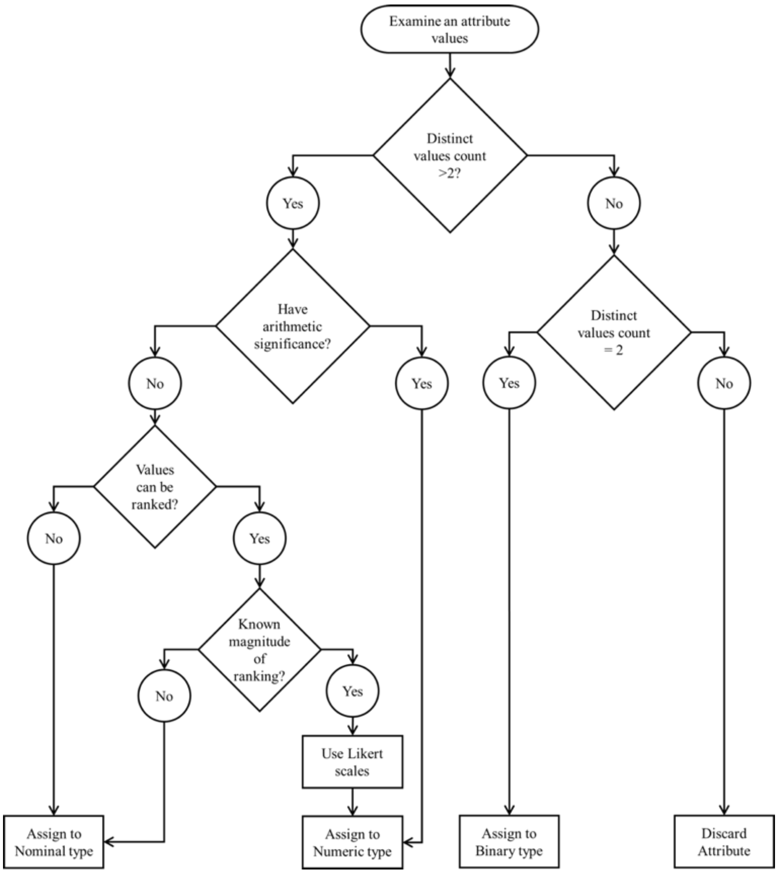


Fig. 4.3 Classification of data measurement scales

D) Transforming large categorical features' methods: For a nominal input features with a large number of categories, one may convert its categorical values into multiple attributes. This transformation is done by applying the One-Hot encoding binarisation technique on each category, transforming it into a new feature.

E) Features summary declaration. Our review in chapter two, regarding variables measurement scales, impacts variables' summary presentation. When comparing variables within a dataset, a standard recommendation is to present descriptive statistics of features as per Table 2.1 in Chapter 2. We also recommend the classical approach of reporting basic statistics of variables and the Information Gain evaluation to indicate the association of various features to the class attribute.

However, when comparing multiple datasets in classification, including training and test sets, we recommend reporting Information Gain per variable as a measure of association with the class variable. Machine learning algorithms tend to be biased towards variables with higher purity. It is interesting to evaluate whether the test data is easier to predict than the training data. The measure of information gain should be used in cases where both datasets are equal in their count of instances. Purity in the data may change due to applying various data pre-processing techniques, including imputation, certain types of scaling, resampling and others.

F) Data visualisation methods: The literature review chapter investigated various methods to explore and visualise datasets for modelling. Regarding high-dimensional data visualisation, we prefer Targeted Projection Pursuit (TPP). Unlike Principal Components Analysis, our choice is based on preserving the view of original variables. TPP offers advantages over other methods, including the automatic blind projection of interestingness. In addition, TPP allows the users to explore these projections by manipulating individual data points or groups of points directly in a multidimensional scatter plot.

G) Precision of measurements (features rounding): Multiple types of feature rounding are presented in chapter three. One, in particular, is rounding up, centred on the tie-break value of 0.5 [135].

Through we acknowledge that the precision of an observation value is related to the significance and interpretation of that value within the health domain. Some value precision might be critical in the case of radiation doses, while others are less important, for example, in the case of Body Mass Index (BMI) values. Therefore, using random rounding is not advised here.

As per feedback from the collaborating physicians seem to accept the use of rounding up for some variables whose precision is not vital, disregarding any bias this may cause towards the higher values. However, this bias may skew the patient's characteristics towards the better or worst outcome, painting a pessimistic or optimistic prediction picture. Therefore, features rounding should be performed under domain experts' supervision and remains optional in our framework as long as it does not introduce unrealistic or conflicting values.

H) Missing observations mitigation methods: Chapter two presented and discussed multiple methods to address missing data by seeking its cause, description and resolution. We combined what we learnt from the literature into a new *Multi-Method Imputation (MMI)* framework to handle missingness. And we recognise that our recommendation to use MMI for handling missingness is a guide similar to *Equivalent Class Partitioning (ECP)* and *Boundary Value Analysis (BVA)* for handling errors.

Practitioners are advised to document key outputs of the overall MMI process flow due to the highly iterative nature of the process. Furthermore, MMI decision flow is a lengthy process; therefore, one may not want to document the output of every single decision path when practically applied. Our new framework, MMI, accounts for the cause, pattern and proportion of missingness. In addition, it pays attention to the domain of the problem at hand and the acceptability of specific missingness treatments in that domain (healthcare in our case). MMI then suggests the best course of action to solve missing observations in a particular dataset.

MMI has five stages, *Prerequisite*, *Avoid*, *Ignore*, *Top-up* and *Treat*. The MMI process flow in Figure 4.4 ensures that any artificial observations inference via simple or model imputation is a last resource after exploring all other non-statistical options to compensate for missing observations with actual, valid, logical or informed values.

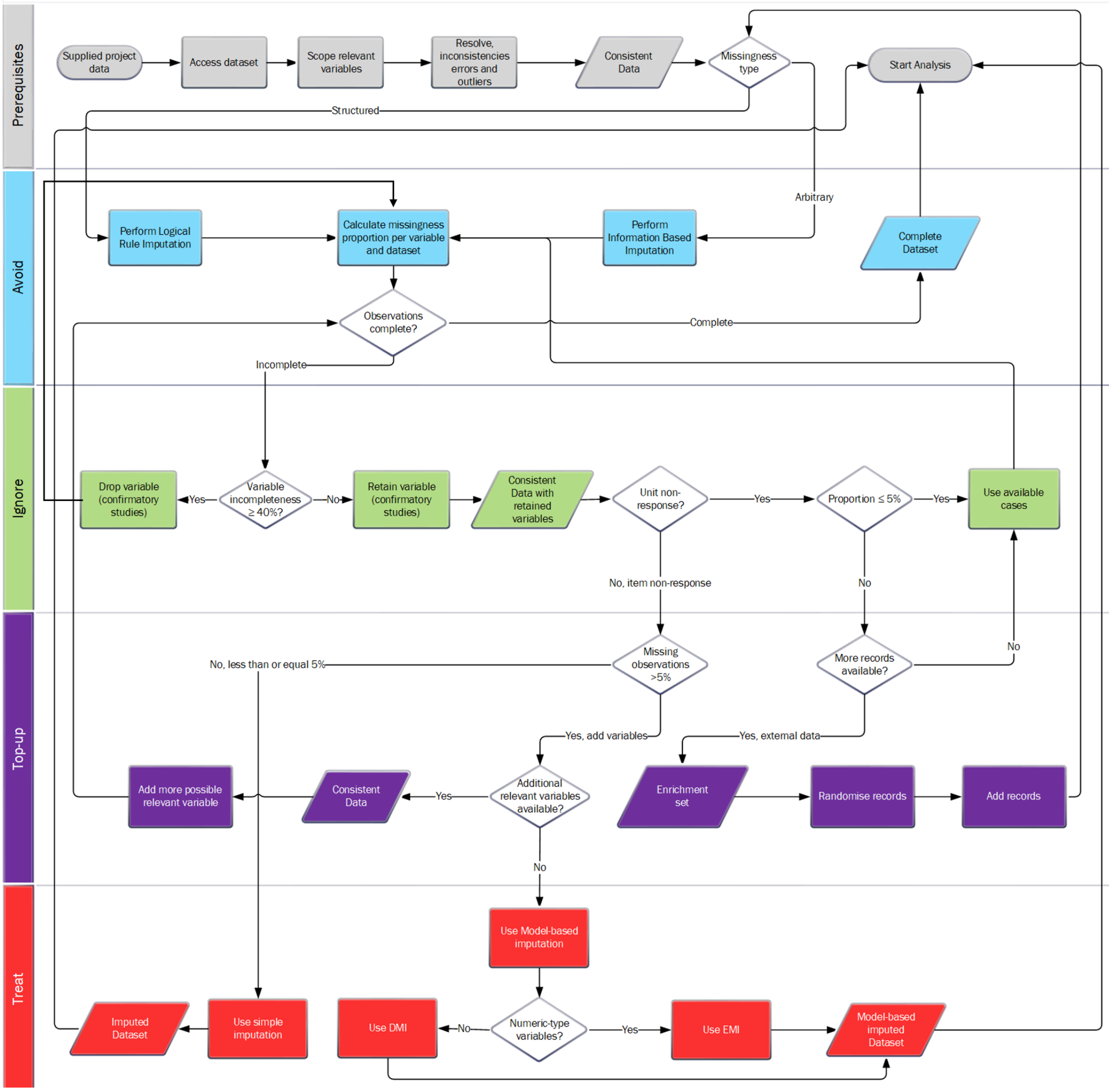


Fig. 4.4 Multi-Methods Imputation (MMI) framework

We acknowledge that finding the actual values for missing observations has no perfect answer since true missing data is unknown forever. Thus, there is no way to validate the replacement values.

We can only put so much effort into following reasonable steps to reduce any potential impact on the analyses. We prefer following MMI to increase our possibility of compensating missing observations with values that are true or of true-like. Thus, MMI ensures that any exploitation of statistical characteristics comes last, hoping to reduce the potential use of unreal data in classifiers' decisions.

We note that the absence of consensus on many of the discussed concepts, the limitations of the current methods, the differences in data characteristics and the varied domain priorities make it almost impossible to draft a unified approach that fits all cases of handling missing data. The Multi-method imputation (MMI) is a new heuristic map that tries to provide a logical flow to handling missing data under certain conditions. MMI's priority is achieving the acceptability of the collaborating healthcare professionals and physicians in this interdisciplinary thesis.

These processes in Figure 4.4 ensure that any artificial observations inference via simple or model imputation is made as a last resource after exploring all other non-statistical options to compensate for the missing observations with real, logical or informed values. We try to minimise the exploitation of statistical data characteristics as the last resource due to the potential impact of their disruption on classification learning [116].

The mapped conditions aim to assist researchers in substituting missing values based on linking various recommendations from the data science community and healthcare research. We try not to distort the variable's distribution to prevent building overly optimistic or overfitted models which fail to generalise on external test datasets.

In MMI, The **Prerequisite** stage assembles all the elementary requirements to start an informed discovery of missing data. This stage includes accessing the data, scoping the relevant (applicable) variables (from literature and domain experts' knowledge), ensuring present values are consistent, and resolving any measurement or formatting errors. In essence, a technically correct dataset is produced and deemed suitable for missingness analysis and treatment. The **Avoid** stage has a conditional gate that ensures real/realistic data values are found first to narrow down the proportion of missing observations. This stage establishes if the missingness can be substituted with logical values extracted from the survey design and information within other variables. The **Ignore** stage identifies the portion of the missingness that is ignorable if gone untreated by recommending a threshold of 5%. Depending on the type of missingness spread (non-response), a suitable path is recommended to discard, top-up, or treat the data. Whichever method is used at this stage ensures that no bias is introduced in the analysis (based on literature) for the available records and variables. A ceiling of 40% missingness proportion within a variable was also chosen to decide whether a variable is retained for the analysis to conceive confirmatory results in health studies.

The fourth **Top-up** stage suggests a feasible route to consider when dealing with missingness that is not ignorable. It considers various elements, such as the availability of additional data and whether to add instances or suspected relevant (applicable) predictors capable of minimising the portion of missingness. The final **Treat** stage recommends suitable approaches to imputing missing data while considering the portion and the type of missing observations. Logically, we acknowledge that using variable dropping in the process could potentially create more latent variables, hence recommending using EMI to handle potential latent numeric variables caused by variable dropping. If nominal values also existed in the data, the DMI with embedded EMI takes care of imputation within such datasets. Finally, through our MMI approach, the MMI approach conditionally groups various the use of imputation techniques with practical consideration and decisions driven by referenced our previous research in literature, which is supported by mathematical proof. The aim is to produce datasets suitable for cross-sectional confirmatory studies analysis as far away from biased analysis as possible in the eyes of domain experts.

There is no right or wrong way to go around this problem. Any chosen method will eventually get the data to the intended analysis pipeline. However, the end purpose of the analyses and their acceptance in the domain's community are the jury who gives the verdict.

I) Data scaling and irregularities mitigations methods: Previously, we reviewed the advantages and limitations of multiple scaling techniques in many key areas, including robustness and sensitivity to outliers and extreme values, the change of the underlying distribution of the data leading to changes in associations and improvement and deterioration of the modelling performance. It was also mentioned that in research, choosing the right scaler is often a trial-and-error process, and no single best scaler works every time. One may try all possible combinations of scaling methods on any dataset to develop a deeper understanding of the effect of normalisation on the modelling performance for that particular data. However, despite that being a time-consuming approach, any observations of improvement or deterioration won't be sufficient to generalise across all datasets.

Furthermore, there is no unified approach to data scaling for all applications. Thus, no single scaling approach can be ranked as the best among all the techniques. We acknowledge that Feature scaling is an essential step in data pre-processing. And thus, it is imperative to decide which feature scaling to use. Our research believes that in the healthcare domain, the machine learning model's interpretability, in some cases, is important. Therefore, the interpretability of the scaled features in a model should be mitigated. In the literature review, we emphasised focusing on feature scaling if seeking an enhanced output when considering specific algorithms for model-based imputation or predictive modelling, such as the KNN algorithm.

Considering all the above, we recommend using monotonic scaling that converts the features from having a variety of different dimensional units and magnitudes of measurement into unified dimensionless features.

Not only is our recommendation widely acceptable in the healthcare community, but it also scopes standardisation and minimum-maximum normalisation. We do not have a preference for which technique to use. We emphasise utilising either method with at least certain types of learners, such as curve fitting, hyperplane and distance-based learners, in cases where features' magnitudes largely vary. In addition, using either of the scaling techniques with other learners, such as C 4.5 decision trees, imposes no issues on performance beyond interpretability. Both methods preserve the information association among variables and the class.

Nevertheless, we accept these scaling techniques' susceptibility to outliers and extreme values. When it comes to outliers, we addressed that there is no standard statistical definition of outliers and extreme values. Therefore, we recommend using only magnitude-independent detection methods, such as the Inter Quartile Range (IQR) test. We look for values with an Extreme Value Factor (EVF) of 6 times the IQR and an Outlier Factor (OF) of 3 times the IQR. However, these factors are only application defaults and not standard, and we could not scope any experiments from the literature to suggest otherwise.

Once irregular data points are identified, they are carefully examined by domain experts for their validity. On the one hand, if found valid, they should be retained. On the other hand, if such abnormal observations are suspected errors and not a phenomenon, deleting the outliers is the most straightforward approach. In case of deletion, this thesis recommends following the 5% missingness threshold to avoid statistical bias. However, suppose they are of a higher ratio and cannot be validated. In that case, their retention is advised, and avoid using scaling methods that suppress them or distort the original statistical association between their features and the outcome.

J) Scaling and training-test splits sequence: Data scientists and domain debate the sequence of applying feature scaling (normalisation or standardisation) before or after sampling the training and test sets.

It is a two-way street! Unfortunately, there is no agreement in the data science community on such a debate. Some prefer scaling the training set and then inheriting the scaling parameters from that training set to scale the test set. Others state there is no problem scaling the whole data before the split. Machine learning developers argued that in both cases, scaling parameters are leaked from training into the test [330]. In general, other than solving the problem of magnitudes, the leakage of scaling parameters also serves to have consistency between training and test feature sets. This consistency is achieved because the new scaled values in the training and test subsets behave within unified boundaries estimating the scaled values (i.e., The mean and standard deviation for standardisation and the minimum and maximum values in normalisation).

Feature scaling before splitting ensures no issues in the evaluation step, which could affect the model's performance. However, some data scientists may say this raises concerns about knowing some test data behaviour in training. The question of consistency arises if the test data is scaled separately without dependence on the training data scaling parameters. In such as case, we cannot guarantee that the scaled values remain within the exact boundaries of the training data. A test dataset in a different range or distribution could reasonably make a model perform worse. One common recommendation is to split the data to obtain the test set, so it is not scaled with training examples. Instead, scale the training subset alone, and save the training set scaling parameters. Then scale the test set inputs using the parameters obtained during training, then de-normalizing the predictions with these same parameters to return the user output in the original scale [330]. Others stated, based on their practical experience, that if scaling is required, then it should be done on both the train and test data sets [331]. Also, the approach taken may depend on the data and the type of scaling.

For standardisation, obtaining the Standard Deviation (SD) and the mean of the whole data (train and test together) will allow the scaling transformation before or after the train-test split. However, if minimum-maximum normalisation is used, it is crucial to know the minimum and maximum values of the dataset. In such as case, a problem emerges if the normalisation is done on the train and test sets separately, then we may end up with possible different min-max ranges, and the models may not perform as expected. In production, if new min-max values are constantly observed, this would indicate the need to retrain the model with the new rescaled data. The key is ensuring matched scaling transformations in modelling and production workflow [331].

Based on the above debates, we assess both cases on the grounds that they are forcing the original data values to be a fraction of a unified unit. There are three elements to achieve the unified measurements' unitlessness. The first element is parameters with the minimum-maximum combination (in normalisation) or the mean-SD combination (in standardisation), the second part is the originally recorded observation itself, and the third is the nature of the scaling being unsupervised (label-independent). For consistency, we could state that using the first element assures that the transformed values within a feature in both training and test remain of comparable magnitudes. This is because the actual observation value, before transformation, governs the new transformed value's magnitude within the feature space's constant boundaries. Therefore, as long as the original observation value of the test record does not expose its displacement in the feature space during the training phase, both suggestions are correct. Therefore, reusing scaling parameters only unifies measurements unitlessnessly within a feature.

By considering our same principle, we emphasise that scaling parameters reuse could be beneficial from a consistency point of view of present observations.

K) Discretisation methods: In the literature review, we examined multiple studies that reported an improvement via discretisation to Naïve Bayes (NB) and Bayesian Network models' performance [136][137].

Although we consider the discretisation step optional, we use multi-interval discretisation of continuous-valued attributes for classification learning (MDL) [145] when combined with NB modelling. MDL, a supervised method, overcomes the challenge of selecting the number of bins using the information theory to minimise the entropy within a variable based on the class label.

The recommendation of using MDL is driven by its intuitiveness. It finds the best split, so the bins are as pure as possible. Thus, most values in a bin have the same class label, leading to finding the best degree to influence the separability of the class labels. An efficient discretisation is characterised by finding the split with the maximal information gain, which entails maximising the entropy function over all possible class boundaries.

L) Imbalanced learning methods: Multiple state-of-the-art techniques were systematically reviewed to enhance imbalanced learning modelling. Unfortunately, the literature showed no consensus on a single method when dealing with a class imbalance problem. We highlighted various studies that reported the superiority of Random Under-Sampling (RUS) over other techniques [174]. Others found that Random Over Sampling (ROS) outperformed RUS, although the behaviour differs widely for each data set [177]. Additionally, a recent empirical study concluded that using oversampling techniques, including SMOTE (and its variations), ADASYN, CBO and others provide overly optimistic predictions when tested on unseen data due to class label leakage [181].

Therefore, in this thesis, we acknowledge the concerns surrounding the removal of data points, in particular, those from the positive class. These samples could be, in some cases, partially representative of a phenomenon mainly in medical sciences, such as susceptibility to an infection or a rare response to a treatment associated with influencing factors that cannot be exhaustively gathered. We also act on information that the medical domain seems to accept synthetic data generation as a tool for scientific discovery but not for the adoption of diagnostics [180]. Also, from our review of Cost-Sensitive (CS) learning, CS competed with other resampling approaches, and in some cases, it outperformed them on larger datasets.

To mitigate the imbalanced learning problem, we recommend starting with Random Under-Sampling (RUS), Random Over Sampling (ROS) (although reported many cases of overfitting), SMOTE and Cost-Sensitive learning. The positive class (class of interest) examples remain without elimination in all these techniques.

Furthermore, we recommend applying the CS learning method only to binary classification problems because estimating misclassification costs for new classification problems is tedious. Thus, the number of experiments accounting for all cost matrix variations grows exponentially with the count of class labels. It is also expected that various machine learning algorithms to produce varied performances and behave differently for the same explicit cost matrices.

Finally, we recommend experimenting (where possible) with a novel approach of data resampling, Minority Pattern Reconstruction (MPR) and observing any enhancement it provides to the traditional methods.

The new MPR approach hypothesises that various data points exist in space and contain different patterns based on the weights of their variables, as described earlier in targeted projection pursuit. The weights' modifications allow analysts to see these patterns in different projections. A perfect or near-perfect classification can be achieved if the data points follow a single pattern within a set that makes the classes entirely separable; however, this is uncommon in real-world data. The data likely contains multiple patterns in space; each pattern contains highly separable class labels. In our hypothesis, we believe that extracting each pattern set and using each to train any classifier achieves a perfect or near-perfect classification.

Therefore, we run multiple cycles of the supervised K-Means clustering algorithm to extract such classification patterns from the data. Each run favours the minority class whose label reserves the highest importance in an imbalanced multi-class classification problem. Two datasets emerge from each cycle, a pure pattern set and a set of residuals. We use the blind Targeted Projection Pursuit to project automatic interestingness on the whole data before the first cycle. And then, it is applied to the residual subsets before every consequent clustering cycle. We also hypothesise that each extracted pattern holds an information profile that can be measured by summing the information gain evaluation of its features in bits. These information profiles are also analysed and compared to the original data profile for the whole dataset.

In the end, all extracted pattern sets are appended, forming a new imbalanced dataset to build classifiers or to undergo other sampling techniques such as RUS and SMOTE, illustrated in Figure 4.5.

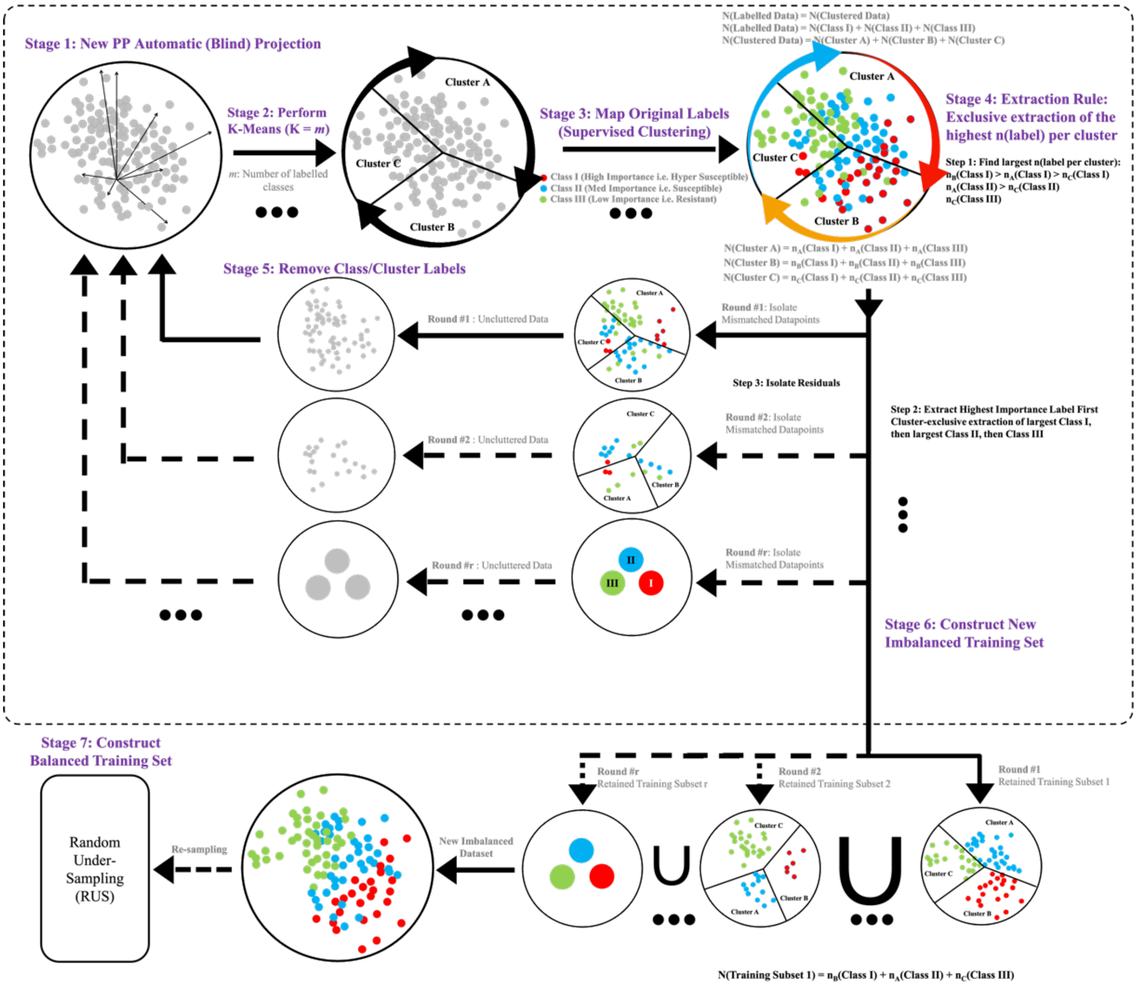

N(Labelled Data) = N(Clustered Data)
N(Labelled Data) = N(Class I) + N(Class II) + N(Class III)
N(Clustered Data) = N(Cluster A) + N(Cluster B) + N(Cluster C)
Stage 1: New PP Automatic (Blind) Projection
Stage 2: Perform K-Means (K = m)
m: Number of labelled classes
Cluster A
Cluster C
Cluster B
Stage 3: Map Original Labels (Supervised Clustering)
Class I (High Importance i.e. Hyper Susceptible)
Class II (Med Importance i.e. Susceptible)
Class III (Low Importance i.e. Resistant)
Stage 4: Extraction Rule: Exclusive extraction of the highest n(label) per cluster
Step 1: Find largest n(label per cluster):
n_B(Class I) > n_A(Class I) > n_C(Class I)
n_A(Class II) > n_C(Class II)
n_C(Class III)
N(Cluster A) = n_A(Class I) + n_A(Class II) + n_A(Class III)
N(Cluster B) = n_B(Class I) + n_B(Class II) + n_B(Class III)
N(Cluster C) = n_C(Class I) + n_C(Class II) + n_C(Class III)
Stage 5: Remove Class/Cluster Labels
Round #1 : Uncluttered Data
Round #1: Isolate Mismatched Datapoints
Step 3: Isolate Residuals
Step 2: Extract Highest Importance Label First Cluster-exclusive extraction of largest Class I, then largest Class II, then Class III
Round #2 : Uncluttered Data
Round #2: Isolate Mismatched Datapoints
Round #r : Uncluttered Data
Round #r: Isolate Mismatched Datapoints
II
III
I
Stage 6: Construct New Imbalanced Training Set
Stage 7: Construct Balanced Training Set
Random Under-Sampling (RUS)
Re-sampling
New Imbalanced Dataset
Round #r
Retained Training Subset r
Round #2
Retained Training Subset 2
Round #1
Retained Training Subset 1
N(Training Subset 1) = n_B(Class I) + n_A(Class II) + n_C(Class III)

Fig. 4.5 The workflow of Minority Pattern Reconstruction (MPR)

In Figure 4.5, we apply simple K-means (k = output class labels) to cluster the blind TPP view, assuming k=3, the output of k-means will be class A, B and C. In each round, the instances within the clusters (A, B and C) are given their original labelled. For example, assume that class 1 represents the minority class of highest importance, the class of interest (i.e., Hyper-susceptible subjects), followed by class 2 (i.e., Susceptible subjects). Class 3 of the least of importance (i.e., Resistant subjects), each cluster is examined to find the highest count of Class 1 labels within; once such a cluster is found, Class 1 labels are extracted, the extraction process is cluster-label exclusive, where once a cluster was found to be the domain of label extraction, such a cluster is excluded from further examination for the extraction of the second label of importance and so on until a sample of data is retained. This retained sample is likely to show perfect classification across all classifiers.

Once all labelled patterns are exclusively extracted, the labels from the remaining points (residuals) are separated. Then the residual records are clustered in a new cycle of clustering-based pattern extraction, and the whole process is repeated until a threshold is met. Finally, all the instances in the retained sets are appended into one training set to undergo Random Under-Sampling (RUS). We hypnotise that the bias towards extracting the class of importance first with the largest count of instances in that class in a specific pattern would force a priority of learning on the classifier when fitting the training data.

M) Data dimensionality decisions: We presented an overview of feature selection methods and the practical debates in the literature with some pros and cons. In our view, whichever feature selection method is used to formulate Octopus Mantle will likely induce a possible limitation in the modelling process.

In addition, we acknowledge the challenge where the structure of the modelling problem at hand could evolve with time or may differ by population, causing "non-stationary" or “heterogenous” data. However, dealing with such a challenge is considered a future work opportunity that may form an extension of our approach.

We explained earlier that this challenge could introduce systematic differences between the training and test sets rendering some predictive models useless. Nevertheless, feature selection could also be a double-edged sword. Models also may be incorrectly specified and vary by modeller biases or arbitrary choices of features. In this instance, an illusion may occur, thinking that the changes in the external samples negatively impact the predictions made. Or, perhaps, the actual reason could be that the model missed a critical predictor and/or included a confounded predictor.

Therefore, to combat both of the above scenarios and while being guided by domain experts' advice, we choose to include all logically applicable useable features (i.e., those with tolerable missingness) in the model, regardless of their definition of relevance by the modellers. This decision may be computationally expensive, but we believe that including all variables could reduce the chances of excluding a current variable that may turn into a confounding predictor due to one of the above scenarios. Thus, this approach may help narrow down the cause of future prediction performance issues during monitoring. Model monitoring is not covered in this thesis.

We provided multiple reviews of the desired sample size and the required overall data dimensionality for modelling. We presented research that supported the *Events Per Variable (EPV)* and the *Samples per Feature Ratios (SFR).* In addition, other literature is found to oppose these ratios, stating that there is no golden rule to follow for every problem.

We also presented the researchers' concerns regarding unique studies where the EPV ratio is constantly violated. Also, we highlighted the debate of the model's reliability; where a perfect or near-perfect classification performance is achieved on both the training data and an independent validation test set, the results could be reliable, indicating a robust classifier. And based on these scientific debates, our choice is to include all features and available instances on their availability in modelling, regardless of any statistical association between the counts of features and examples.

Following from our previous recommendation, the reliability of a practical project output in the eyes of the domain experts is subject to mitigating a trade-off between two key considerations. These are high *generalisation* power, where unknown data points are classified correctly and meeting the required medical/biological *interpretability*.

We tackle the trade-off challenge by achieving some mutual trust between humans and machine learning. One approach that could influence such a trust is the informed involvement of clinicians and physicians in collecting, cleaning, declaring, formatting, preparing, engineering and modelling the data, followed by their evaluation of the models. A skilled data scientist should explore an important concept in human psychology: trying something new requires courage, vanishes fears and expands the human mind to learn more in a new area.

Although this approach may seem to have full domain experts' involvement, the entire data mining activity processes are always governed by the data scientist from the point of data understanding. The primary task for the data scientist here is not limited only to modelling but, more importantly, bridging the knowledge gap between the domain experts and data science. In turn, this may enhance the domain experts' confidence in the built models shifting their evaluation focus away from the model's complexity and more in favour of generalisation performance. Producing any additional indicators of the features' behaviour in the model will become satisfactory. Thus, priority of building classifiers with high generalisation power becomes dominant over the cost of achieving medical/biological interpretability. And the data mining tasks become manageable.

Including all applicable usable predictors may require scoping specific implementations of modelling algorithms known to provide robust behaviour against issues arising under these relaxed trade-off conditions, such as multicollinearity and overfitting. Thus, full features' inclusion has limitations. Therefore, we do not rule out the potential need for feature selection and data resampling methods. In addition, we do not have answers for the feature selection-resampling paradox. "Should the features be selected before or after the resampling techniques?" The answer to that question is empirically addressed by Ramos-Pérez et al. earlier this year [332].

N) Training-test split and validation considerations. In Chapter 3, we reviewed common approaches to performing training-test split for training and evaluation (testing). Here, this chapter acknowledges two scenarios of how these sets originate. In one scenario, the training and test sets are resampled from one big dataset, where the complete set is made available. In the other scenario, in some cases, the evaluation test set is extracted and made available later.

In the data science community, there are different recommendations for approaching training-test splits. Some suggestions are based on ratios relevant to the size of the dataset at hand. Others addressed the problem away from the relevance of records count in each class, relying on a solid dependence on the complexity of the learning process.

An article was presented in our literature review on this topic determined that 32 different adjustable parameters from the learning process play a role in determining the split ratio. In other words, the splitting depends on the capacity of a classification model, which is related to how complicated the model can be. Therefore, different machine learners require different training-test set splits ratios [231].

When considering the appropriate split ratio for our Octopus methodology, we see two issues of performing varied split ratios. First, while it is not a problem in particular that machine learning algorithms are trained with different subsets of the whole data, one issue is that the training performance is not an ultimate evaluator of models' generalisation performance, even if accompanied by cross-validation. Another problem is that the training performance of multiple classifiers built on different data subsets, i.e., resampled sets, cannot be fairly compared since different instances have a distinct influence on tuning the classifier's learning parameters. Thus, the classifiers would not have been equally challenged with the same instance hardness thresholds. This also could mean that cross-studies out there trying to solve the same domain problem cannot be relied upon for benchmarking, and oddly enough, every data mining problem may seem new.

We believe only a single constant test dataset can achieve a fair comparison. Therefore, a comparative evaluation of multiple classifiers is the key issue; thus, a fair comparison requires the same test instances across all trained models. We also acknowledge that when splitting the training and test data, there are two competing concerns, with fewer training samples, the parameter estimates may have greater variance. And with fewer test instances, the classifiers' performance evaluation may have greater variance.

To balance doubts in bias-variance trade-offs, we suggest proceeding with an equal training-test split ratio of ~1.0 when the whole data is available at once. And strive towards the same training-test balance where the test data is made available later, although this is not guaranteed. Our decision follows a similar recommendation by Pawluszek-Filipiak and Borkowski of having the training–test ratio around 1 (with the training portion should be as large as the testing data or very close) [235].

In addition, to provide further assurances on the quality of the classifier's learning, we recommend using a resampling approach to validate (tune) the classifier; the out-of-sample testing approach is reported to estimate how well the classifier generalises on test data. Based on the literature overview in the previous, we recommend using the K-Fold Cross-Validation resampling technique in training, with the most popular *k* value used in applied machine learning equal to 10. This recommendation follows findings from multiple studies reviewed in Chapter 3. *k* =10 was suggested to provide a good trade-off of low computational cost and low bias in estimating model performance [239].

O) Machine learning algorithms selection. Selecting a good-performing machine algorithm is challenging, primarily if the dataset was not modelled in the past and no performance evaluations are available as indicators. Having tens, if not hundreds, of machine learning algorithms to choose from intensifies the problem. But unfortunately, when approached with a new classification problem, there is a tendency in the data science community to use the so-called favourite algorithm such as Random Forest, Xtreme Gradient Boosted Trees (XGB), etc.

Another strategy is scouring the research literature for descriptions of similar problems and attempting to reimplement the algorithms with their configurations without guaranteeing that the domain experts would approve the reported performance metrics for evaluating the new classification problem.

In Octopus, we target the lack of effectiveness in either strategy since they can be hit-or-miss or time-consuming. Furthermore, we recommend that literature on the domain problem be only used to identify techniques and ideas to try, not for reimplementation or comparison purposes.

On a high level, to handle the variety of available techniques effectively, we suggest taking a shortcut by identifying an acceptable model by domain experts on new non-attempted classification problems. Therefore, we recommend systematically evaluating a suite of standard machine learning algorithms of different learning mechanisms to establish which performs well across multiple data preparation strategies. Once the data is prepared, we obtain a baseline performance on the original data samples across the suite of algorithms. By doing so, any specific learning issues will surface from the first modelling run.

The next modelling cycle is then tailored with candidates of strategies, such as data sampling, cost-sensitive learning (as suggested earlier), hyper-parameter tuning, etc., hoping to produce specialised models outperforming the baseline. These models can be presented to domain experts to choose from based on their satisfaction with the empirical results.

One should also ensure that the selected evaluation metrics used in the comparisons translate the success criteria.

As for the algorithms' suite, there is no agreement in the machine learning community on classifying the types of learning. However, in our methodology, we suggest the suite should consist of a mixture of parametric and non-parametric classifiers. Thus, we suggest a probabilistic learner (Naïve Bayes), an instance-based algorithm (K-Nearest Neighbour), a hyper-plane classifier (SVM with SMO implementation), a linear curve fitting algorithm (logistic regression with ridge estimator), a non-linear curve fitting algorithm (Multi-Layer Perceptron), the non-linearity for MLP is driven by the use of multiple layers despite having the sigmoid function is used for neuron activation, a decision tree classifier (C4.5), and ensemble learners (Logistic Model Tree and Random Forest). Our recommendation is to model as many of these algorithms as possible on the same prepared datasets. It is also important to perform the training in the same machine-learning environment. This is because different environments may have different implementations of the same algorithms and randomisation settings, producing varied results.

P) Success criteria and evaluation metrics. Selecting an evaluation metric is the most critical step in a classification problem. The selection of adequate metrics is the foundation for comparing the built models and establishing their abilities and limitations to full fill the predefined success criteria by the domain experts.

The metric must capture the details about a model that are most important to the domain experts. Choosing the unsuitable metric can alter the choice of models, leading to selecting unfit models to full fill the success criteria of the problem at hand. However, they may be suitable for solving different problems. Scoping suitable evaluation measures is challenging due to the variety of performance metrics; as seen in our literature review, each has its own interpretation of the model's performance. This variety of metrics often leads to domain experts' uncertainty about which metric to choose. Therefore, in our Octopus methodology, we do not recommend selecting metrics before forming competitive success criteria as a heuristic for all models to work towards.

Sometimes, the interpretation of such criteria can be covered by one or more metrics. Thus, due to the variation in success criteria among domains, it is difficult to scope metrics that suit every classification problem. Therefore, once the criteria are formed, the data scientist should assist the experts in finding the metrics that best interpret their desired output of interest.

The success criteria could be precise, like observing a specific predictive performance of a particular group; hence the confusion matrix comes in handy or can be more general and interpretable with fundamental or combined metrics.

Heuristically speaking, as discussed and presented in the literature review in Chapter 3, some metrics are proven effective interpreters of a model's success in specific circumstances, such as imbalanced learning, i.e., Balanced Accuracy, Youden's Index, and the G-mean. Some graphical metrics are also popular in the medical domain but may be misleading in certain conditions, such as AUC. Some are voted to be balanced metrics, such as the MCC. Some domain experts may seek the best balance between TRP and TNR. Hence, the term 'best model' is subjective and could have many interpretations. Therefore, in our Octopus methodology, we consider the confusion matrix, including its fundamental metrics as the go-to metric, and from there, we proceed systematically with scoping various combined and graphical metrics to interpret models' performances based on defined success criteria by real-world life sciences and healthcare experts.

In addition, we will also perform a proof-of-concept ranking of multiple binary classification models using our newly adapted graphical evaluation metric, *XDistance*.

XDistance is a new proposed graphical performance metric that ranks different models in terms of inaccuracy and balance. It allows for a fast comparison of multiple models. In Figure 4.6, we replace the FAR with FPR and the FRR with FNR curves. The cross point between the FPR and the FNR is the new CER. The x-axis is the threshold of the conventional ROC curve, and the y-axis is the error rate.

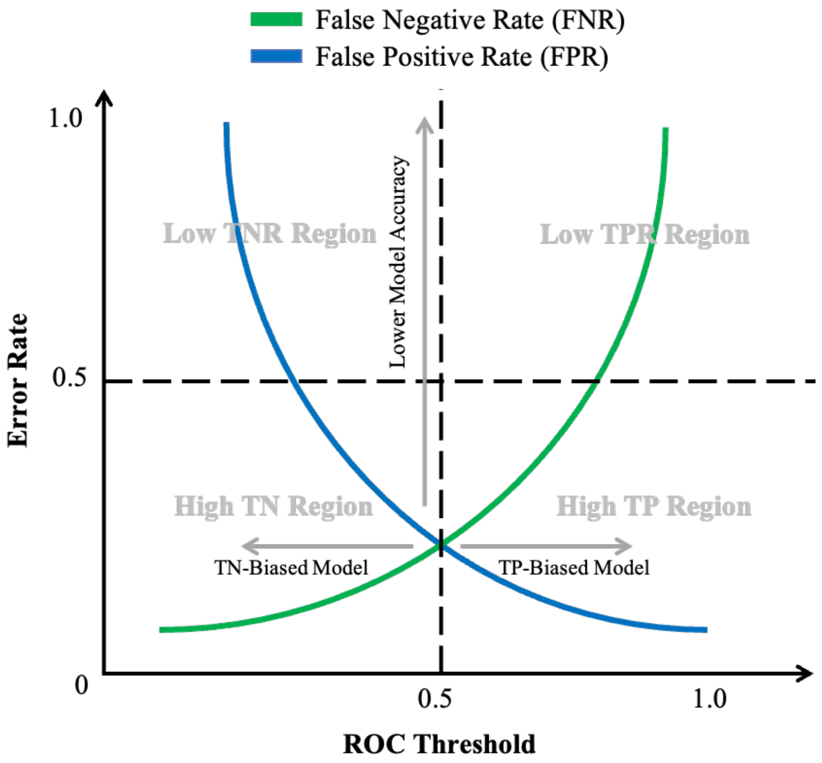


Fig. 4.6 Illustration of the newly adapted Cross Error Rate (CER)

Unlike the usability decision being made on the CER based on the y-axis direction (the lower the CER, the better), we test in imbalanced data classification, the cross-point (X) shift into both directions, threshold and error rates.

Therefore, for a classifier shifting in the direction of the error rate, the closer X from the threshold line, the more accurate classifier. However, the shift of the crossover point (X) in the direction of the threshold line indicates the state of error balance that influenced the classifier's accuracy.

Each model typically produces a Crossover point. By observing the placement of the models' Crossover point in each model's CER plot (from testing), we hypothesise that X for the minority-biased resampled models would reside at the right side of the plot ($XThreshold > 0.5$). On the contrary, the behaviour of the crossover curves for the majority-biased models would highly drift to the left of the CER graph ($XThreshold < 0.5$). The CER curves behaviour corresponds to true-positive instances counts of 73 and 1 out of 96 total positives, respectively. Thus, a model would show less imbalance between TP and TN if $\boldsymbol{X}$ resides at $Xthreshold \approx 0.5$.

Furthermore, to overcome the potentially deceiving high accuracy in imbalanced learning. The new Crossover Error Curves allow for a simplified geometric graphical comparison by replacing FPR-FNR curves with their crossover X points. The plot comparison forms a potential opportunity to consider a new criterion for empirical performance evaluation. It can be called a graphical **X-Matrix** (see Figure 4.7).

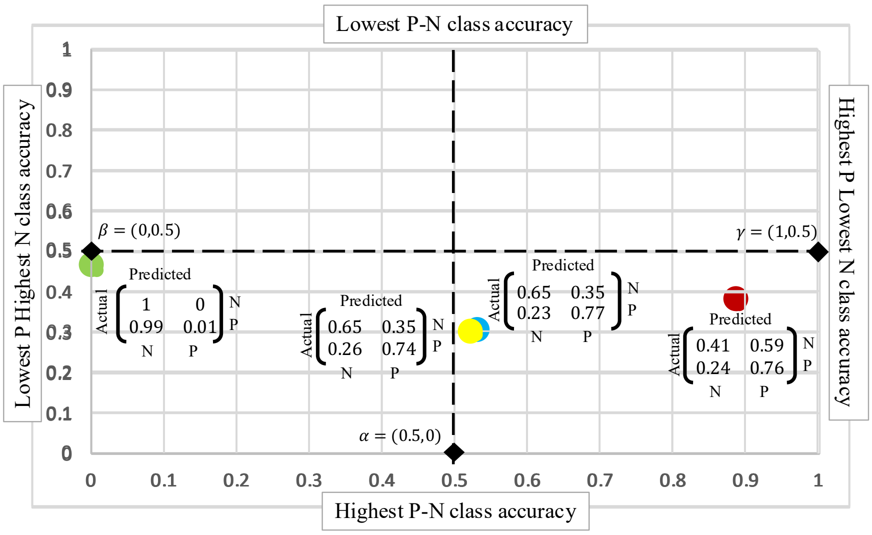


Fig. 4.7 Graphical illustration of X-Matrix for balanced and imbalanced tested models

From Figure 4.7, we assume that based on the Euclidean Distance for each model δ, the shorter the XDistance, the more balanced the model is. Hence the XDistance may indicate a balanced performance ranking metric. The model of balanced performance TP-TN can be extracted from the X-Matrix by calculating the model's geometric XDistance of a model (δ) to the XThreshold reference point $\alpha(0.5,0)$.

For a model's XThreshold $\boldsymbol{x_i}$ and CER $\boldsymbol{y_i}$, the geometric X-Distance can be calculated using the Euclidean Distance $\delta(x_i, y_i)$ by:

$$d_{Eucl}(\delta, \alpha) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2} = \sqrt{(x_i - 0.5)^2 + (y_i - 0)^2}$$

$$XDistance = \sqrt{(XThreshold_i - 0.5)^2 + (CER_i)^2}$$

An indicative qualitative sign (±) could be attached to the XDistance to indicate the models' class bias based on the XThreshold value x_i of 0.5:

$$\text{XDistance:} \begin{cases} is\ (+) & \forall\ x_i > 0.5\ \textit{Positive biased model} \\ is\ (\pm) & \forall\ x_i = 0.5\ \Leftrightarrow \textit{Balanced model} \\ is\ (-) & \forall\ x_i < 0.5\ \Leftrightarrow \textit{Negative biased model} \end{cases}$$

4.2.3 Structured applications at Octopus head

This part of the Octopus methodology is the part which validates the effectiveness of the researched methods at the arms and scopes at the mantle, forming the framework for practical applications in data-driven preventive healthcare. When applied, it usually contains a structured approach to applying such methods to healthcare and life sciences case studies.

These methods should collaborate to perform predictions of adverse health events to benefit preventive medicine. In one case study, the predictions of adverse events of having excess amounts of Visceral Fat could be vital in our lifestyle adjustments and treatments to detect and prevent potential health complications. Although lifestyle modifications should be a collaborative effort between physicians, health professionals and the patient and sound natural and medically independent for many, the reality is that some medical or surgical interventions might be required to support, speed up or sustain these changes.

In relation to obesity, treatment-independent lifestyle modifications are numerous, including diet, physical activity, weight reduction, smoking, and stress. Nevertheless, lifestyle modifications may include medical interventions, such as prescribed weight loss drugs, or even surgical procedures, such as weight loss surgeries (i.e., gastric banding). In many cases, weight loss surgery isn't for everyone who is overweight or even obese. Surprisingly, to qualify for such surgical intervention, the patient must have already developed and diagnosed chronic health complications related to obesity, such as heart disease, diabetes, or high blood pressure [324].

Octopus methodology governs building new tools for the early detection of susceptible individuals to such diseases who are in most need of proactive lifestyle modifications before developing such chronic diseases. This early detection should help individuals to follow self-induced lifestyle changes. In addition, it indeed triggers different considerations for physicians when offering preventative medical and surgical treatment. It provides both primary and secondary prevention improving large-scale interventions for the population's overall health and provides an inexpensive routine screening at the patient level. This is a case where our new machine learning models step in to take action to assist in growing a healthier community.

In the second case study, Octopus goes beyond proactive lifestyle adjustments with a similar collection of methods to govern building new clinically valid models. These models are approved by a collaboration of national and international scientists. These new models can potentially then be used as new screening tools to predict those who may develop severe side effects to certain types of treatment, namely advanced cancer radiotherapy in breast cancer patients. The new models aim to help flag potential impairment to patients' Quality of Life (QoL).

In the radiation therapy case, some cancer patients will likely experience various side effects on the severity scale during radiotherapy. A more specific scenario is observed in some treated breast cancer patients as they develop severe skin toxicity leaving them with a lifelong lasting impact on their lives. The current protocols in advanced radiotherapy include positioning, imaging, radiation doses and fractions so that the treatment is effective and safe. Still, they do not account for specific acute side effects such as severe pain, ulcers, scars, liquid build-up under the skin, or other potentially life-threatening conditions.

Sometimes, the protocols leave radiotherapy physicists considering two options: either continuing the treatment while their patient endures such acute side effects or interrupting the therapy. However, if interrupted, the cancer cells, after a halt, may become agitated, increasing the risk of cancer recurrence or even with a faster spread, reducing survival rates [325]. Our models give radiotherapy clinicians the capability of early patient screening. They are a new line of help to initiate discussions with the susceptible patients to such an adverse event about other options available to them.

Chapter 5

Octopus Framework Application for Visceral Fat Associated Diseases Prediction

This chapter covers

- *Case Study 1: Predicting patients' susceptibility to visceral fat-Associated Diseases.*

This chapter applies our OCTOPUS data-driven framework concluded from our research in the previous chapters to develop new predictive models as new tools to help detect patients' susceptibility to adverse events in healthcare, Visceral Adipose Tissue (VAT) associated diseases. We use the methods in the novel framework with trivial differences due to the nature of the data in this case study, the domain success criteria, and the execution timeline. These new tools are built using new real-world datasets. Our models are new in the field. At the time of execution, the modelling of the endpoints was not attempted by other case studies in the field using our data or predictors.

* **Disclaimer:** *The author and collaborators assume no responsibility or liability for any errors or omissions in the data provided by the UK-Biobank. The data is provided from their repositories on an "as is" basis with no guarantees of completeness, accuracy, usefulness or timeliness. The authors assume no liability for the reliability of the results obtained from the use of this data. All data in these studies were engineered and modelled from the provided "as is" data at best with collaborators' review. This practical work was accomplished by the author and collaborators in their personal academic and professional capacities. The opinions expressed in this study are their own and do not reflect the view of the UK Biobank or the University of Westminster in the United Kingdom.*

** **Disclaimer**: *This chapter reuses the published content of the following articles:*

❒ *Aldraimli, M., Soria, D., Parkinson, J. et al. Machine learning prediction of susceptibility to visceral fat associated diseases. Health Technol.* ***10****, 925–944 (2020). https://doi.org/10.1007/s12553-020-00446-1*

❒ *Aldraimli, M., Soria, D., Parkinson, J., Thomas, E.L., Bell, J.D., Dwek, M.V. and Chaussalet, T.J., 2020. Machine learning prediction of susceptibility to visceral fat associated diseases. Health and Technology, 10(4), pp.925-944.*

Our focus in this chapter is to use the methods in our OCTOPUS framework, formulated by our heuristic-systematic research and critique, to demonstrate its ability to create new inexpensive screening methods with Machine Learning approved by the life sciences and healthcare domain experts.

This case study models individuals' susceptibility to visceral-fat-associated diseases, with two distinct cases, on females and males. Both studies are a tri-class nominal classification problem. Thus, the models are developed to predict three susceptibility class labels: Healthy subjects, those who can be described as resistant to developing such diseases; individuals with moderate risk (susceptible subjects) with a medium likelihood to develop Visceral Adipose Tissue (VAT) associated diseases; and finally, the high-risk group who can be considered hyper-susceptible (highly-likely) to developing chronic health conditions related to high amounts of VAT.

Females and Males cohorts are modelled separately due to the differences in the labels' endpoint definitions. This labelling difference ought to show the physiological composition differences for each gender. Visceral fat or visceral adipose tissue (VAT) is 'hidden' fat stored deep inside the belly, wrapped around the organs, including the liver and intestines. Figure 5.1 illustrates the different types of adipose tissue (AT) in the human body and the increased susceptibility to adverse diseases associated with increased amounts of visceral fat in the human body.

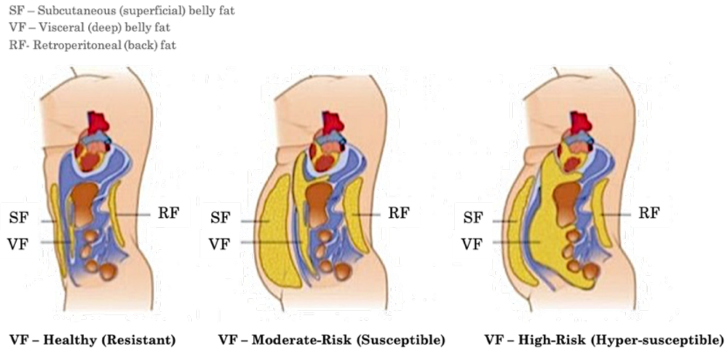

SF – Subcutaneous (superficial) belly fat
VF – Visceral (deep) belly fat
RF- Retroperitoneal (back) fat
SF
VF
RF
SF
VF
RF
SF
VF
RF
VF – Healthy (Resistant)
VF – Moderate-Risk (Susceptible)
VF – High-Risk (Hyper-susceptible)

Fig. 5.1 Illustration of visceral fat levels and associated susceptibility to diseases

The problem of VAT in the body cannot always be visible to the naked eye. Individuals can be slim on the outside and thick on the inside, with dangerous visceral fat deposits (see Figure 5.2). This phenomenon is known as Thin on the Inside and Fat on the Outside (TOFI). Doctors can examine the body's composition thanks to Magnetic Resonance Imaging (MRI). MRI images reveal how much 'internal fat' even slim people carry and raise questions about how healthy people are. Doctors recently are increasingly concerned that people can look slim on the outside but still have a problem with fat [333]. Performing an MRI does not come cheap. Therefore, this method is not pragmatic to screen a whole population. Thus, we apply our new data-driven framework, OCTOPUS, to create new data-driven tools to offer an affordable pragmatic and potential substitute to MRI scanning for this specific problem.

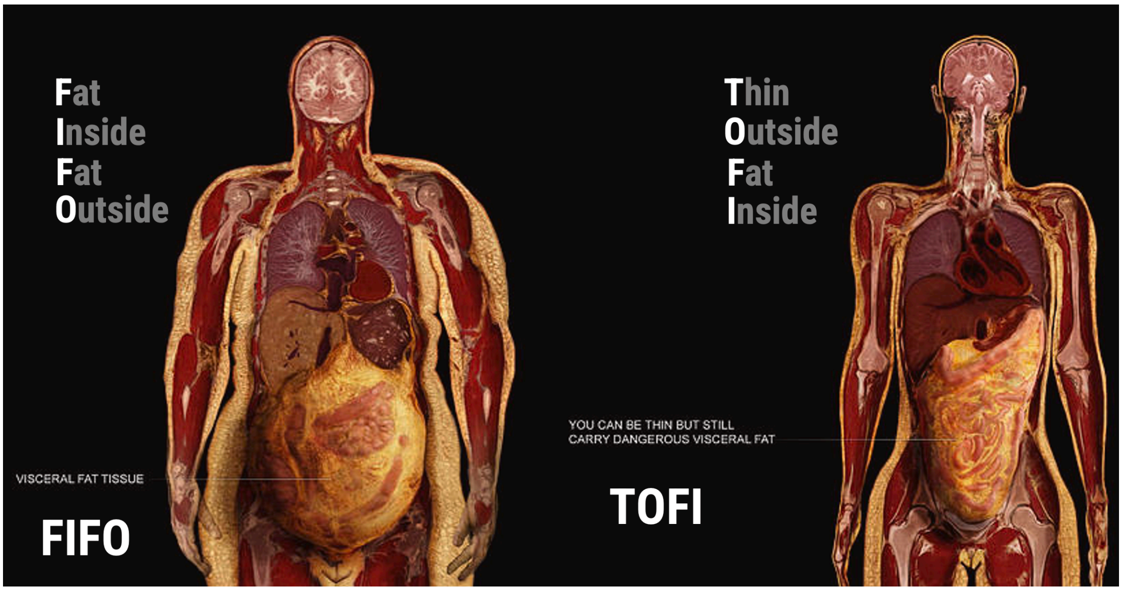


Fig. 5.2 Coronal plane illustration of Thin on the Outside Fat of the Inside (TOFI)

5.1 Introduction

The deployment of machine learning modelling in this case study aims at tackling a long-term real-world disease burden; Obesity affects an increasing number of adults in the UK [334], with obesity-associated changes in adipose tissue (AT) predisposing to metabolic dysregulation [335] and other disorders. Distribution of AT, in particular the accumulation of visceral adipose tissue (VAT) and liver fat, is a critical factor in determining susceptibility to diseases [336] [337]. Excess VAT and liver fat play a significant role in the pathogenesis of type 2 diabetes, dyslipidaemia, hypertension and cardiovascular disease [338]. Current strategies for treating obesity and its associated co-morbidities have focused on lifestyle improvements [339] [340]. Such a focus aims to reduce VAT and liver fat via calorie restriction with or without exercise. The impact of this is associated with improved insulin sensitivity, decreased blood pressure and lower circulating lipid levels [335] [341] [342]. However, large-scale analysis of the compartmental distribution of AT is often limited due to the expense and time required to employ the requisite imaging techniques.

The UK Biobank (UKBB) provides comprehensive means of assessing the relationship between body composition and lifestyle in a large population-based cohort of adults. Having such a large dataset could increase the presence of a pattern in the data. Without it, machine learning algorithms can't sufficiently learn to produce effective results. The primary goal of these two case studies is to identify the best classification models as new tools that can be linked to applications for inexpensive early screening of three VAT levels. This way, a subject's susceptibility to developing potential chronic adverse health conditions such as non-communicable diseases is detected. The prediction is performed on females and males. A positive detection prompts the need for lifestyle modifications.

This case study is a cross-sectional assessment of individuals from the UKBB multi-modal imaging cohort [343], aged 40-70 years and scanned chronologically between August 2014 and September 2016.

Here we apply our formulated data-driven framework. Then we assess the classification performance of six machine learning algorithms (Naïve Bayes, Logistic Regression, Artificial Neural Network, Decision Tree, Logistic Model Tree and Random Forest) in predicting discretised visceral fat ranges associated with the susceptibility of developing long-term diseases. The dataset is class imbalanced, which may cause classifier learning issues.

The new models are built using the Imbalanced Dataset. In addition, we apply two sampling techniques, Random Under-Sampling (RUS) and Synthetic Minority Over Sampling Technique (SMOTE). Both methods are used to resample the highly imbalanced training data (in the female cohort case) and the less severe class imbalance (in the male cohort case) to achieve a state of class balance.

Each case suggests the most suitable models meeting the domain experts' success criteria. The data imbalance characteristic causing the transition in classifier training performance was captured visually by Adaptive Projection Analysis (APA) in the Targeted Projection Pursuit (TPP) and numerically via Information Gain (IG) attribute evaluation.

5.2 Predicting females' susceptibility to visceral fat-associated diseases workflow

Three-class nominal classification models were applied for VAT prediction to predict susceptibility (risk) to adverse diseases based on the discretised amount of VAT. A group of 2292 female subjects was used to train eight ML algorithms using 10-fold cross-validation in three different scenarios. In relation to their cohort, the trained models were tested on a new group of external data of 2035 female cases. With training – test ratio ~1 (57:50), Figure 5.3 shows the study workflow: multiple imbalanced datasets with the same predictor variables were modified with sampling techniques and used for modelling using the six ML algorithms. Selected performance metrics of the models were compared after training in the evaluation phase. IG was monitored for all predictor variables at every stage.

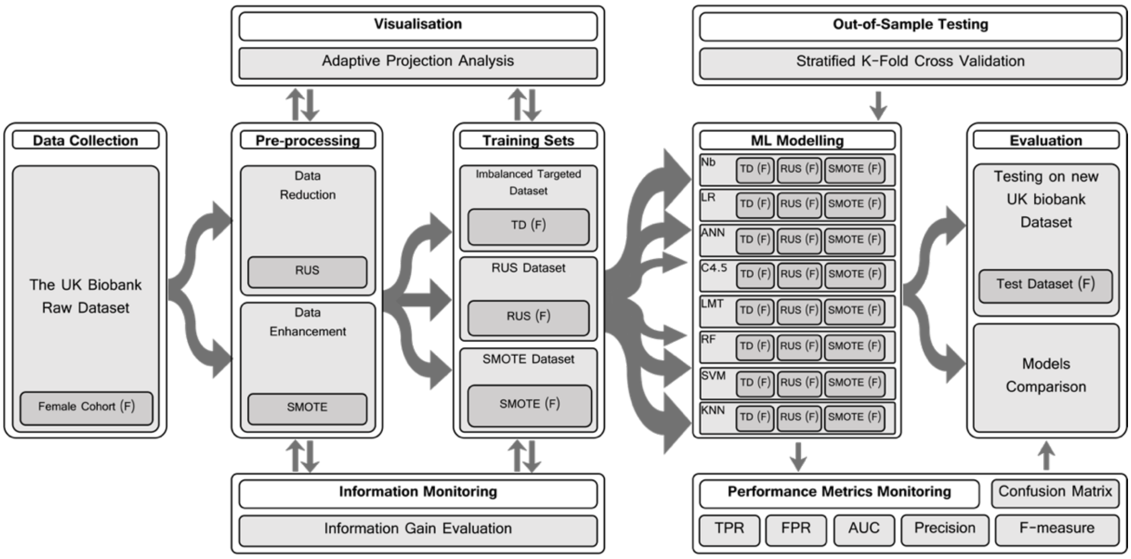


Fig. 5.3 The females' VAT case study methods and workflow, showing the used techniques Where TD = Targeted dataset, RUS = Random Under Sampling, SMOTE = Synthetic Minority Oversampling Technique, ML = Machine Learning, NB = Naïve Bayes, LR = Logistic Regression, ANN = Artificial Neural Network, C4.5, LMT = Logistic Model Tree, RF = Random Forest, TPR = true-positive rate, FPR = false-positive rate, AUC = Area under receiver operator characteristic curve

5.2.1 Data collection protocol

This cross-sectional study includes data from 4327 females in the UKBB multimodal imaging cohort. The UKBB had approval from the North West Multi-Centre Research Ethics Committee (MREC), and written consent was obtained from all participants before their involvement. The data was acquired through the UK Biobank Access Application number 23889. The age range for inclusion was 40-70 years, with exclusion criteria being: metal or electric implants, medical conditions that prohibited MRI scanning or planned surgery within six weeks before the scanning date. The subjects were scanned chronologically between August 2014 and September 2016. The visceral adipose tissue (VAT) volumes were acquired as part of the UKBB dataset.

Anthropometry measurements were collected at UKBB assessment centres; height was measured using the Seca 202 height measure (Seca, Hamburg, Germany). The average of two blood pressure measurements, taken moments apart, was obtained using an automated device (Omron, UK). Images were acquired at the UK biobank imaging Centre at Cheadle (UK) using a Siemens 1.5T Magnetom Aera. The participants' height and weight were recorded before imaging screening which later was utilised to calculate the Body Mass Index (BMI).

For physical activity assessment data, a touch screen questionnaire was used to collect information on sociodemographic characteristics and lifestyle exposures (http://www.ukbiobank.ac.uk /resources/). Specific questions on the frequency and duration of walking (UK biobank field ID: 864, 874), moderate physical activity (884, 894) and vigorous physical activity (904, 914) events allowed the calculations of metabolic equivalent-minutes per week (MET-min/week) for each individual. Participants were excluded from the calculations and analysis if they selected 'prefer not to answer' or 'do not know' to any of the possible six questions on physical activity used to calculate the MET score.

5.2.2 Experimental design

VAT-related disease susceptibility is based on the following MRI response labels: Healthy (Resistant), Moderate (Susceptible) and Risk (Hyper-Susceptible) defined according to VAT volume. In females, a VAT volume of ≤2 litres was deemed 'Healthy' (H); VAT volume >2 litres but ≤5 litres was classed as 'Moderate' (M); VAT volume >5 litres was classified as 'Risk' (R) [344]. The training datasets contained ten data variables reported in Table 5.1, with the VAT in litres being the class determination response variable. All nine predictor variables in Table 5.1 were selected as input features by domain experts based on their low cost to obtain and associations with VAT prediction in previous studies.

The choice of a limited number of anthropometry and physical activity variables is also driven by the desire to produce a model with inexpensive reading collected from the subjects. The new model can be integrated into a new application for screening.

Table 5.1. Descriptive statistics of variables in the Targeted Dataset (TD)

Female Cohort (n=2292)			
Numeric selected dataset variables	Median	Mean	(Min, Max)
Response variable			
Visceral adipose tissue volume (VAT in litres)	2.2	2.5	(0.1, 9.7)
Predictor variables			
Waist Circumference (WC in cm)	80.0	81.6	(55.0, 126.0)
Pre-imaging Weight (W in Kg)	66.0	68.3	(42.0, 128.0)
BMI (in kg/m^2)	24.8	25.7	(15.5, 48.0)
Hip circumference (HC in cm)	100.0	100.9	(77.0, 147.0)
Standing height (H in cm)	163.0	163.0	(141.0, 194.0)
Systolic blood pressure (SBP in mmHG)	133.0	134.5	(87.0, 225.0)
Diastolic blood pressure (DBP in mmHG)	77.0	77.8	(45.0, 120.0)
Physical Activity Index (PAI)	0.5	0.6	(-12.0, 15.5)
Age at recruitment (AGE in years)	55.0	54.6	(40.0, 70.0)

5.2.3 Physical activity index (PIA) feature construction

The UK Biobank Physical Activity Index (UKBB PAI or PAI) was created by domain experts [345] using data collected during physical activity assessment; comprising a total of 27 outcomes, 23 outcomes reflecting activity and four reflecting inactivity (see Table 5.2). An individual's response to questions was scored with values between -1 and +1 and combined cumulatively to give a final score with an increasingly negative score implying a progressively unhealthier phenotype. For binary variables, 0 indicates the absence of the parameter, 1 the presence.

Table 5.2. UK Biobank outcomes used in creating the physical activity index

UKB ID	Outcome	Units
816	Job involves heavy lifting	Categorical
864	Days/week walked 10+ minutes	Days/Week
874	Duration of Walks	Minute/Day
884	Days/week moderate physical activity 10+ minutes	Days/Week
894	Duration of moderate activity min	Min/Day
904	Days/week vigorous physical activity 10+ minutes	Days/Week
914	Duration of vigorous activity	Minute/Day
924	Usual walking pace	Categorical
943	Frequency of stair climbing in last 4 weeks	Categorical
971	Frequency of walking for pleasure in last 4 weeks	Categorical
981	Duration of walking for pleasure	Categorical
991	Frequency of strenuous sports in last 4 weeks	Categorical
1001	Duration of strenuous sports	Categorical
1011	Frequency of light DIY in last 4 weeks	Categorical
1021	Duration of light DIY	Categorical
2624	Frequency of heavy DIY in last 4 weeks	Categorical
2634	Duration of heavy DIY	Categorical
3637	Frequency of other exercises in last 4 weeks	Categorical
3647	Duration of other exercises	Categorical
6164	Types of physical activity in past 4 weeks	Categorical
104900	Time spent doing vigorous physical activity	Categorical
104910	Time spent doing moderate physical activity	Categorical
104920	Time spent doing light physical activity	Hours
806	Job involves mainly standing or walking	Categorical
1070	Time spent watching television	Hour/Day
1080	Time spent using computer	Hour/Day
1090	Time spent driving	Hour/Day

5.2.4 Inspecting irregular training examples

Inter Quartile Range (IQR) test is used to discover abnormal values with an Extreme Value Factor (EVF) of 6 times the IQR and an Outlier Factor (OF) of 3 times the IQR. A total of 10 abnormal values are verified and retained. All abnormal values were outliers; no extreme values were found within the definition.

These abnormal observations may be part of a phenomenon. Figure 5.4 shows the outliers and extreme values per variable in the imbalanced targeted dataset TD. Nine outliers are found in the moderate and risk groups, while one outlier in the height variable belongs to a healthy subject.

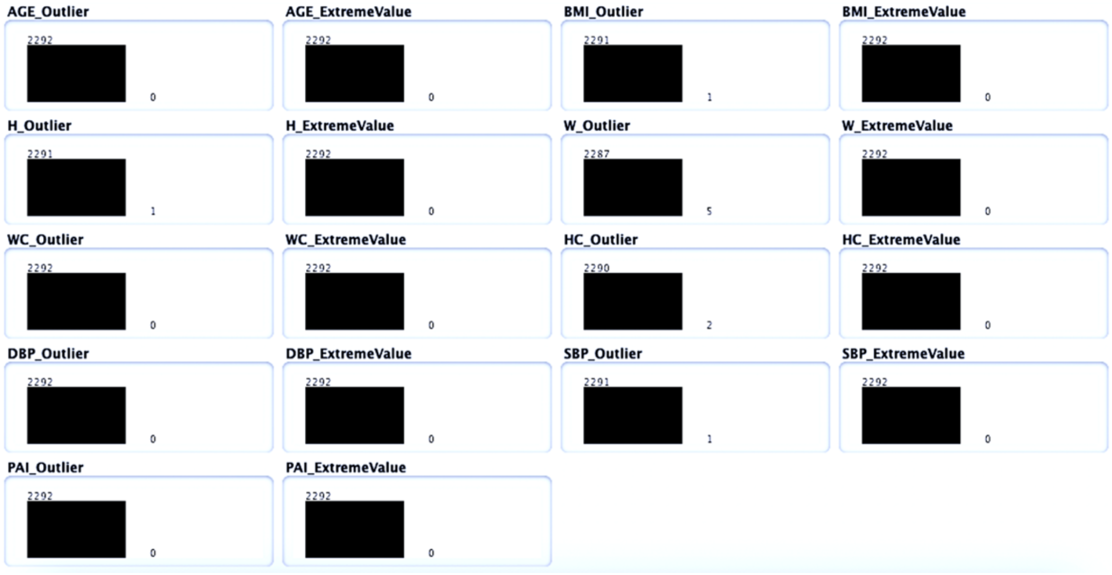


Fig. 5.4 Abnormal values detected in the imbalanced Targeted Dataset (TD)

5.2.5 Training instances resampling

The targeted dataset (TD) was the first dataset modelled. The TD contained 2292 female records from the UKBB cohort. Table 5.1 shows the summary statistics of all TD's variables. The TD was highly imbalanced in the female cohort in relation to records numbers per class: In the females' TD, class H had 1002 subjects, class M had 1128 subjects, and class R contained only 162 subjects. Random under-sampled (RUS) dataset is a reduced subset of TD. A subset of each majority class was randomly removed to balance the data. As a result of applying RUS to the females' TD, each of the H, M and R classes ended up with 162 subjects. Synthetic Minority Over-Sampled (SMOTE) dataset was obtained as a result of applying SMOTE to the numeric data variables of TD. By doing so, the three VAT classes became more closely balanced. In the female cohort, class H had 1002 subjects, class M had 1128 subjects, and class R contained 1296 subjects. The class imbalance of TD and the effect of RUS and SMOTE on TD can be observed in the TPP adaptive projection analysis (APA) visualisation in Figure 5.5.

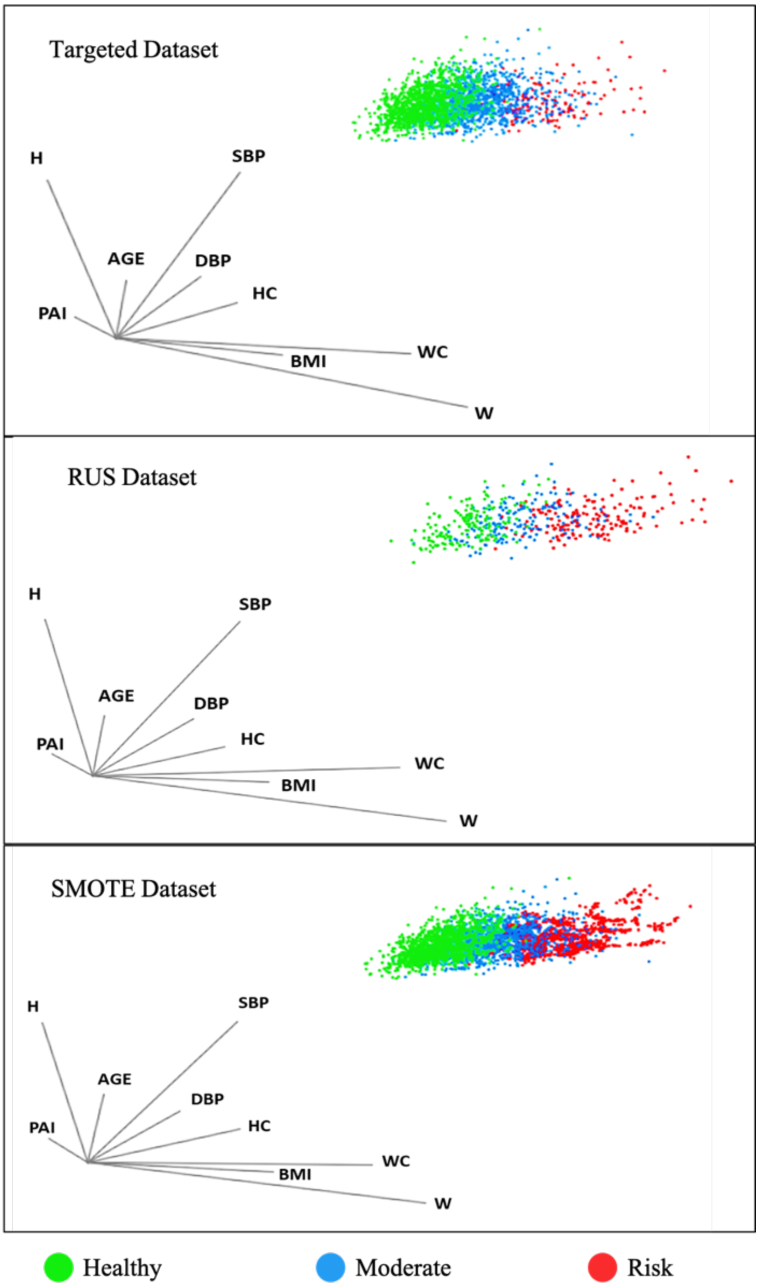


Fig. 5.5 Adaptive projection visualisation of Targeted Dataset (TD), Random Under Sampled (RUS) dataset and SMOTE dataset variables

5.2.6 Test data characteristics

The ML models were tested on a new group of 2035 females from the UKBB female cohort from the UKBB male cohort. Table 5.3 shows their summary statistics. Like the TD, the female Test Dataset was also highly imbalanced: class H had 823 subjects, class M had 1039, and class R contained only 173 subjects.

Table 5.3 Descriptive statistics of variables in the females' test set

Female Cohort (n=2035)			
Numeric test dataset variables	Median	Mean	(Min, Max)
Response variable			
Visceral adipose tissue volume (VAT in litres)	2.4	2.7	(0.2, 10.0)
Predictors variables			
Waist Circumference (WC in cm)	80.0	81.6	(55.0, 142.0)
Pre-imaging Weight (W in Kg)	67.0	68.7	(39.0, 136.0)
BMI (in kg/m^2)	25.2	25.9	(14.4, 54.5)
Hip circumference (HC in cm)	100.0	101.3	(73.0, 156.0)
Standing height (H in cm)	163.0	162.7	(145.0, 195.0)
Systolic blood pressure (SBP in mmHG)	129.0	130.4	(87.0, 196.0)
Diastolic blood pressure (DBP in mmHG)	76.0	76.6	(45.0, 115.0)
Physical Activity Index (PAI)	0.0	0.1	(-12.5, 18.0)
Age at recruitment (AGE in years)	55.0	54.6	(40.0, 70.0)

5.2.7 Inspecting irregular test records

In a similar approach to outliers' detection in TD, the Inter Quartile Range (IQR) test was used per test data attribute to look for defined values with an Extreme Value Factor (EVF) of 6 times the IQR and Outlier Factor (OF) of 3 times the IQR.

A total of 14 abnormal data values are identified, verified and retained, and no extreme values are found. However, these abnormal observations may be part of a phenomenon. Figure 5.6 shows the outliers and extreme values detection in the test dataset. Thirteen outlier data points are found in the risk examples, while one was in the height variable for a healthy subject.

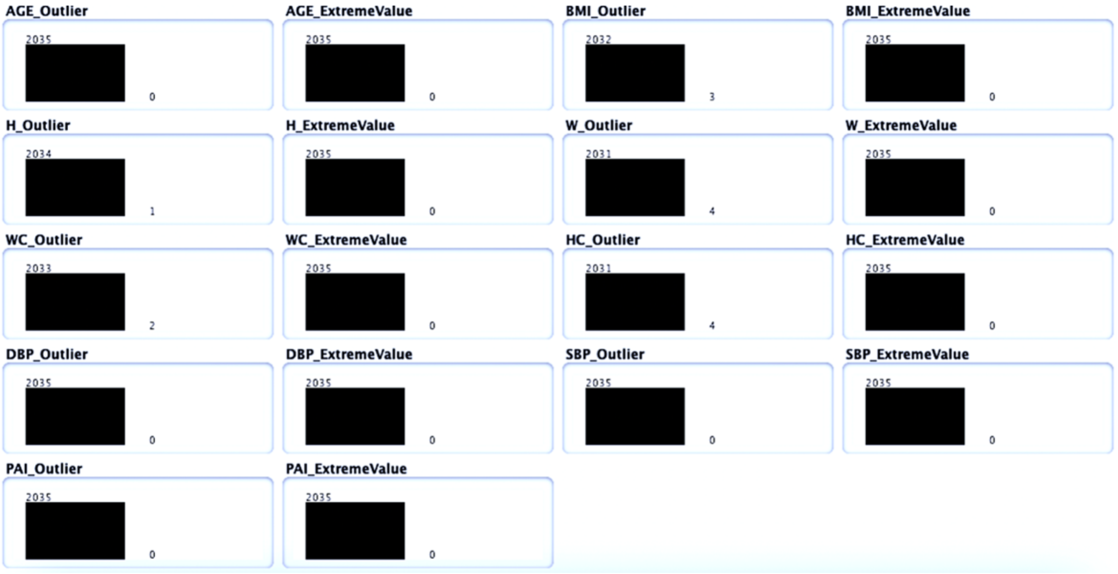


Fig. 5.6 Abnormal data values detection in the test set

5.2.8 Modelling females' susceptibility to VAT diseases.

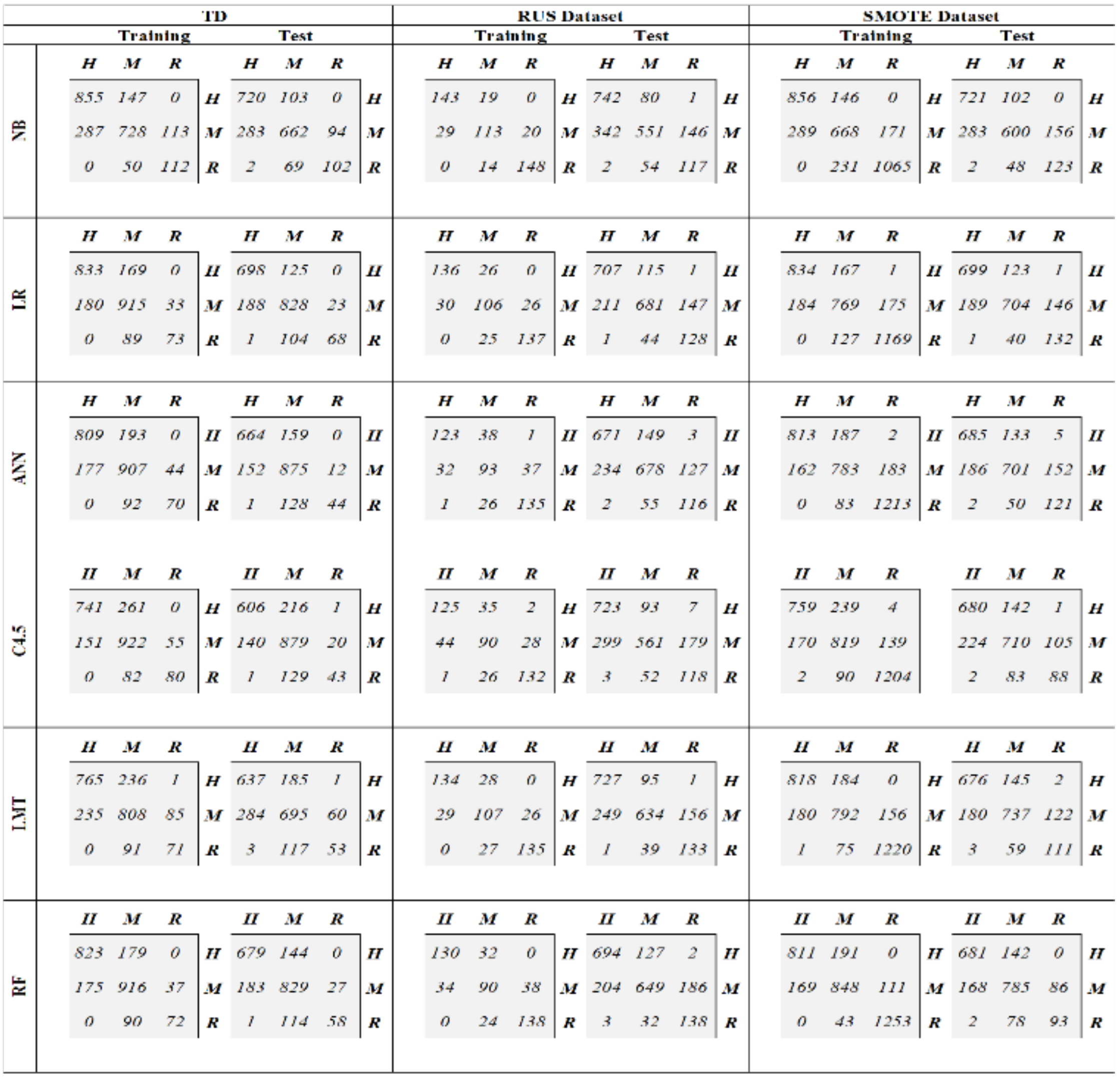
The classification modelling females' susceptibility to adverse health conditions associated with visceral fat uses six algorithms. The modelling algorithms are described in Chapter 4. And these are Naïve Bays (NB), Logistic Regression (LR) with Ridge Regularization, Artificial Neural Networks (ANN) with Multi-Layer Perceptron (MLP) architecture and built-in feature scaling (Range Normalisation), C4.5 Decision Tree algorithm, and ensemble learners Logistic Model Tree (LMT) and Random Forest (RF).

5.2.9 Models training results

From the confusion matrices in Table 5.4, the model training accuracies for the female cohort, presented as Correctly Classified Instances ratio (CCI) or True Positive Rate (TPR), of all methods, were computed, they showed that resampling methods resulted in an improvement in CCI compared to the original TD. When the performance of the LR, ANN, C4.5 and RF models for the female cohort was evaluated, it was apparent that the RUS dataset was poorer than when the TD data set was used, see Figure 5.7.

The AUC for each of the trained models were in the range of 0.783 (for RF on SMOTE) to 0.96 (for C4.5 on TD). These values indicate that the trained models did not sacrifice much precision to achieve a good recall value on the observed data points. The RF model achieved the highest TPR (0.850) when trained on the SMOTE dataset, while the C4.5 model achieved the lowest TPR (0.714) when trained on the RUS dataset.

Table 5.4. Female cohorts VAT prediction models training and test confusion matrices



	TD								RUS Dataset								SMOTE Dataset							
	Training				Test				Training				Test				Training				Test			
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
NB	855	147	0	H	720	103	0	H	143	19	0	H	742	80	1	H	856	146	0	H	721	102	0	H
	287	728	113	M	283	662	94	M	29	113	20	M	342	551	146	M	289	668	171	M	283	600	156	M
	0	50	112	R	2	69	102	R	0	14	148	R	2	54	117	R	0	231	1065	R	2	48	123	R
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
LR	833	169	0	H	698	125	0	H	136	26	0	H	707	115	1	H	834	167	1	H	699	123	1	H
	180	915	33	M	188	828	23	M	30	106	26	M	211	681	147	M	184	769	175	M	189	704	146	M
	0	89	73	R	1	104	68	R	0	25	137	R	1	44	128	R	0	127	1169	R	1	40	132	R
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
ANN	809	193	0	H	664	159	0	H	123	38	1	H	671	149	3	H	813	187	2	H	685	133	5	H
	177	907	44	M	152	875	12	M	32	93	37	M	234	678	127	M	162	783	183	M	186	701	152	M
	0	92	70	R	1	128	44	R	1	26	135	R	2	55	116	R	0	83	1213	R	2	50	121	R
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
C4.5	741	261	0	H	606	216	1	H	125	35	2	H	723	93	7	H	759	239	4		680	142	1	H
	151	922	55	M	140	879	20	M	44	90	28	M	299	561	179	M	170	819	139		224	710	105	M
	0	82	80	R	1	129	43	R	1	26	132	R	3	52	118	R	2	90	1204		2	83	88	R
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
LMT	765	236	1	H	637	185	1	H	134	28	0	H	727	95	1	H	818	184	0	H	676	145	2	H
	235	808	85	M	284	695	60	M	29	107	26	M	249	634	156	M	180	792	156	M	180	737	122	M
	0	91	71	R	3	117	53	R	0	27	135	R	1	39	133	R	1	75	1220	R	3	59	111	R
	H	M	R		H	M	R		H	M	R		H	M	R		H	M	R		H	M	R	
RF	823	179	0	H	679	144	0	H	130	32	0	H	694	127	2	H	811	191	0	H	681	142	0	H
	175	916	37	M	183	829	27	M	34	90	38	M	204	649	186	M	169	848	111	M	168	785	86	M
	0	90	72	R	1	114	58	R	0	24	138	R	3	32	138	R	0	43	1253	R	2	78	93	R

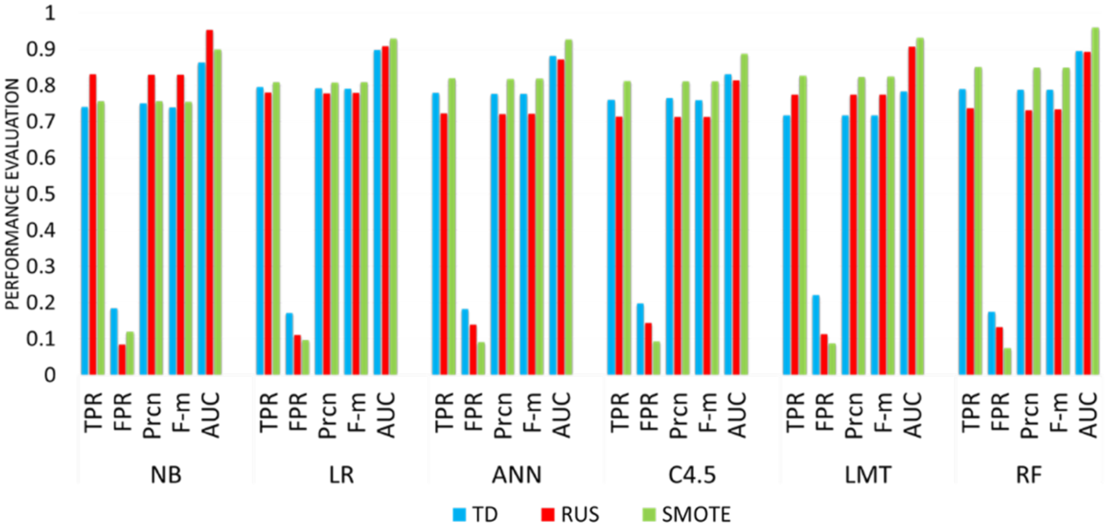


Fig. 5.7 Comparison of training performance metrics across trained models in the female cohort

The AUC for each of the trained models were in the range of 0.783 (for RF on SMOTE) to 0.96 (for C4.5 on TD). These values indicate that the trained models did not sacrifice much precision to achieve a good recall value on the observed data points. The RF model achieved the highest TPR (0.850) when trained on the SMOTE dataset, while the C4.5 model achieved the lowest TPR (0.714) when trained on the RUS dataset.

By observing the confusion matrices for all models after training on all the TD and RUS datasets, it is clear that the number of incorrectly classified instances for class R highly decreased for the models trained on the RUS dataset compared to those trained on the TD. However, when evaluating the minority class accuracy performance in Figure 5.8, it is notable that all trained models benefitted from the sampling methods, exhibiting consistent TPR improvement for class R in each model.

5.2.10 Models test results

The models derived above were tested on a further dataset (female, n = 2035). When the CCI values for all models were compared using the female cohort, the CCI decreased to a maximum degradation of 6.2% when testing the C4.5 model trained on the RUS dataset against the same algorithm trained on the original TD. The LMT model built with SMOTE dataset achieved an overall test accuracy improvement of 6.83% when compared to TD.

In the female cohort (see Figure 5.9), RF models achieved the best TPR of 0.770 when trained on the TD dataset. The LMT model achieved the least TPR of 0.681 when trained on the TD dataset. The ROC area across all tested models ranged between 0.786 (for C4.5 on SMOTE dataset) and 0.889 (for LR on TD).

These values indicate hardly any loss of precision whilst achieving a good recall value on the observed data points. For evaluating risk class, R, the TPR performance (Figure 5.8) classified the risk group with the highest level of 0.798 was achieved by RF on RUS. RF also achieved the greatest TPR improvement in test with a difference of 0.463 between RUS and TD. NB ranked last, with just 0.121 in minority class TPR improvement between NB on SMOTE and TD. These results can be visualised in the confusion matrixes in Table 5.4. The RF model trained on SMOTE correctly classified the highest number of instances (138 of the original 173) in class R. The model which performed the worst in TPR performance for the class R was C4.5 trained on TD, which only correctly classified 43 instances.

The effect of using a variety of machine learning algorithms with different learning schemes is examined. At a model level, Figure 5.9 shows a small difference between the minimum and the maximum TPR test performances per dataset. In the females, tested TD, RUS and SMOTE models showed only differences of 0.1, 0.06 and 0.05, respectively, between the highest and the lowest-performing algorithms.

At a class level, taking the risk group into account for this comparison, Figure 5.8 demonstrates relatively large differences between the minimum and the maximum TPR test performances for the R class in each cohort. In the females, tested TD, RUS and SMOTE models showed high R class accuracy differences of 0.34, 0.13 and 0.25, respectively, between the highest and the lowest-performing algorithms.

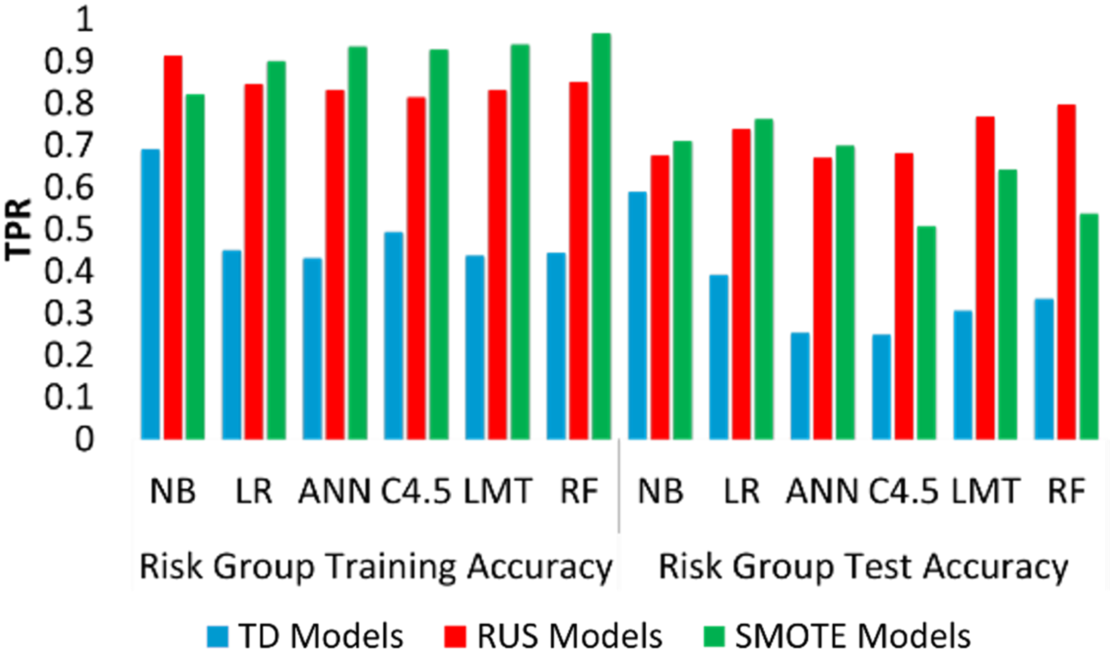


Fig. 5.8 Risk class TPR performance for trained and tested models – Female Cohort

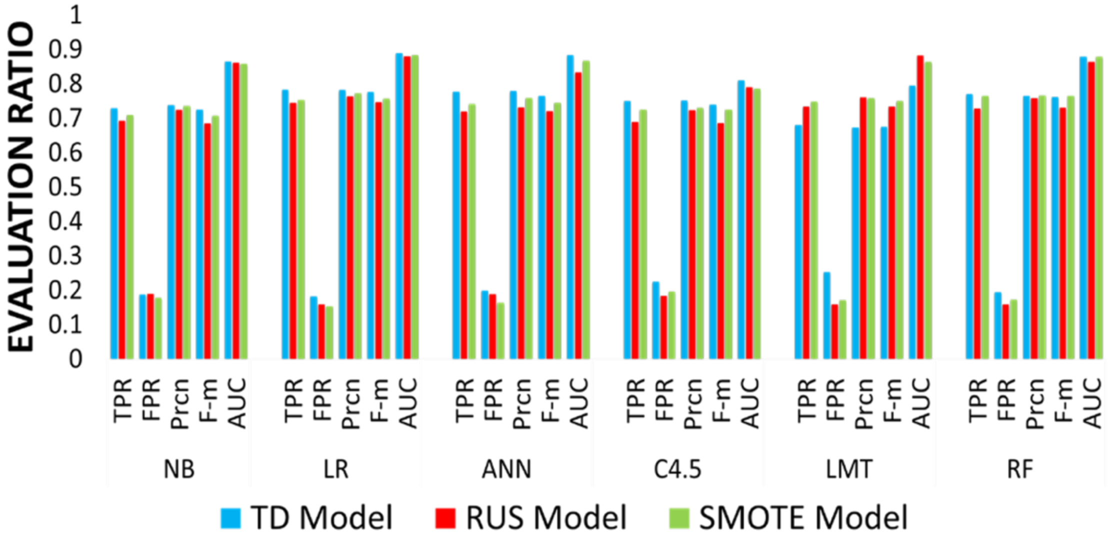


Fig. 5.9 Comparison of performance metrics across all tested models

5.2.11 Attributes information gain analysis

In the female cohort training datasets, when considering the information gain (IG) for each variable across all datasets (Figure 5.10), the IG increased in each attribute for RUS and SMOTE datasets compared to the TD. By comparing the IG ranking of variables in each dataset, it is apparent that WC achieved the highest IG value in all three datasets. The dominance in WC ranking was also accompanied by an increase in its values (from TD to RUS and SMOTE). Such an increase correlates directly with the increase in class R TPR performance in all trained models except for NB where RUS model overtook SMOTE by a small TPR positive margin of 0.092. From Figure 5.10, SMOTE boosted the information within each variable (Table 5.5). This boost, in turn, increased the ability to differentiate class R from other classes in the TD, which in turn increases the class R TPR (see Figure 5.8). The APA multi-dimensional visualisation (Figure 5.11) shows the improved class R discrimination per dataset.

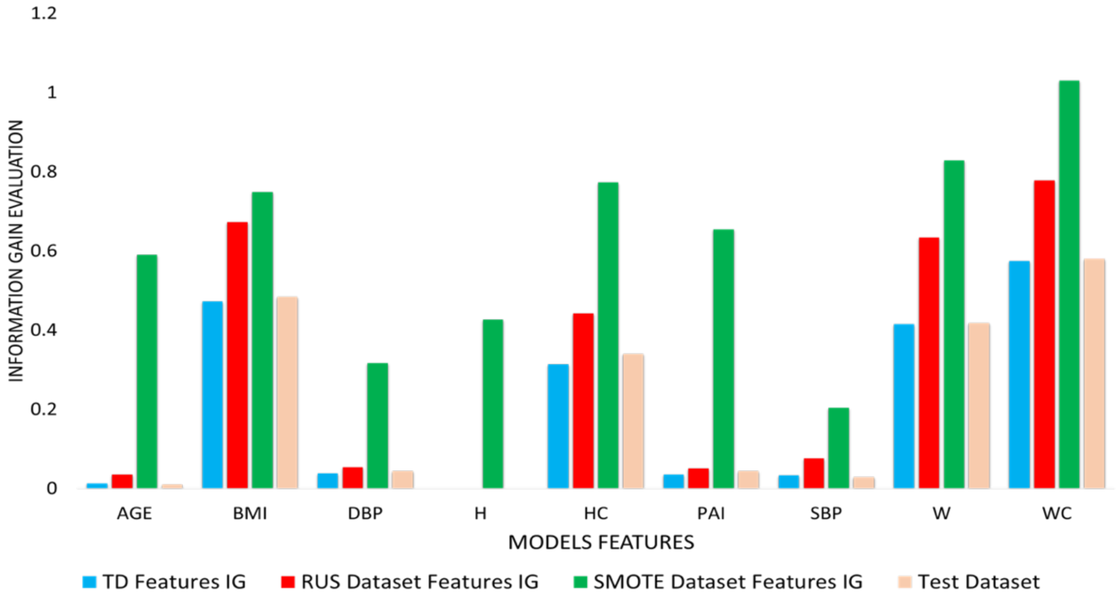


Fig. 5.10 Information Gain evaluation comparison of all variables per female cohort datasets

Table 5.5. The Information Gain evaluation of all features per dataset

Dataset / Features	Female Cohort Datasets			
	TD	RUS	SMOTE	Test Dataset
AGE	0.0125	0.0346	0.5900	0.0092
BMI	0.4717	0.6725	0.7488	0.4839
DBP	0.0375	0.0537	0.3163	0.0435
H	0.0000	0.0000	0.4266	0.0000
HC	0.3130	0.4420	0.7725	0.3399
PAI	0.0346	0.0503	0.6546	0.0435
SBP	0.0331	0.0755	0.2034	0.0278
W	0.4148	0.6341	0.8284	0.4172
WC	0.5743	0.7781	1.0294	0.5806

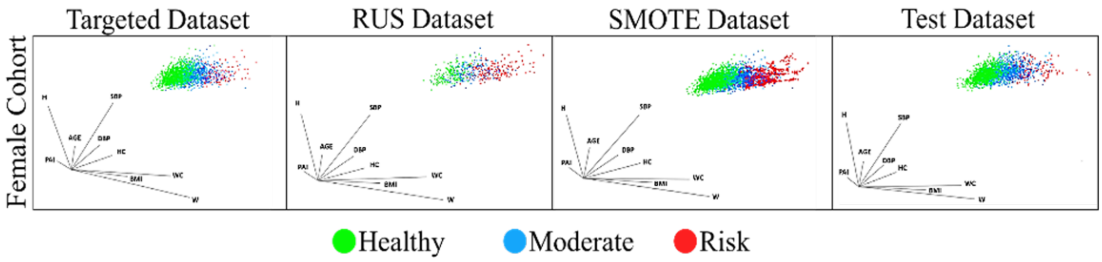


Fig. 5.11 Adaptive projection visualisation of all classes and the effect of sampling methods

5.2.12 Models evaluation and selection

The misclassification of healthy subjects by a predictive model could result in costly and unnecessary follow-up examinations, whilst false-negative misclassifications might result in an individual not receiving an important intervention. In this application, apart from potential cost, there would be few adverse effects associated with misclassified healthy/moderate risk subjects, as such subjects would be encouraged to undertake lifestyle-based interventions to improve their health. Therefore, in this scenario, the best models to adopt would be those which minimise the number of subjects misclassified at 'risk', so they may initiate interventions at an appropriate time.

Confusion matrices are the suitable choice of metric and play an essential role in helping researchers define the best-suited model for use in future trials in this particular case. When analysing the confusion matrices (Table 5.4) from the female cohort, three models were identified as satisfying the domain experts' criteria. These models are reported in Table 5.6. However, they may not necessarily occupy the highest ranks when their other performance metrics are compared to the others.

Table 5.6. Domain-compliant prediction models for females. n (LMT RUS Trained) = 486; n (LR SMOTE Trained) = 3426; n (RF RUS Trained) = 486. n (all Tested) = 2035.

Model	Algorithm	LMT						LR						RF					
	Dataset	RUS						SMOTE						RUS					
Performance	Metrics	CCI	TPR	FPR	Pcrn	F-m	AUC	CCI	TPR	FPR	Pcrn	F-m	AUC	CCI	TPR	FPR	Pcrn	F-m	AUC
	Training	77.37	0.774	0.113	0.774	0.774	0.907	80.91	0.809	0.096	0.807	0.808	0.929	73.66	0.737	0.132	0.731	0.733	0.892
	Test	73.41	0.734	0.159	0.761	0.734	0.883	75.43	0.754	0.154	0.773	0.757	0.884	72.78	0.728	0.159	0.758	0.731	0.864
Confusion Matrixes	Trained	Predicted						Predicted						Predicted					
		H	M	R				H	M	R				H	M	R			
		134	28	0	H	Actual		834	167	1	H	Actual		130	32	0	H	Actual	
		29	107	26	M			184	769	175	M			34	90	38	M		
		0	27	135	R			0	127	1169	R			0	24	138	R		
	Tested	Predicted						Predicted						Predicted					
		H	M	R				H	M	R				H	M	R			
		727	95	1	H	Actual		699	123	1	H	Actual		694	127	2	H	Actual	
		249	634	156	M			189	704	146	M			204	649	186	M		
		1	39	133	R			1	40	132	R			3	32	138	R		

5.2.13 Modelling minority pattern reconstructed data

From the previous analysis, fortunately, we observed the class imbalance learning problem when modelling the female cohort. Therefore, here we examine the effect of using our new cyclic idea of data resampling to reconstruct the imbalanced training dataset (TD) in favour of the minority Hyper-susceptible (Risk group), the class of interest. Also, we examine the effect of modelling the imbalanced MPR reconstructed dataset in combination with RUS and SMOTE.

We reconstruct the training dataset with MPR to favour the minority class of interest. Each cycle has two parts, obtaining a blind (automatic) TPP projection on the sets and then applying *k*-means supervised clustering with (k=3) to cluster the automatic TPP.

The output of k-means will be class A, B and C. In each round, the instances within the clusters (A, B and C) are given their original labelled (R, M and H). Class R represents the minority class of highest interest (Hyper-susceptible subjects), followed by class M (Susceptible subjects), then class H (Resistant subjects). Each formed cluster is examined for the highest count of Class R instances, and once such a cluster is identified, then class R instances are extracted. The extraction subprocess is cluster-label exclusive; this exclusion means that once a cluster is identified as the domain of the label extraction subprocess, the same cluster is excluded from further examination to extract different labels, i.e., the second label of interest class M and so on until a subsample of data is retained.

Unlike the established practice that each feature in a dataset has an Information Gain evaluation of a value greater or equal to zero bits (Kullback–Leibler divergence bits), we take a different thought and assume that an isolated subsample with perfect separable classes can have an information gain profile (information profile) equivalent to the sum of Information Gain for all features n in the isolated subsample S; hence the conditional entropy of all predictors is given by:

$$\boldsymbol{H_n(Y|H) = -\sum_n \sum_x \sum_y p_n(x,y)\log\big(p_n(y|x)\big)}$$

Therefore, the Information Gain Evaluation for a pure subsample *S* with *n* features is given by:

$$\boldsymbol{I_S(Y,X) = \sum_n H_n(Y) - H_n(Y|X)}$$

Each extracted (isolated) and retained subsample for reconstruction has a perfect purity per the subsample information profile. Classifiers are known to be impacted by higher Information Gain. Therefore, these isolated subsamples produce very high variance models (almost perfect classifiers) in training with almost perfect class separation across all classifiers.

Let's look at the isolated subsample with its own context. The high variance classification from training on the isolated subsample may indicate that any considerations for potential interactions among the predictor are ignorable, hence, eliminating the potential of absent confound predictors that might add any further improvement to the classifier training performance on such *S*.

However, the above assumption is only theoretical and lacks mathematical proof. Producing mathematical proof is not in the scope of this thesis. Nevertheless, this is a future research opportunity to explore the validity of this assumption mathematically and on different datasets. Figure 5.12 shows the sequence and the visual impact of applying Minority Pattern Reconstruction on the imbalanced training female cohort TD n(2292) to produce a final resampled set S$_{\text{MPR}}$ n(1190). The resampling clustering cut-off threshold was set in relation to the instant drop of total information gain in the residuals subsample below the sum of information gain for the original imbalanced training set $I_R(Y,X) < I_{TD}(Y,X)$.

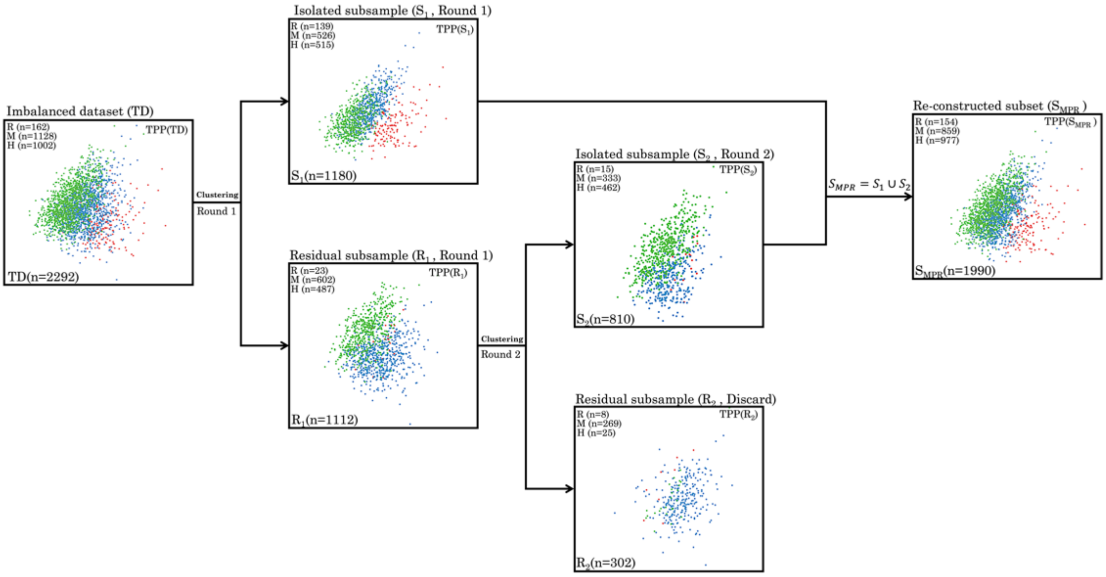


Fig. 5.12 Resampling TD with Minority Pattern Reconstruction (MPR)

The effect of the imbalanced MPR subsample, S$_{\text{MPR}}$ n(1190), on classification modelling can be compared to the imbalanced classifiers' learning on *TD* n(2292) when all produced models are tested on the same test set n(Test) = 2035. Table 5.7 shows the test confusion matrices for modelling on TD and S$_{\text{MPR}}$.

Table 5.7. TD vs MPR models confusion matrices. n(Test) = 2035

	Predicted																		
	NB			LR			MLP			C4.5			LMT			RF			
	R	M	H	R	M	H	R	M	H	R	M	H	R	M	H	R	M	H	Actual
Tested TD Models	102	69	2	68	104	1	44	128	1	43	129	1	53	117	3	58	114	1	R
	94	662	283	23	828	188	12	875	152	20	879	140	60	695	284	27	829	183	M
	0	103	720	0	125	698	0	159	664	1	216	606	1	185	637	0	144	679	H
	R	M	H	R	M	H	R	M	H	R	M	H	R	M	H	R	M	H	
Tested MPR Models	126	46	1	127	45	1	129	42	2	117	54	2	123	49	1	127	44	2	R
	176	673	190	133	724	182	144	644	251	138	741	160	135	721	183	134	727	178	M
	3	143	677	1	131	691	1	96	726	1	210	612	1	127	695	1	145	677	H

From Table 5.7, the ratio of R events in TD is 7.1% and 7.7% in the MPR subset. However, when comparing the tested imbalanced models for each classifier, we observe a large increase in the number of correctly classified instances in the Hyper-susceptible (R) in all the imbalanced MPR models reaching a maximum improvement of 66% in MLP, see Figure 5.13.

These results show that the MPR models outperform the TD models on the R class true positives (TP).

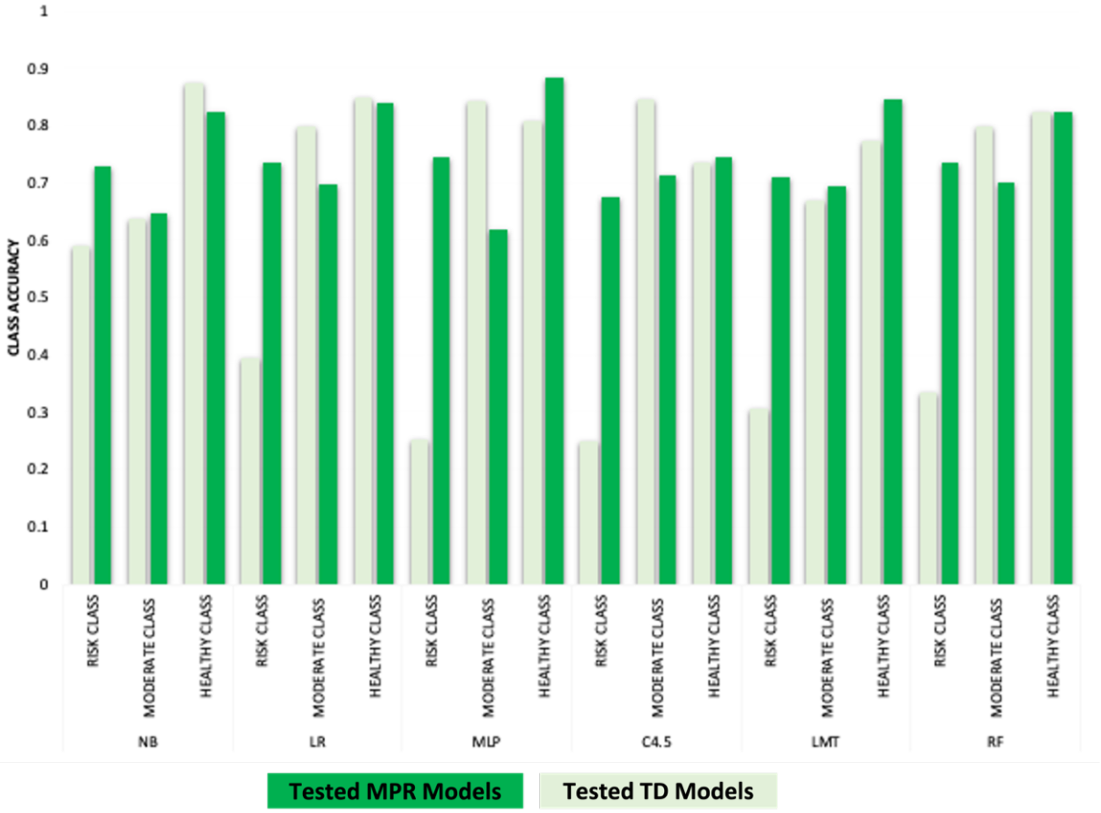


Fig. 5.13 Test class accuracy comparison MPR vs Imbalanced-TD models

From Table 5.7, in MPR models, a slight loss is observed in the correct classifications in the susceptible M classes for all models. The lost portion in class M class TP is heavily classified as false R, except for the NB model, which witnessed a slight improvement for class M.

Similarly, in MPR models, the resistant subjects (H) had a slight drop in the TP cases across NB, LR and RF models but a noticeable improvement in MLP, C4.5 And LMT. The test results on both sets, TD and MPR, show that our resampling tactic to reconstruct the training data with MPR enhances the hyper-susceptible R group predictions and adds value to the Moderate (susceptible) M group when compared to TD models.

We believe that Hyper-susceptible subjects require urgent intervention. Therefore, an increase in TP performance in R is needed. Also, susceptible group M subjects require lifestyle modifications; the filtered reduction in their TP into more False R predictions and fewer H predictions only contributes to identifying more Moderate subjects that require interventions.

In terms of the TP transition in MPR models for the H class, we observe improved TP predictions in the MLP, C4.5 and LMT cases. And a slight deterioration of TP classifications in NB, LR and RF. This reduction in TP only indicates that new MPR models are of similar cost to the imbalanced TD model.

However, when explicitly observing the MPR performance over the given success criteria by the domain experts, The MPR seem to meet the requirements closer than TD. We follow an enhanced identification of R subjects in the MPR models. And we observe a similar model behaviour between MPR and TD in minimising the misclassification of R subjects into the H group and vice versa. These observations lead us to believe that MPR subsample models, despite their imbalance, overtake TD models.

Furthermore, MPR can be combined with Random Under-Sampling (RUS), forming a Hybrid MPR-RUS resampled dataset $S_{MPR-RUS}$ (n=462). See Figure 5.14.

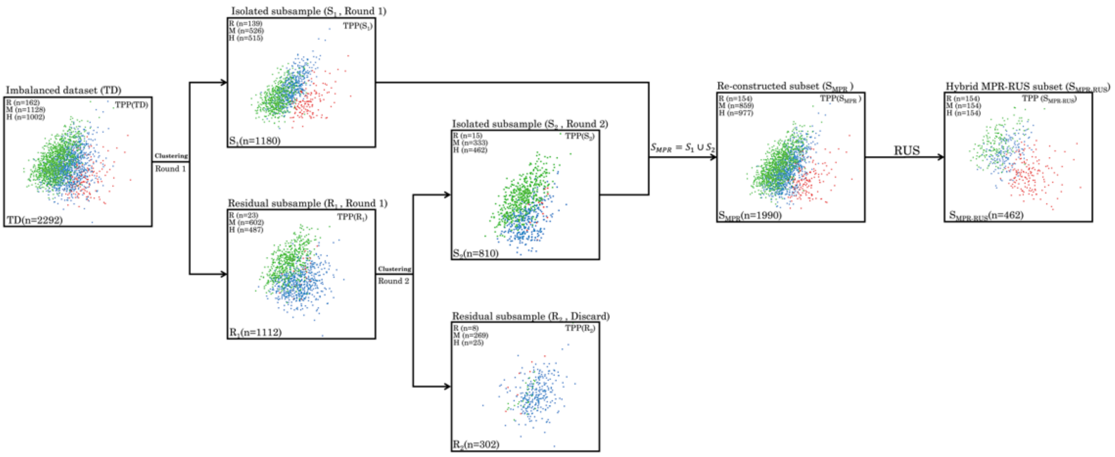
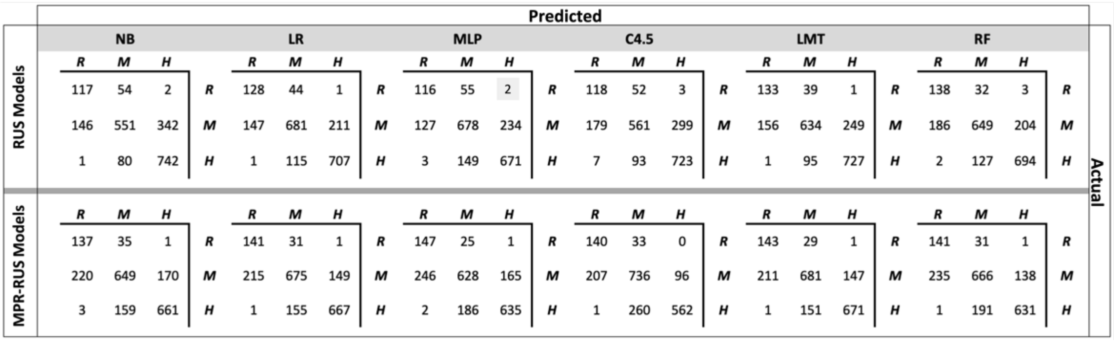


Fig. 5.14 Resampling TD with Hybrid MPR-RUS

Table 5.8 compares the test confusion matrixes for RUS and hybrid MPR-RUS models on the same test set (n=2035). The matrixes show a noticeable improvement in the MPR-RUS classifiers R class TP in all models compared to RUS.

Table 5.8. RUS vs Hybrid MPR-RUS models confusion matrices. n(Test) = 2035



	Predicted																							Actual
	NB				LR				MLP				C4.5				LMT				RF			
RUS Models	*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*	
	117	54	2	*R*	128	44	1	*R*	116	55	2	*R*	118	52	3	*R*	133	39	1	*R*	138	32	3	*R*
	146	551	342	*M*	147	681	211	*M*	127	678	234	*M*	179	561	299	*M*	156	634	249	*M*	186	649	204	*M*
	1	80	742	*H*	1	115	707	*H*	3	149	671	*H*	7	93	723	*H*	1	95	727	*H*	2	127	694	*H*
MPR-RUS Models	*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*		*R*	*M*	*H*	
	137	35	1	*R*	141	31	1	*R*	147	25	1	*R*	140	33	0	*R*	143	29	1	*R*	141	31	1	*R*
	220	649	170	*M*	215	675	149	*M*	246	628	165	*M*	207	736	96	*M*	211	681	147	*M*	235	666	138	*M*
	3	159	661	*H*	1	155	667	*H*	2	186	635	*H*	1	260	562	*H*	1	151	671	*H*	1	191	631	*H*

When comparing the models in Figure 5.15, we observe a considerable uplift in the TPR in the Hyper-susceptible (R) in all MPR-RUS models. Additional improvement is shown in the susceptible (M) classes for all models but LR and MLP.

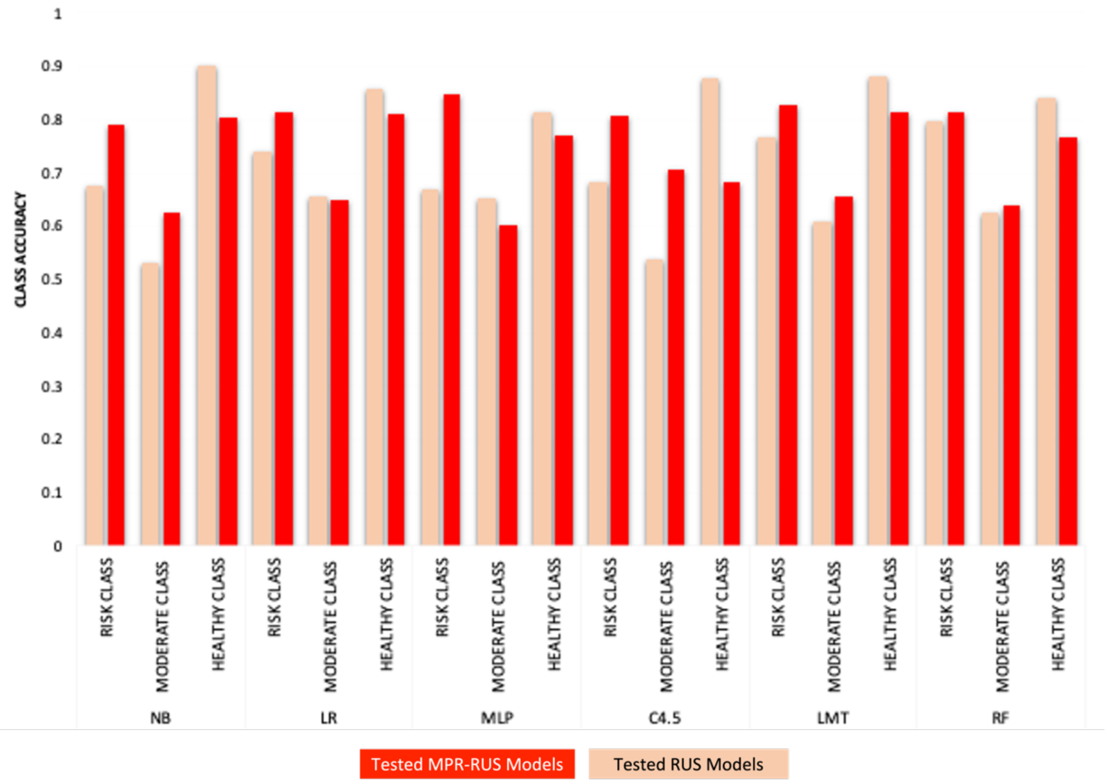


Fig. 5.15 Test class accuracy comparison MPR-RUS vs RUS models

Figure 5.15 shows that our hybrid MPR-RUS has reconstructed the training data in such a way as to prioritise classifier learning the data regions representing those who require lifestyle intervention. However, the resistant subjects (H) had a drop in their TP cases across all models. However, despite the MPR-RUS models better meeting the success criteria with the C4.5 decision tree model with no false misclassifications in class H, it seems that it is a more expensive model. However, in some medical screening cases, healthcare professionals tend to favour models of higher sensitivity.

MPR is not just limited to combining with RUS; it can also be combined with SMOTE. See Figure 5.16. Forming MPR-SMOTE hybrid resampled set n(2760).

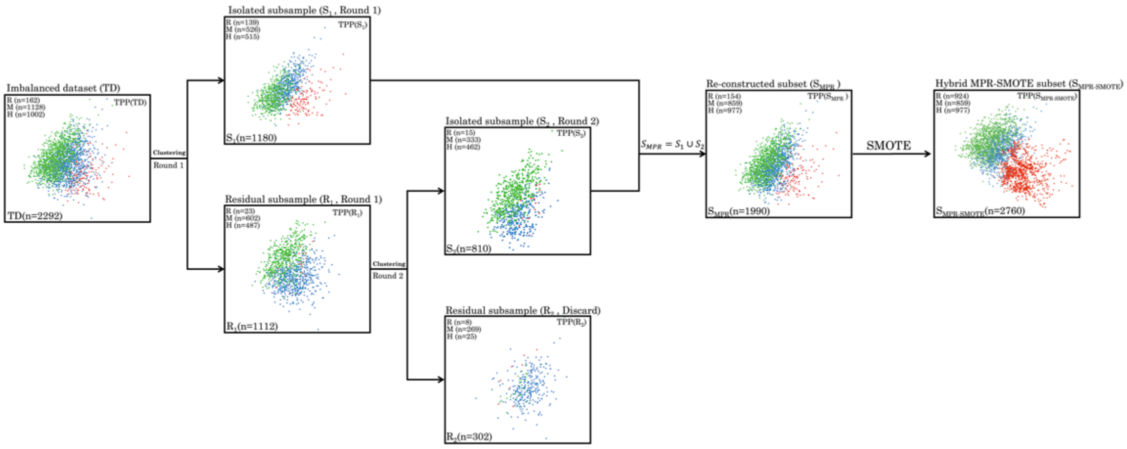
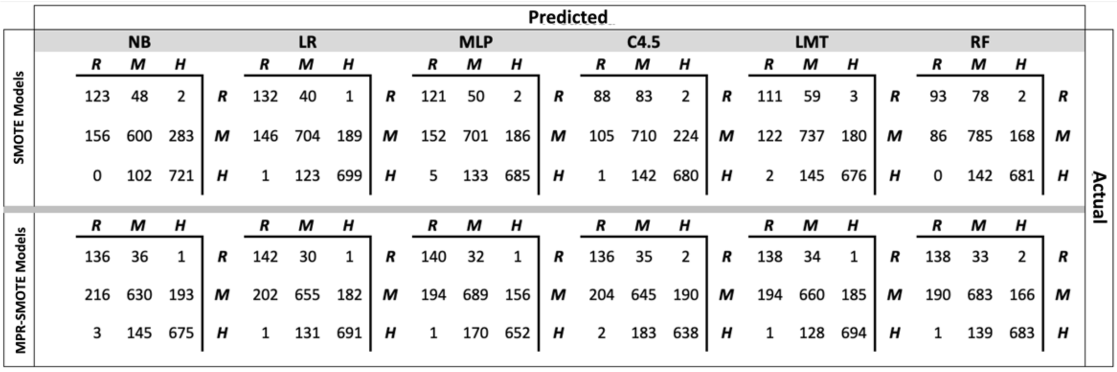


Fig. 5.16 Resampling TD with Hybrid MPR-SMOTE

Table 5.9 shows the test confusion matrixes of the MPR-SMOTE vs SMOTE models. When comparing MPR-SMOTE models to SMOTE only, we still notice improvements in the TP predictions for class R, and a slight loss of TP cases in the M class predicted as the false R. We also see a deterioration of the TP examples in the H group in all models except LMT and RF.

Table 5.9. SMOTE vs Hybrid MPR-SMOTE models confusion matrices. n(Test) = 2035



	Predicted																							
	NB				LR				MLP				C4.5				LMT				RF			Actual
SMOTE Models	**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**	
	123	48	2	**R**	132	40	1	**R**	121	50	2	**R**	88	83	2	**R**	111	59	3	**R**	93	78	2	**R**
	156	600	283	**M**	146	704	189	**M**	152	701	186	**M**	105	710	224	**M**	122	737	180	**M**	86	785	168	**M**
	0	102	721	**H**	1	123	699	**H**	5	133	685	**H**	1	142	680	**H**	2	145	676	**H**	0	142	681	**H**
MPR-SMOTE Models	**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**		**R**	**M**	**H**	
	136	36	1	**R**	142	30	1	**R**	140	32	1	**R**	136	35	2	**R**	138	34	1	**R**	138	33	2	**R**
	216	630	193	**M**	202	655	182	**M**	194	689	156	**M**	204	645	190	**M**	194	660	185	**M**	190	683	166	**M**
	3	145	675	**H**	1	131	691	**H**	1	170	652	**H**	2	183	638	**H**	1	128	694	**H**	1	139	683	**H**

From Figure 5.17, when comparing the models, we observe an uplift in the TRP in the Hyper-susceptible (R) in all MPR-SMOTE models.

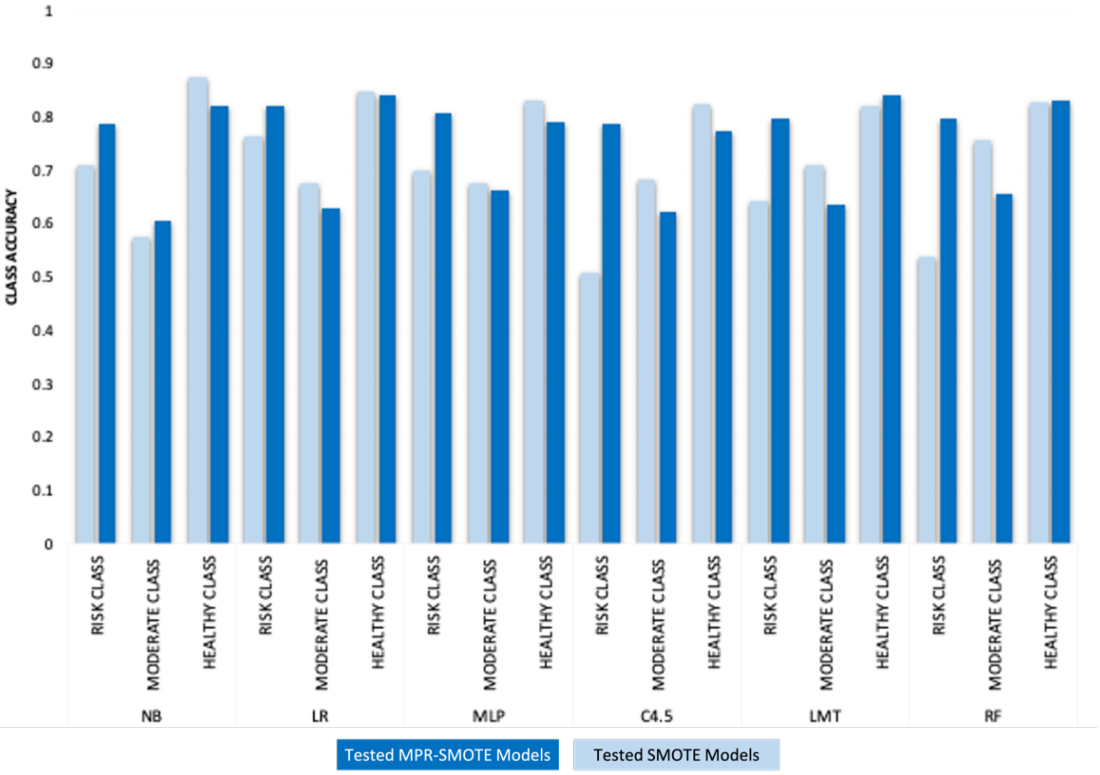


Fig. 5.17 Test class accuracy comparison MPR-SMOTE vs SMOTE models

Also, from Figure 5.17, a slight TPR drop is observed in the susceptible (M) classes for all models but not severe. Similarly, the resistant group (H) had a reduced TPR performance across all models except LMT and RF witnessed a slight TPR increase. These MPR-SMOTE models still outperform the direct SMOTE model in predicting R subjects. However, when the MPR,

MPR-RUS and MPR-SMOTE models compete to meet the success criteria, The MPR-RUS C4.5 model retains its superiority in predicting susceptible patients fully, followed by the MPR-RUS MLP model.

5.2.14 Minority pattern reconstructed data information analysis.

The previous section demonstrated how the minority pattern reconstruction (MPR) improved susceptibility prediction in the imbalanced and balanced (hybrid) sampling settings. Also, we previously mentioned that the total Information Gain governs MPR in extracting the isolated subsamples. Figure 5.18 shows the information gain profile per subsample in the MPR resampling cycles.

We see from Figure 5.18 that the total information gain for TD is 1.89. The total information gain (IG) is expected to increase for the extracted subsample S1 in MPR first clustering cycle, which provides a perfect classification training pattern. S1 IG is 3.46 bits. The residual R1 subsample still shows a greater total IG than the original training set TD; thus, it proceeds to a second clustering cycle where S2 is isolated. S2 IG is still high at 2.83 bits but less than S1. R2 IG is now 0.46 bits less than TD IG, hence discarded, and the MPR clustering cycles are terminated. S1 and S2 subsamples are combined to form the imbalanced MPR subset with an IG of 2.18 bits greater than TD IG by 0.29 bits.

Figure 5.19 gives an idea of the impact of sampling on the features. In the exact figure, the IG per feature is normalised and stacked within its subset profile. Their comparison shows any feature IG proportion leakage or shrinkage within a sampled subset compared to unsampled data.

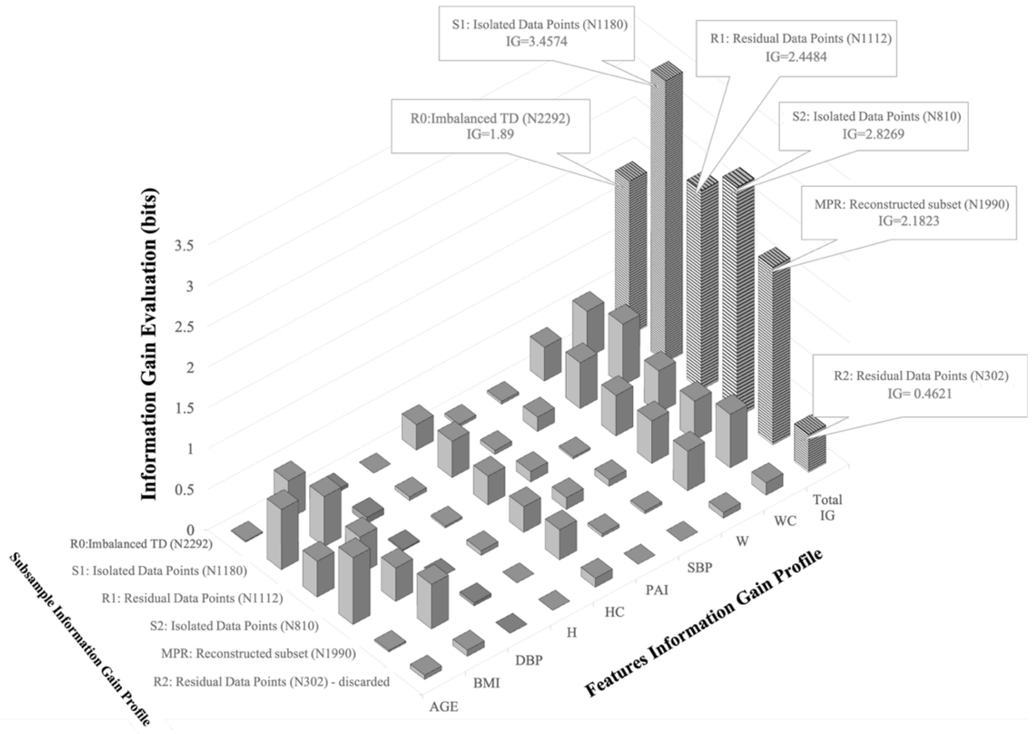


Fig. 5.18 Features and total subsample IG profiles in MPR

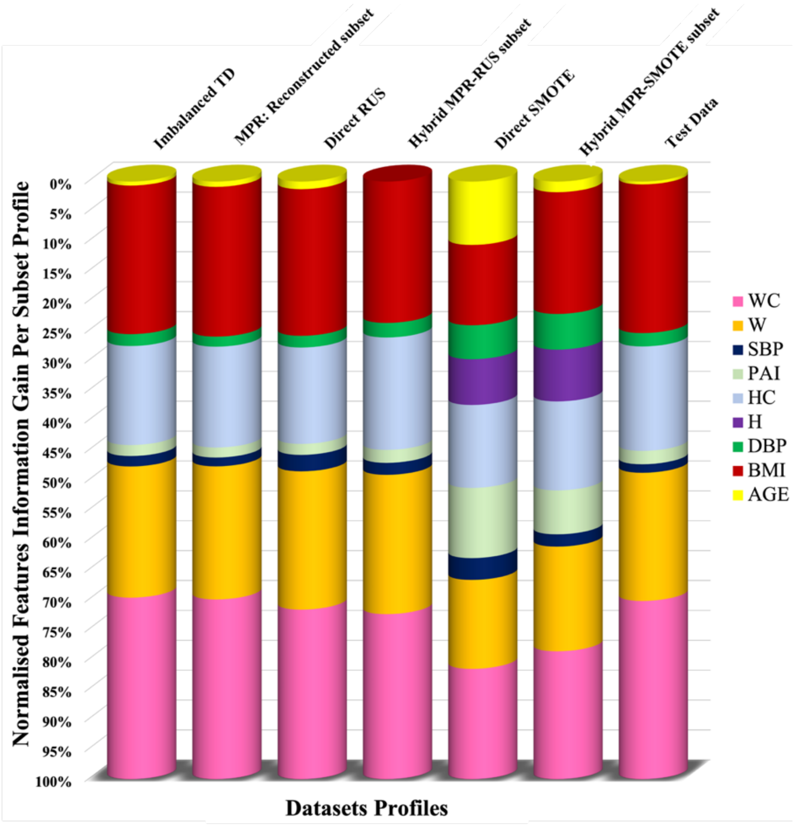


Fig. 5.19 Features' ratio-normalised IG leakage and shrinkage per set

MPR preserved the features information gain proportions close to those in TD. Both TD and the Test sets look homogenous; the comparable sample size ($TD : Test = 57 : 50$), predictors distribution of both sets and their extraction from the same population could be the driver of this homogeneity, resulting in a close proportional purity on both sets. There seems to be minimal leakage or shrinkage of features information gain when comparing ITD, MPR and RUS sets. Unlike the latter three, SMOTE subset shows severe IG leakage and shrinkage in every feature, while MPR-SMOTE shows less leakage and shrinkage behaviour of features IG compared to SMOTE. However, MPR-RUS demonstrated the least leakage and shrinkage severity. The observed behaviour of features IG preservation could be the driver of the better-fitted classifiers with improved performance when applying MPR.

5.3 Predicting males' susceptibility to visceral fat-associated diseases workflow

Three-class multi-nominal machine learning classification models were applied for VAT prediction to predict susceptibility (risk) to disease based on the discretised amount of VAT. A group of 2191 male subjects were used to train eight ML algorithms using 10-fold cross-validation in three different scenarios. The trained models were tested on a new group of external data of 1935 male cases in relation to their cohort. With training to test ratio ~1 (57:50). Figure 5.20 shows the methodology: multiple imbalanced datasets with the same predictor variables were modified with sampling techniques and used for modelling using the eight ML algorithms. Selected performance metrics of the models were compared after training in the evaluation phase. IG was monitored for all predictor variables at every stage.

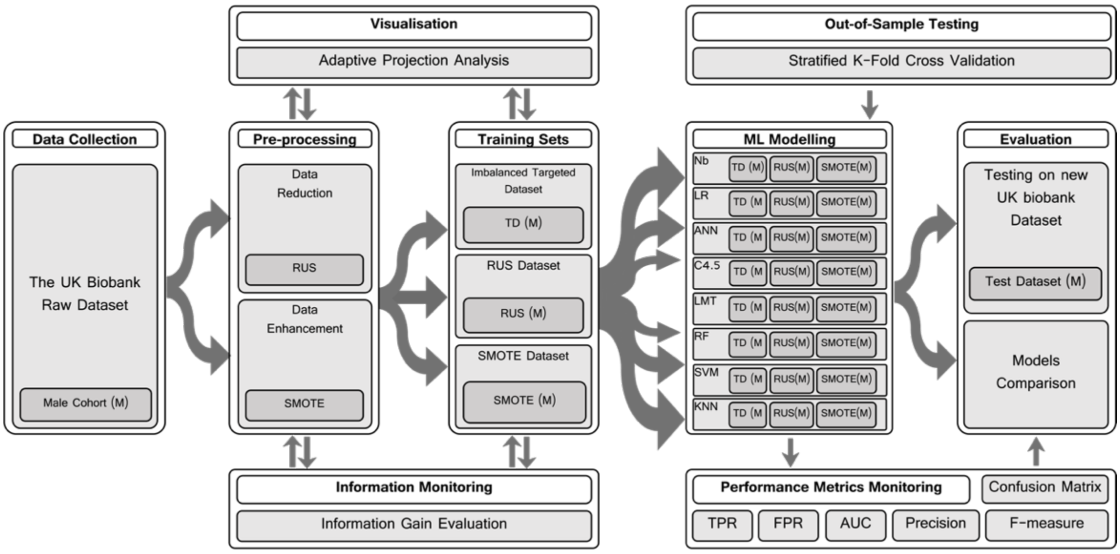


Fig. 5.20 The males' VAT case study methods and workflow. Where TD = Targeted dataset, RUS = Random Under Sampling, SMOTE = Synthetic Minority Oversampling Technique, ML = Machine Learning, NB = Naïve Bayes, LR = Logistic Regression, ANN = Artificial Neural Network, C4.5, LMT = Logistic Model Tree, RF = Random Forest, TPR = true-positive rate, FPR = false-positive rate, AUC = Area under receiver operator characteristic curve

5.3.1 Data collection protocol

This cross-sectional study includes data from 4126 males included in the UKBB multimodal imaging cohort. The UKBB had approval from the North West Multi-Centre Research Ethics Committee (MREC), and written consent was obtained from all participants before their involvement. The data was acquired through the UK Biobank Access Application number 23889. The age range for inclusion was 40-70 years, with exclusion criteria being: metal or electric implants, medical conditions that prohibited MRI scanning or planned surgery within six weeks before the scanning date. The subjects were scanned chronologically between August 2014 and September 2016. The visceral adipose tissue (VAT) volumes were acquired as part of the UKBB dataset.

Anthropometry measurements were collected at UKBB assessment centres; height was measured using the Seca 202 height measure (Seca, Hamburg, Germany). The average of two blood pressure measurements, taken moments apart, was obtained using an automated device (Omron, UK). Images were acquired at the UK biobank imaging Centre at Cheadle (UK) using a Siemens 1.5T Magnetom Aera. The participants' height and weight were recorded before imaging screening which later was utilised to calculate the Body Mass Index (BMI). For physical activity assessment data, a touch screen questionnaire was used to collect information on sociodemographic characteristics and lifestyle exposures (http://www.ukbiobank.ac.uk/ resources/). Specific questions on the frequency and duration of walking (UK biobank field ID: 864, 874), moderate physical activity (884, 894) and vigorous physical activity (904, 914) events allowed the calculations of metabolic equivalent-minutes per week (MET-min/week) for each individual. Participants were excluded from the calculations and analysis if they selected 'prefer not to answer' or 'do not know' to any of the possible six questions on physical activity used to calculate the MET score.

5.3.2 *Experimental design*

VAT-related disease susceptibility was based on the following MRI response labels: Healthy (Resistant), Moderate (Susceptible) and Risk (Hyper-Susceptible) defined according to VAT volume. In males, VAT volume of ≤3 litres was deemed 'Healthy' (H); VAT volume >3 litres but ≤6 litres was classed as 'Moderate' (M); VAT volume >6 litres was classified as 'Risk' (R) [344]. The training datasets contained ten data variables reported in Table 5.10, with the VAT in litres being the class determination response variable. All nine predictor variables in Table 5.10 were selected as input features by domain experts based on their low cost to obtain and associations with VAT prediction in previous studies.

The choice of the limited number of anthropometry and physical activity variables is also driven by the desire to produce a tool using the least expensive variable to collect by the subjects themselves. Thus, the new tool can be integrated into a new web application for professional and domestic use. Therefore, it is unlikely that additional raw variables are required to be collected within the domain of this study.

Table 5.10. Descriptive statistics of variables in the males' Targeted Dataset (TD)

Male Cohort (n=2191)			
Response variable	Median	Mean	(Min, Max)
Visceral adipose tissue volume (VAT in litres)	5.6	4.7	(0.35, 9.63)
Predictor variables			
Waist Circumference (WC in cm)	102.0	92.5	(66.0, 138.0)
Pre-imaging Weight (W in Kg)	104.0	82.5	(53.0, 155.0)
BMI (in kg/m^2)	33.0	26.6	(18.0, 48.0)
Hip circumference (HC in cm)	116.5	101.1	(83.0, 150.0)
Standing height (H in cm)	176.0	176.1	(152.0, 200.0)
Systolic blood pressure (SBP in mmHG)	159.0	141.9	(99.0, 219.0)
Diastolic blood pressure (DBP in mmHG)	84.5	80.8	(51.0, 118.0)
Physical Activity Index (PAI)	3.0	0.5	(-12.0, 18.0)
Age at recruitment (AGE in years)	55.0	56.4	(40.0, 70.0)

5.3.3 Physical activity index (PIA) feature construction

The UK Biobank Physical Activity Index (UKBB PAI or PAI) was created by domain experts [345] using data collected during physical activity assessment; comprising a total of 27 outcomes, 23 outcomes reflecting activity and four reflecting inactivity (see Table 5.11). An individual's response to questions was scored with values between -1 and +1 and combined cumulatively to give a final score with an increasingly negative score implying a progressively unhealthier phenotype. For binary variables, 0 indicated the absence of the parameter, 1 the presence.

Table 5.11. UK Biobank outcomes used in creating the physical activity index

UK BB ID	Outcome	Units
816	Job involves heavy lifting	Categorical
864	Days/week walked 10+ minutes	Days/Week
874	Duration of Walks	Minute/Day
884	Days/week moderate physical activity 10+ minutes	Days/Week
894	Duration of moderate activity min	Min/Day
904	Days/week vigorous physical activity 10+ minutes	Days/Week
914	Duration of vigorous activity	Minute/Day
924	Usual walking pace	Categorical
943	Frequency of stair climbing in last 4 weeks	Categorical
971	Frequency of walking for pleasure in last 4 weeks	Categorical
981	Duration of walking for pleasure	Categorical
991	Frequency of strenuous sports in last 4 weeks	Categorical
1001	Duration of strenuous sports	Categorical
1011	Frequency of light DIY in last 4 weeks	Categorical
1021	Duration of light DIY	Categorical
2624	Frequency of heavy DIY in last 4 weeks	Categorical
2634	Duration of heavy DIY	Categorical
3637	Frequency of other exercises in last 4 weeks	Categorical
3647	Duration of other exercises	Categorical
6164	Types of physical activity in past 4 weeks	Categorical
104900	Time spent doing vigorous physical activity	Categorical
104910	Time spent doing moderate physical activity	Categorical
104920	Time spent doing light physical activity	Hours
806	Job involves mainly standing or walking	Categorical
1070	Time spent watching television	Hour/Day
1080	Time spent using computer	Hour/Day
1090	Time spent driving	Hour/Day

5.3.4 Inspecting irregular training examples

Inter Quartile Range (IQR) test was used per attribute to look for defined values within an Extreme Value Factor (EVF) of 6 times the IQR and an Outlier Factor (OF) of 3 times the IQR. A total of 15 abnormal observations were found, all of which were outliers and had no extreme values. These data values were verified and retained. These abnormal observations might be part of a phenomenon. Figure 5.21 shows the outliers and extreme values detection per variable in the imbalanced targeted dataset TD.

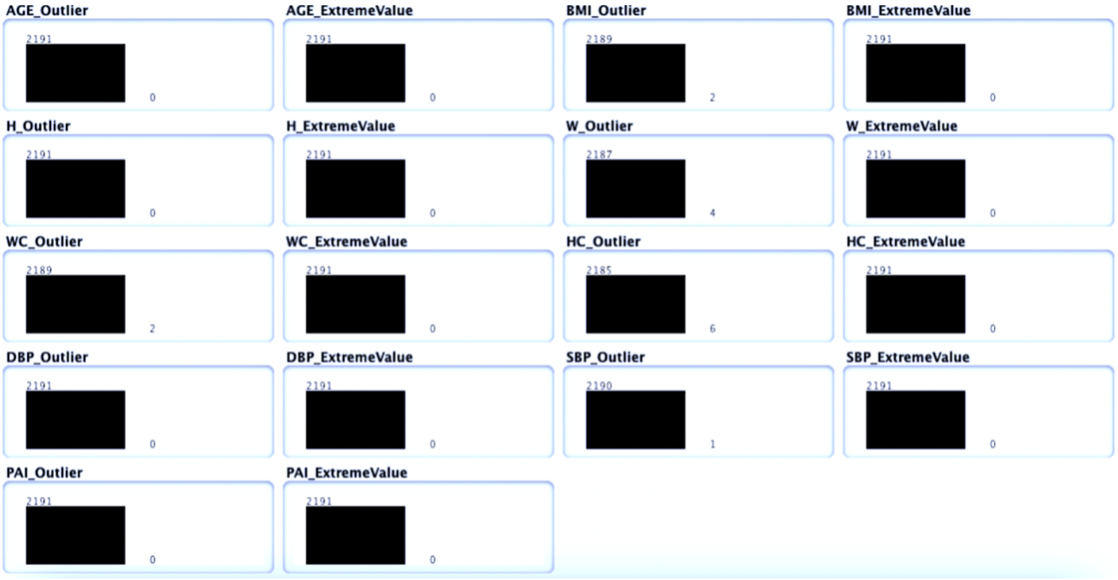


Fig. 5.21 Outliers and extreme values detection per attribute in the training dataset TD

5.3.5 Training instances resampling

Targeted dataset (TD). The TD was the first dataset modelled. The TD contained 2191 male records from the UKBB cohort. Table 5.10 shows the summary statistics of all TD's variables. Unlike the previous case study on females, the TD was less severely imbalanced in the male cohort in relation to records numbers per class: In the males' TD, class H had 489 subjects, class M had 1125 subjects, and class R contained 577 subjects. The class imbalance of TD can be observed via APA visualisation in Figure 5.22.

Random under-sampled (RUS) dataset: this dataset was a reduced subset of TD. A subset of each majority class was randomly removed to balance the data. As a result of applying RUS to the males' TD, each of the H, M and R classes ended up with 489 subjects. The effect of RUS can be observed in APA visualisation in Figure 5.22.

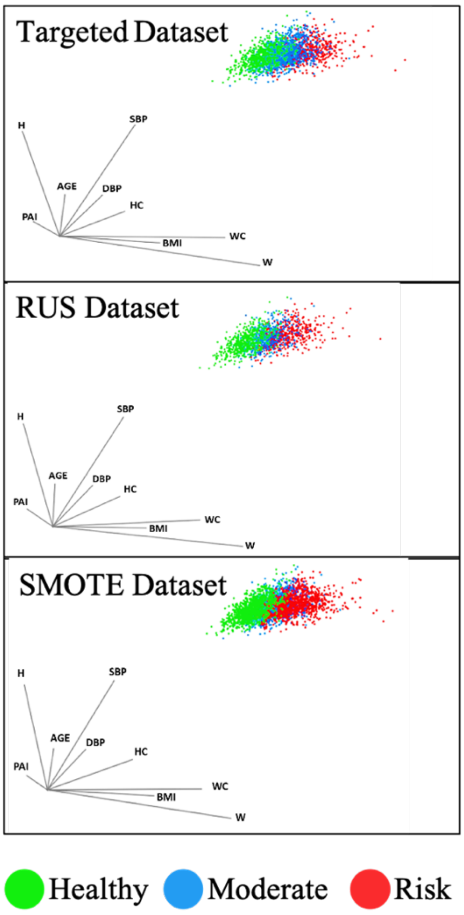


Fig. 5.22 Adaptive projection visualisation of Targeted Dataset, Random Under Sampled dataset and SMOTE dataset variables – Male Cohort

Synthetic Minority Over-Sampled (SMOTE) dataset: This dataset was obtained as a result of applying SMOTE to the numeric data variables of TD. By doing so, the three VAT classes became more closely balanced. In the male cohort, class H had 1125 subjects, class M had 1125 subjects, and class R contained 1125 subjects. The effect of SMOTE can be observed via APA visualisation in Figure 5.22.

Similar to the previous case study for females, IG Evaluation Algorithm was used to measure the information levels for independent variables in relation to the class variable. The measurement and ranking of IG in each independent variable in TD, RUS and SMOTE training sets are presented in the discussion.

5.3.6 *Test Data Characteristics*

The ML models were tested on a new group of 1935 males from the UKBB male cohort. Table 5.12 shows the test set variables summary statistics. Like the TD, the class imbalance in the males' test dataset was less severe: class H had 468 subjects, class M had 906, and class R contained 561 subjects.

Table 5.12. Descriptive statistics of variables in the males' test set

Male Cohort (n=1935)			
Numeric test dataset variables	Median	Mean	(Min, Max)
Response variable			
Visceral adipose tissue volume (VAT in litres)	7.2	4.9	(0.3, 14.1)
Predictors variables			
Waist Circumference (WC in cm)	101.0	93.3	(63.0, 139.0)
Pre-imaging Weight (W in Kg)	100.0	83.4	(50.0, 150.0)
BMI (in kg/m^2)	32.5	26.9	(17.0, 48.0)
Hip circumference (HC in cm)	109.5	101.4	(78.0, 141.0)
Standing height (H in cm)	178.5	175.8	(156.0, 201.0)
Systolic blood pressure (SBP in mmHG)	142.0	137.1	(75.0, 209.0)
Diastolic blood pressure (DBP in mmHG)	83.5	79.9	(47.0, 120.0)
Physical Activity Index (PAI)	2.8	0.5	(-12.0, 17.5)
Age at recruitment (AGE in years)	55.0	56.0	(40.0, 70.0)

5.3.7 *Inspecting irregular test records*

In a similar approach to outliers' detection in TD, the Inter Quartile Range (IQR) test was used per test data attribute to look for defined values with an Extreme Value Factor (EVF) of 6 times the IQR and Outlier Factor (OF) of 3 times the IQR. A total of 13 abnormal observations were identified, all of which were outliers and had no extreme values. All points were verified and retained. The abnormal observations may be part of a phenomenon. Figure 5.23 shows the test dataset's outliers and extreme values per variable. All thirteen outlier data points are found in the risk groups.

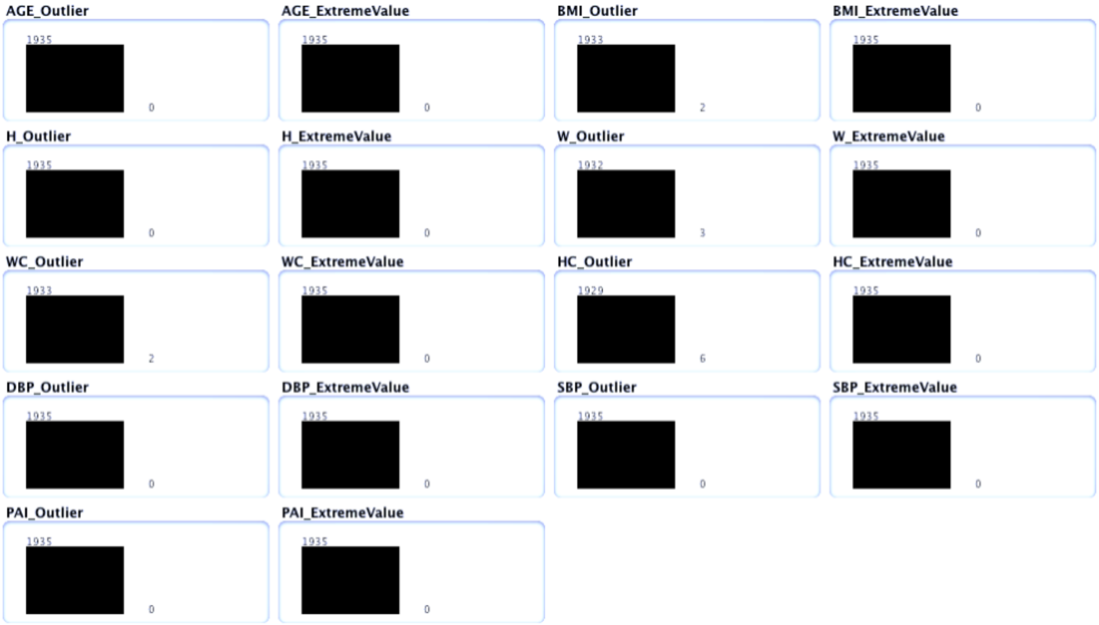


Fig. 5.23 Outliers and extreme values detection per attribute in the males' test dataset

5.3.8 Modelling males' susceptibility to VAT diseases

The classification modelling males' susceptibility to adverse health conditions associated with visceral fat uses six algorithms. The modelling algorithms are described in Chapter 3. And these are Naïve Bays (NB), Logistic Regression (LR) with Ridge Regularization, Artificial Neural Networks (ANN) with Multi-Layer Perceptron (MLP) architecture and built-in feature scaling (Range Normalisation), C4.5 Decision Tree algorithm, and ensemble learners Logistic Model Tree (LMT) and Random Forest (RF).

5.3.9 Models training results

The accuracies (CCI) of the models for the male cohort were calculated from Table 5.13. SMOTE resampling resulted in a consistent improvement in CCI compared to the original TD. SMOTE resampling resulted in a consistent improvement in CCI compared to the original TD. The training performance of all models for the male cohort using the RUS dataset was reduced compared to the same algorithms trained on the TD.

Table 5.13. Male cohort VAT Prediction Models Confusion Matrices

		TD						RUS Dataset						SMOTE Dataset					
		Training			Test			Training			Test			Training			Test		
		R	*M*	*H*	*R*	*M*	*H*	*R*	*M*	*H*	*R*	*M*	*H*	*R*	*M*	*H*	*R*	*M*	*H*
NB	*R*	331	242	4	361	198	2	308	171	10	403	154	4	738	371	16	398	156	7
	M	143	762	220	117	637	152	86	278	125	162	528	216	200	620	305	160	526	220
	H	2	127	360	1	127	340	5	98	386	3	89	376	6	203	916	2	85	381
LR	*R*	350	226	1	383	177	1	372	114	3	451	108	2	862	258	5	451	108	2
	M	111	901	113	108	739	59	113	272	104	193	539	174	237	650	238	193	544	169
	H	0	180	309	1	170	297	0	106	383	3	97	368	1	207	917	3	101	364
MLP	*R*	371	206	0	405	155	1	367	119	3	433	126	2	904	216	5	435	125	1
	M	158	880	87	131	701	74	117	286	86	180	588	138	290	629	206	198	624	84
	H	0	211	278	2	191	275	5	117	367	3	125	340	6	240	879	4	150	314
C4.5	*R*	346	228	3	379	181	1	371	106	12	438	110	13	445	108	8	445	108	8
	M	221	775	129	157	687	62	156	201	132	251	430	225	268	517	121	268	517	121
	H	6	205	278	6	215	247	18	113	358	11	104	353	9	172	287	9	172	287
LMT	*R*	329	247	1	377	183	1	387	99	3	469	90	2	918	191	16	468	91	2
	M	110	905	110	104	746	56	126	259	104	237	487	182	270	648	207	228	568	110
	H	0	179	310	1	178	289	5	96	388	5	93	370	14	193	918	7	131	330
RF	*R*	345	232	0	383	177	1	380	107	2	458	97	6	951	164	10	421	139	1
	M	142	868	115	131	717	58	129	252	108	250	478	178	201	764	160	180	634	92
	H	3	200	286	5	179	284	7	115	367	7	95	366	14	149	962	7	151	310

The AUC for each of the trained models were in the range of 0.729 (for C4.5 on TD) to 0.923 (for RF on SMOTE). These values indicate that the trained models did not sacrifice a lot of precision to obtain a good recall value on the observed data points. The RF model trained on the SMOTE dataset achieved the highest TPR (0.793), while the C4.5 model trained on the RUS dataset achieved the lowest TPR (0.631).

Examination of the confusion matrices for all models (Figure 5.24) trained on the TD vs the RUS datasets demonstrated that the number of subjects incorrectly classified as class H instead of class R increased for models trained on the RUS dataset compared with those trained on the original TD despite the removal of 88 subjects from the original R group as a result of RUS.

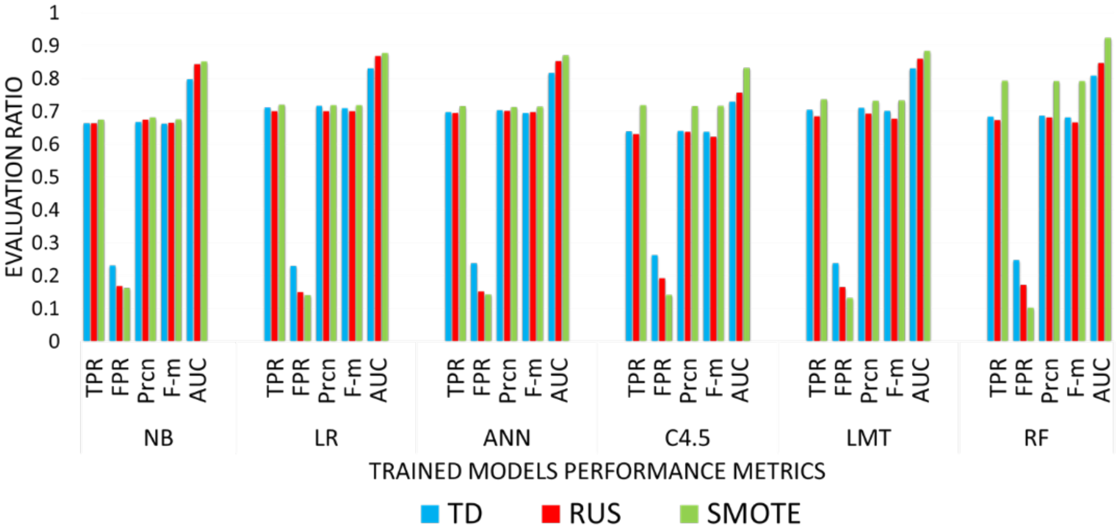


Fig. 5.24 Comparison of performance metrics across trained models in the male cohort

The number of correctly classified instances for class H increased. However, when evaluating class R accuracy performance (see Figure 5.25), it is notable that all trained models benefitted from the sampling methods, exhibiting consistent TPR improvement for class R in each model.

5.3.10 Models test results

The models derived above were tested on a further dataset (male n=1935). In male cohort subjects, when comparing the CCI for all models, CCI decreased with a maximum degradation of 11.9% when testing the RF model trained on the SMOTE dataset compared to the same model built on the TD. All models built on the TD showed an overall model accuracy improvement on test datasets, the highest model accuracy improvement of 4.0% was achieved with the C4.5 model trained on the TD dataset compared to all other models. The models' overall accuracy improvements in test were also observed for NB, LR and MLP models trained on the RUS dataset, with the greatest improvement of 1.2% on NB compared to all models built with the

RUS dataset. All models built with SMOTE dataset suffered an overall model accuracy degradation in test except for NB overall accuracy, which remained unchanged.

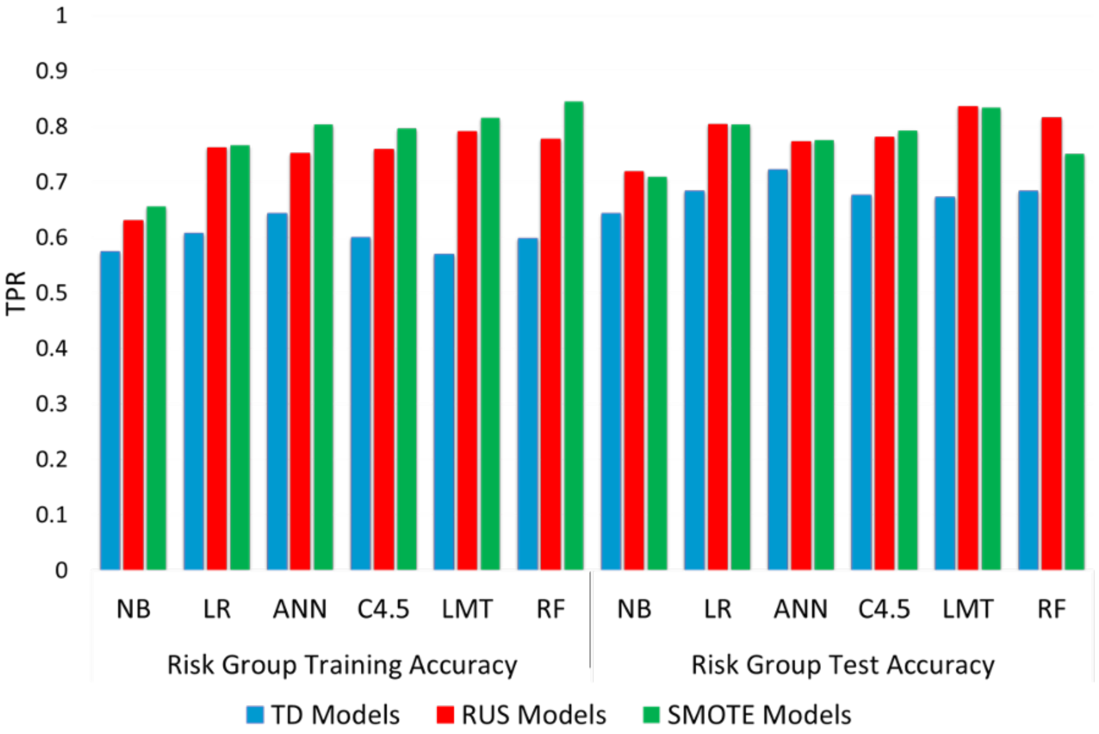


Fig.5.25 Risk class TPR performance for trained and tested models – male cohort

In the male cohort, it was observed that in test, LR models achieved the best TPR of 0.733 when trained on the TD dataset (see Figure 5.26). The LMT model achieved the least TPR of 0.730 when trained on the TD dataset. The ROC area across all tested models ranged between 0.753 (for C4.5 on SMOTE) and 0.864 (for LR on both TD and SMOTE, and LMT on TD). These values indicate that also, the tested models do not sacrifice much precision to obtain a good recall value on the observed data points.

When observing class R, the TPR performance results in Figure 5.25 show that consistent improvements were made in classifying the risk group, with the highest level of 0.836 achieved by LMT on RUS.

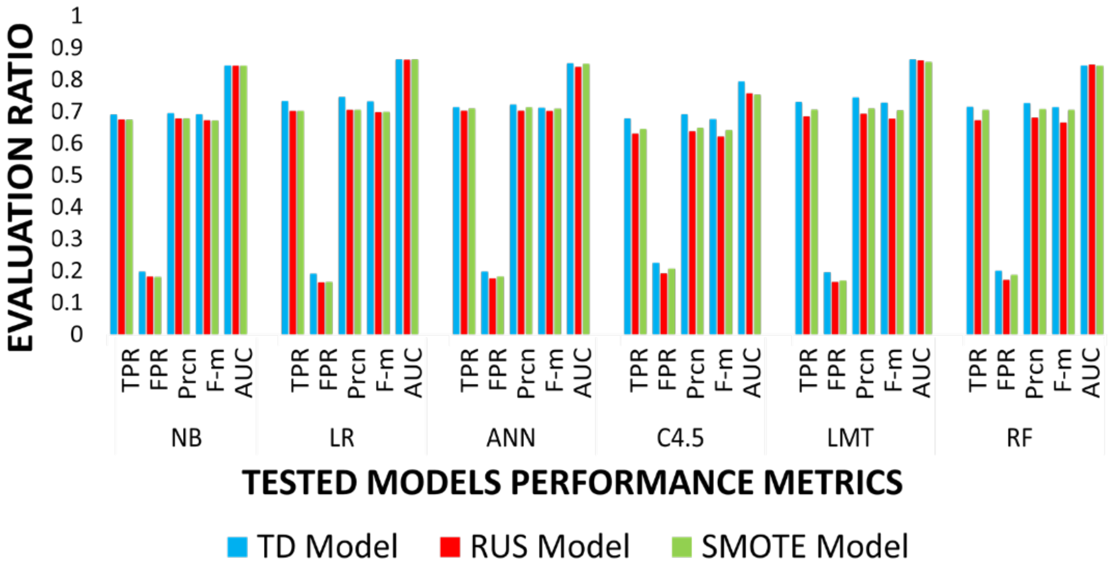


Fig. 5.26 Comparison of performance metrics across all tested models

LMT also achieved the greatest TPR improvement in test with a difference of 0.164 between LMT on RUS and LMT on TD, while MLP ranked last, with just 0.05 in class R TPR improvement between MLP on RUS and TD. This comparison is demonstrated in the confusion matrixes in Table 5.13. The LMT model trained on RUS correctly classified the highest number of instances (469 of the original 561) in class R. The model which performed the worst in TPR performance for class R was NB trained on TD, which only correctly classified 361 instances.

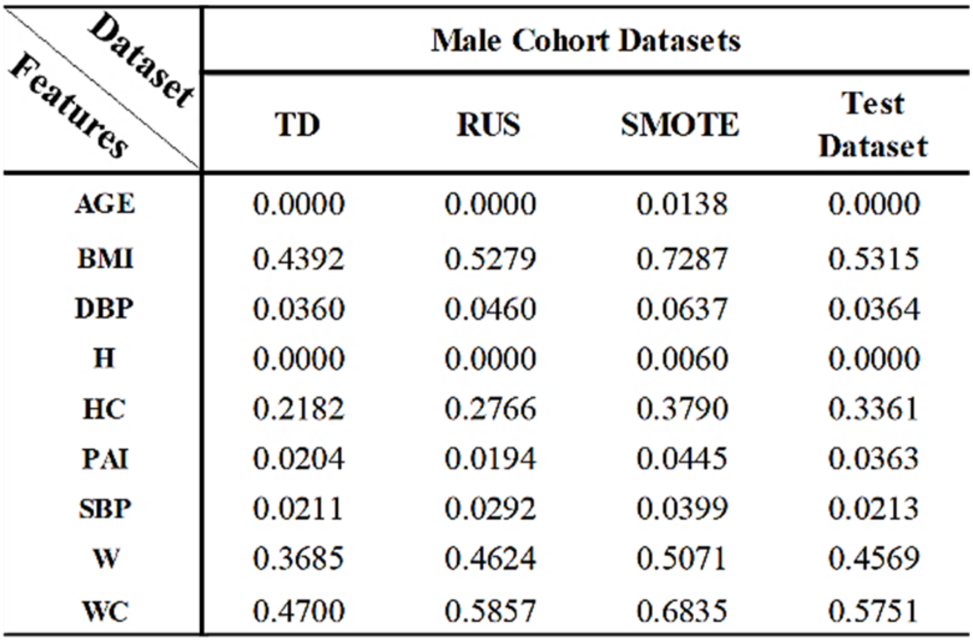
The effect of using a variety of ML algorithms with different learning schemes is examined. At a model level, figure 5.26 shows a small difference between the minimum and the maximum TPR test performances per dataset in each cohort. In the males, tested TD, RUS and SMOTE models showed differences of 0.06, 0.07 and 0.07, respectively, between the highest and the lowest-performing algorithms. C4.5 showed consistency in achieving the least TPR among all tested models.

At a class level, taking the risk group into account for this comparison, Figure 5.25 demonstrates some differences between the minimum and the maximum TPR test performances for the R class in the male cohort. TPR differences were found in the males' TD, RUS and SMOTE models of 0.08, 0.13 and 0.12, respectively, between the highest and the lowest-performing algorithms. NB showed consistency in scoring the lowest TPR among all tested models.

5.3.11 *Attributes' information gain analysis*

The Information Gain evaluations are found in Table 5.14. In the male cohort training datasets, when considering the measured IG for each variable across all datasets (Figure 5.27), it is observed that the IG increased in each attribute for SMOTE dataset and some of the attributes for the RUS dataset compared to the TD. By comparing the IG ranking of variables in each dataset, it was apparent that waist circumference (WC) achieved the highest IG value in all the TD and RUS datasets, while BMI achieved the highest IG value in the SMOTE dataset.

Table 5.14. The Information Gain evaluation of all features per dataset



Dataset / Features	Male Cohort Datasets			
	TD	RUS	SMOTE	Test Dataset
AGE	0.0000	0.0000	0.0138	0.0000
BMI	0.4392	0.5279	0.7287	0.5315
DBP	0.0360	0.0460	0.0637	0.0364
H	0.0000	0.0000	0.0060	0.0000
HC	0.2182	0.2766	0.3790	0.3361
PAI	0.0204	0.0194	0.0445	0.0363
SBP	0.0211	0.0292	0.0399	0.0213
W	0.3685	0.4624	0.5071	0.4569
WC	0.4700	0.5857	0.6835	0.5751

The advancement in BMI ranking in SMOTE dataset correlates directly with the increase in class R TPR performance in all trained models. SMOTE resampling technique amplified the information within each variable (Figure 5.27). This amplification, in turn, increased the class R border density with other classes in the training dataset, which in turn increased class R TPR in training (see Figure 5.25). The APA visualisation showing the enhancement in class R borders density per dataset is shown in Figure 5.28.

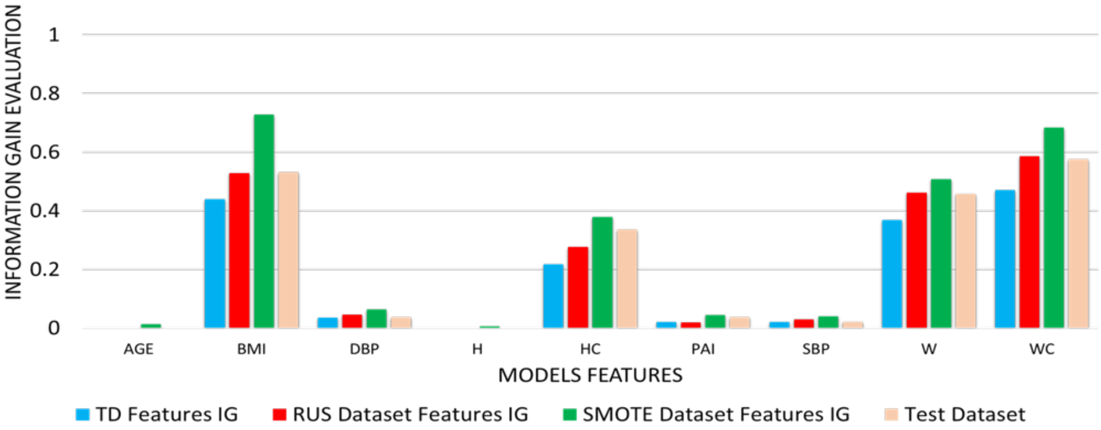


Fig. 5.27 IG evaluation comparison of all variables per dataset

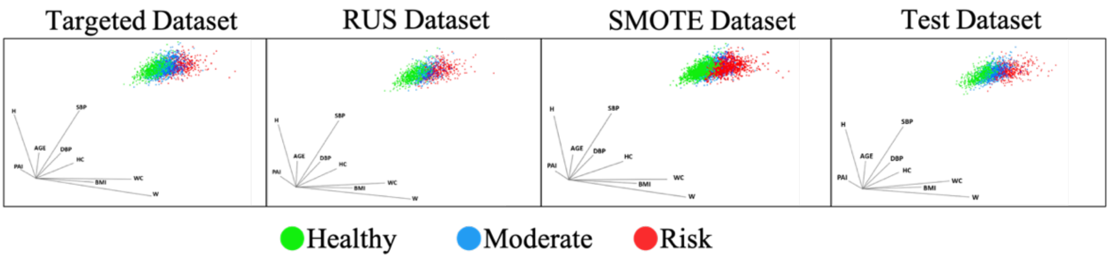
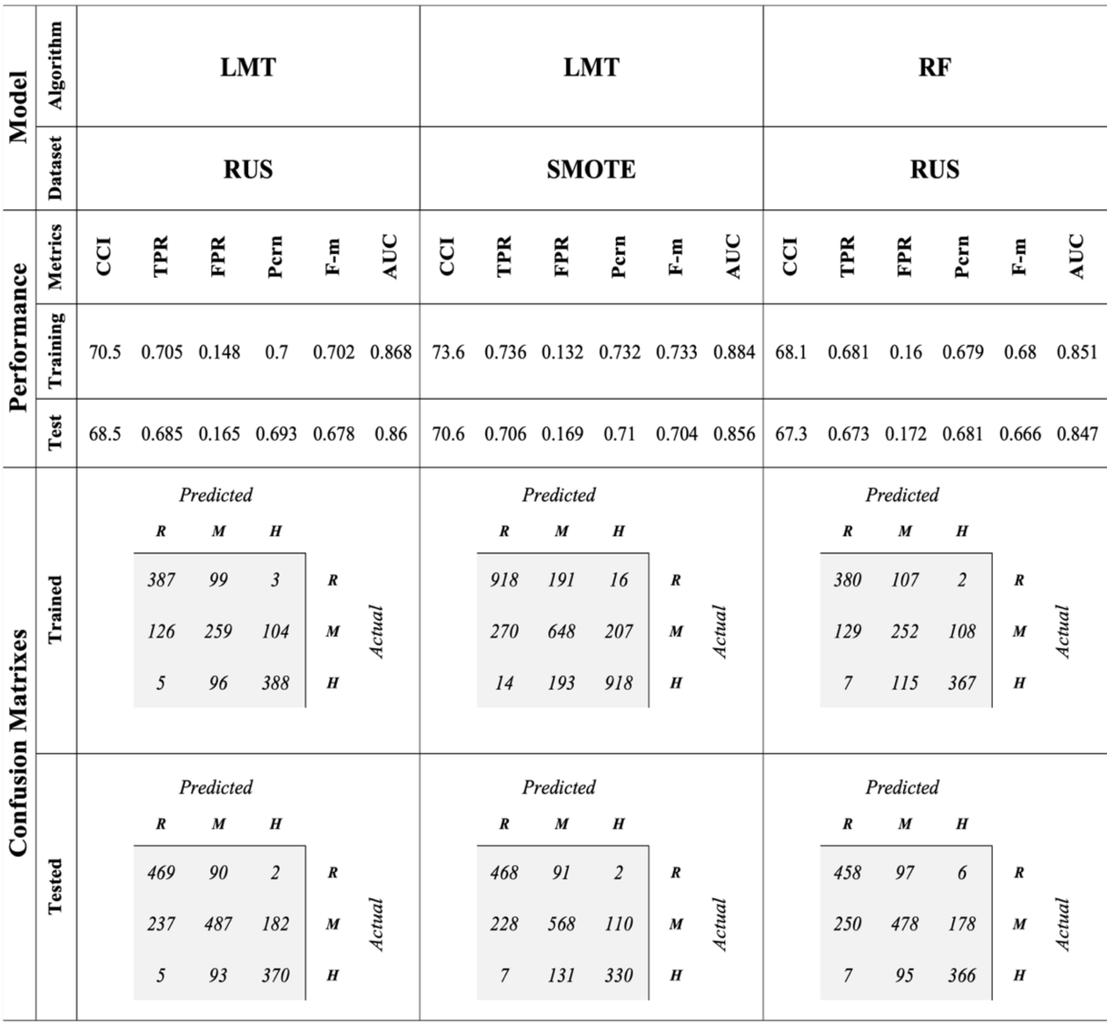


Fig. 5.28 Adaptive projection of all classes and the effect of sampling methods

5.3.12 Models' evaluation and selection.

In a similar approach to evaluating the models in the females' case study, the misclassification of healthy subjects by a predictive model could result in follow-up examinations, whilst false-negative misclassifications might result in an individual not receiving an important intervention. In this application, apart from potential cost, there would be few adverse effects associated with misclassified healthy/moderate risk subjects, as such subjects would be encouraged to undertake lifestyle-based interventions to improve their health. Therefore, in this scenario, the best models to adopt would be those which minimise the number of subjects misclassified as at ‘risk’, so they may initiate interventions at an appropriate time. Confusion matrices play an essential role in helping researchers define the best-suited model for use in future trials. Three models were identified as satisfying the domain experts' criteria when analysing the confusion matrices (Table 5.13) from the male cohort. These models are reported in Table 5.15. However, they may not necessarily occupy the highest ranks when their performance metrics are compared to the others.

Table 5.15. Domain-compliant prediction models for males. n (LMT RUS Trained) = 1467; n (LMT SMOTE Trained) = 3375; n (RF RUS Trained) = 1467. n (all Tested) = 1935. For the F-m metric, m=1



Model	Algorithm	LMT						LMT						RF					
	Dataset	RUS						SMOTE						RUS					
Performance	Metrics	CCI	TPR	FPR	Pcrn	F-m	AUC	CCI	TPR	FPR	Pcrn	F-m	AUC	CCI	TPR	FPR	Pcrn	F-m	AUC
	Training	70.5	0.705	0.148	0.7	0.702	0.868	73.6	0.736	0.132	0.732	0.733	0.884	68.1	0.681	0.16	0.679	0.68	0.851
	Test	68.5	0.685	0.165	0.693	0.678	0.86	70.6	0.706	0.169	0.71	0.704	0.856	67.3	0.673	0.172	0.681	0.666	0.847

Confusion Matrixes		Predicted				Predicted				Predicted			
		R	M	H		R	M	H		R	M	H	
Trained		387	99	3	R	918	191	16	R	380	107	2	R
		126	259	104	M Actual	270	648	207	M Actual	129	252	108	M Actual
		5	96	388	H	14	193	918	H	7	115	367	H
Tested		469	90	2	R	468	91	2	R	458	97	6	R
		237	487	182	M Actual	228	568	110	M Actual	250	478	178	M Actual
		5	93	370	H	7	131	330	H	7	95	366	H

5.4 VAT-associated diseases prediction case study discussion

The overall goal of this case study was to predict visceral adipose tissue (VAT) content in UKBB participants and apply machine learning methods to classify these subjects into risk categories. VAT has consistently been shown to be associated with the development of metabolic conditions such as coronary heart disease and type-2 diabetes. The ability to predict and classify this variable using simple anthropometry without the need for costly MRI scanning, will have a significant impact on the identification of subjects likely to benefit most from lifestyle-based interventions [346]. The models tested here input features that include age, waist and hip circumferences, weight, height, BMI, blood pressure and level of physical activity; all variables previously demonstrated to significantly correlate with VAT [347].

Previous UKBB studies [344][345] have demonstrated significant correlations of anthropometry measurement and physical activities with VAT, with significant gender differences in the distribution of VAT, as well as by age; Hence separate case studies were completed for female and male participants. In this case study, an index of physical activity, the UKBB-PAI, was proposed, which correlated more strongly with VAT outcomes than established questionnaires, such as the International physical activity questionnaire (IPAQ) and Lifestyle Index. Additionally, its findings challenged previous studies [347][348][349], and described only a weak correlation between age with VAT, even after adjusting for BMI and UKBB-PAI. It was also noted that the influence of UKBB-PAI parameters was comparable to that of age, and that it provided more effective means representing the physical activity measures to discriminate between Health, Moderate and Risk classes.

With domain experts' advice, the current case study selected the variables (age, blood pressure, body mass index, height, hip circumference, physical activity index, waist circumference and weight) as input features on which to base the machine learning VAT prediction models. However, to understand the influence of each feature on VAT and their reliability in predicting three distinct ranges associated with various long-term conditions, information theory was used to evaluate each feature in relation to the 3 different classes, Healthy, Moderate and at Risk. The IG evaluation algorithm was utilised to evaluate the worth of each input feature independently [194] against the class, unlike correlation analysis carried out in previous studies [344]. One way to interpret the calculated IG values is the possible presence of associations between each feature and the class labels in each cohort. In the female cohort, the strength of the association in the IG varies (see Tables 5.5), with HC, WC, BMI and W providing the greatest contribution, whilst the physical activity, age, H, SBP and DBP showed the least in both TD and RUS training datasets. A similar dominance in IG ranking is observed in the SMOTE dataset, with HC, WC, BMI and W showing the strongest associations with VAT. An analogous pattern was found in male subjects (Tables 5.14). This information theory approach into the models' features adds an additional layer of details to observed correlations reported in previous studies by describing the strength of each feature to discriminate between the Health, Moderate and at-Risk classes.

When considering the TD and RUS datasets which contain observations from the participants rather than generated artificial synthetic data, there was no association between age and VAT (given zero IG) in the males and only a weak association in the female cohort, with the discretised VAT ranges similar to the weak correlation described in previous studies. Though this may challenge previous studies that have reported a linear relationship between age and VAT [347] [348] [349], our results may reflect the somewhat smaller age range included in the UKBB (40-70yrs), compared with previous studies (17-70yrs) [347]. However, this may also relate to a data problem in the machine learning community known as data heterogeneity [350].

The lack of association of UKBB-PAI with discretised VAT classes reflects the previously reported [351] [352] low correlation between physical activity and these fat deposits and may, in part, arise from the poor reliability of the recorded frequencies and durations of physical activities. The level of granularity in the input data variables always determines the level of detail in the prediction model's possible outputs. Depending on the assessment design, detailed observations may be grouped during or after data collection into frequencies, categories and scores. This grouping is considered a variable transformation. Variable transformation aims to create better features at exposing patterns in the data. However, the transformation process could also lead to engineering a new feature that is less powerful, suppressing important trends offered by its detailed (raw) components.

It could also be argued that the implementation of such low-cost measures may lack the susceptibility to errors if studied within large populations [353]. However, there are many newly developed physical activities questionnaires (PAQs) which do not appear to perform substantially better than existing tests with regards to reliability and validity [354][355][356]. The variability of these PAQs and their ineffectiveness leads to a cause known in the data science community as *detail aggregation*. Variables in datasets often fall within two types; either detailed (Granular) or aggregated (Summaries). ML modelling prefers detailed variables over summary variables. Detailed data often represent summary variables and are better at showing patterns. Take daily walking which forms part of PAI calculations, as an example, previous studies [357] [358] showed that daily walking is linked to reductions in VAT. However, its significance is curbed when combined with other variables in PAI calculations. Data granularity is a macro structural feature. Granularity refers to the amount of detail captured in any measurement, such as time to the nearest minute, the nearest hour, or simply differentiating morning, afternoon, and night, for instance. Decisions about macro structure have an essential impact on the amount of information that a data set carries, which, in turn, significantly affects the resolution of any model built using that data set [53]. Therefore, we must acknowledge that physical activity is a complex behaviour that is hard to measure accurately, even at a low degree, in the case of memory recollection, or a high degree, by using electronic monitoring devices. However, it is a real challenge to record the interactions among physical activity's various elements (variables). The PAI structure that combines sets of variables with transformed scores could introduce bias, which stresses the natural structure of the original variables' states in a dataset so that the data is distorted. Hence, the PAI may be less representative of the real world than the original, unbiased variables form.

The understanding of the effect of data aggregation by domain experts enhances feature selection strategies of how variables are used in predictive modelling. Some derived (aggregated) variables may increase the representation of trends within a dataset which, in turn, show higher IG evaluation and pose as a stronger predictor in modelling. For example, BMI is directly obtained from height and weight (calculated as weight in kilograms (W) divided by height (H) in meters squared). From our analysis, H maintained its IG evaluation to zero in both RUS and TD datasets, by dividing body mass over two exponents of the base H, this seems to reveal better trends. Aggregated variables may require checking for calculation integrity from detailed variables. For numeric features, aggregated variables come in many forms, such as averages, sums, multiplication and ratios. Categorical features can be combined into a single feature containing combinations of different categories. Variable aggregation must not be overdone as not to overfit models due to misleading combined features. Wrongly derived variables may show false significance or insignificance in the analysis [53].

For machine learning modelling, tackling the imbalanced class problem has an important impact on the learning performance of standard machine learning algorithms. Classification performance in the training phase is severely impacted by class separability. Training standard ML algorithms with highly imbalanced overlapping classes without any adjustment to the training set results in an accuracy bias towards the majority class. In this case study, we observed that applying the two methods (RUS and SMOTE) was used to adjust the class imbalance in the classification training phase at the dataset level, which in turn, amplified the IG in many input features. It remains unclear as to whether other remedies for imbalanced data classifications, such as Cost-Sensitive and Ensembles Learning (which is implemented at an algorithmic level) or deep learning, could result in better performances [359] [360]. The advantages of sampling techniques evaluated here however include simplicity and transportability. Nevertheless, they are limited by the amount of IG manipulation due to their application resulting in biased predictions toward the minority class. Furthermore, the excessive use of such techniques could result in overfitting the models.

In this case study, for the female cohort case, the original dataset was highly imbalanced. Traditional ML algorithms were sensitive to higher information gains. As a result, they tended to produce superb performance results in training, but when testing the models, the overall model accuracy often dropped below the training phase performance.

However, for the male cohort, the class imbalance in the original dataset was less severe; therefore, traditional ML algorithms were less sensitive to higher information gains and tended to produce close performance results in training and test. The overall model accuracy often dropped below the training phase performance, which was the case for all models trained with the SMOTE dataset. On the contrary, the models' test accuracy outperformed the training accuracy when each algorithm was trained on TD; this situation also occurred in NB, LR and MLP trained with the RUS dataset. The cause of such competitive accuracy test results may be attributed to the increase in IG per feature in the test dataset compared to the TD (Figures 5.24 and 5.27). A higher IG in a variable indicates higher observations' purity per class. Having higher IG in multiple features enhances class separability and leads to improvement in classification accuracy. In other words, the higher the IG in a dataset, the easier the dataset to be learned and predicted.

In both cohorts, the UKBB datasets utilised in this case study showed that applying some level of sampling with slight disruption to the original data distribution, together with the desired choice of performance metrics and slight manipulation of IG levels, produced a good prediction solution which may be enhanced further with algorithmic modifications. Among all eighteen models for each cohort presented in this case study, six models satisfied the domain experts' success criteria for this specific domain problem. For the female cohort, these were LMT and RF built with RUS sampled dataset, and LR built with SMOTE sampled dataset. For the male cohort, they were LMT and RF built with RUS sampled dataset, and LMT built with SMOTE sampled dataset.

The difference in algorithms' learning schemes proved to have a minimal impact on the models' overall accuracy. This is because machine learning algorithms are biased towards achieving the highest model's accuracy. But the effect of the learning mechanism becomes primarily noticeable in imbalanced datasets when the minority classes' accuracies are compared (for example, Figure 5.8). In the testing results analysis, the learning schemes impact was seen to increase with the class imbalance severity in datasets compared to balanced datasets. But this difference in classifiers' performance may be driven by some underlying assumptions per algorithm learning mechanism design.

Observing the assumptions about the impact of IG manipulation has led us to make further investigations on the algorithms learning behaviour from datasets with features offering maximum class purity. Our analysis resulted in developing a new sampling approach, Minority Pattern Reconstruction (MPR). MPR improved the modelling accuracy of the Risk groups across all models by extracting multiple subsamples from the original training data TD. Each subsample would provide a perfect classification learning pattern with its own context in space. The isolation of the subsamples was followed by appending them into one new training set and discarding the residuals whose features' sum of IG is below the original IG sum in TD.

MPR was only applied to the females' case study since its fundamental assumption is limited to a multi-class problem, with some gradual attention to their classes, where the risk group (hyper-susceptible), the class of interest, is a minority governing the extraction of patterns leading the moderate risk subjects then the healthy. Part of the MPR assumption cannot be met in the males' case since the males' risk group are not a minority, hence not applicable, and it may not reach its potential. Perhaps with further research and experiments, MPR could reach new limits.

The modelling of the MPR set and its hybrid variants massively improved the classification TPR of the risk and moderate groups. MPR efficiency might be driven by its maintained proportion levels of information gain per feature compared to the TD.

The MPR resampled models reached a highly competitive performance level that also fits the success criteria. This competitiveness made it harder for domain experts to reselect/reassess the best-performing models within the given success criteria. While data scientists can observe MPR's massive improvement in the females' models, the domain experts may need to refine their success criteria further; perhaps by making them more restrictive. One way to restrict the selection criteria further is by obtaining the actual monetary costs of the additional tests required for H and M subjects upon misclassification. The availability of cost information may influence a data scientist to rethink new performance metrics alongside the fundamental measures (i.e., the confusion matrix), which could interpret future refined success criteria.

Moreover, estimating such costs lies outside this case study collaborating experts' knowledge and capacity and requires clinical expertise. Therefore, if a new collaboration is formed with clinicians in the field, it could become a natural extension to restart a new data mining cycle in this case study.

At the time of this study, this domain problem is the first to use the discretised MRI VAT variable ranges to describe participants' health status and label instances. Therefore, at the time, it would have been impractical to compare the results of this study to any other research from the same domain. However, this work will be followed by further analyses where additional methods to improve the outcomes will be investigated.

5.5 VAT-associated diseases prediction case study conclusion

Our study shows that applying our new OCTOPUS framework to model datasets of phenotype variables offers a fast and inexpensive solution to predict dangerous visceral fat levels as an adverse event in healthcare. This is achieved by aligning the classification task to predict specific VAT ranges. The selection of a multi-class prediction task in this study is strategic. It identifies individuals at higher risk of developing metabolic conditions and is more likely to benefit from focused lifestyle intervention to reduce visceral fat. The design of the case study of a multi-class prediction, by separating the risk group from a moderate group, helped select models that minimise incorrect classification of those at high risk as healthy. Achieving a zero False Negative Rate (FNR) when classifying risk patients as healthy guarantees that any individual to miss treatment intervention belongs to the moderate group rather than the risk group. Training various machine learning algorithms with 10-Fold Cross-Validation and testing the models with external class-stratified groups of females and males makes this study suitable for follow-up research in medical screening to identify subjects that may require treatment intervention.

Finally, the promising results of the selected models make them suitable for deployment as tools in real-life web applications. However, due to data limitations, we still think they should be preceded by further validation (testing) on an external cohort outside the UK Biobank population. In addition, the models' misclassifications may not immediately impact the subjects (modifications to lifestyle take time) but prompt them to make drug-free lifestyle adjustments. Nevertheless, should these applications ever be used for medical screening to prescribe medications for drug-induced lifestyle modifications, further external testing before deployment on external cohorts becomes mandatory, and performance monitoring becomes vital.

Chapter 6

Octopus Framework Application for Radiotherapy Side Effects Prediction

This chapter covers

- *Case Study 2: Predicting Breast Cancer Patients' Susceptibility to Radiotherapy Toxicity.*

This chapter applies our Octopus data-driven framework from our research to develop new predictive models as new tools to help detect patients' susceptibility to adverse events in healthcare, Advanced Radiation Therapy Acute Skin Toxicity Side Effect. We use the methods in the novel framework with trivial differences due to the nature of the data in this study, the domain success criteria, and the execution timeline. These new tools are built using new real-world datasets. Our models are new; at the time of execution, the modelling of the endpoints was not attempted by other case studies in the field using our data or predictors. Our focus in this chapter is to use the methods in our OCTOPUS framework, formulated by our heuristic-

* **Disclaimer:** *The author and collaborators assume no responsibility or liability for any errors or omissions in the data provided by the REQUITE consortium. The data is provided from their repositories on an "as is" basis with no guarantees of completeness, accuracy, usefulness, or timeliness. The authors assume no liability for the reliability of the results obtained from the use of this data. All data in these case studies were engineered and modelled from the provided "as is" data, at best, with collaborators' review. This practical work was accomplished by the author and collaborators in their personal academic, and professional capacities. The opinions expressed in this study are their own and do not reflect the view of the National Health Services (NHS), or the Department of Health in the United Kingdom.*

** **Disclaimer**: *This chapter reuses the published content of the following articles:*

- ❒ *Aldraimli, M., Soria, D., Grishchuck, D., Ingram, S., Lyon, R., Mistry, A., Oliveira, J., Samuel, R., Shelley, L.E., Osman, S. and Dwek, M.V., 2021. A data science approach for early-stage prediction of Patient's susceptibility to acute side effects of advanced radiotherapy. Computers in biology and medicine, 135, p.104624.*
- ❒ *Aldraimli, M., Osman, S., Grishchuck, D., Ingram, S., Lyon, R., Mistry, A., Oliveira, J., Samuel, R., Shelley, L.E., Soria, D. and Dwek, M.V., 2022. Development and Optimization of a Machine-Learning Prediction Model for Acute Desquamation After Breast Radiation Therapy in the Multicentred REQUITE Cohort. Advances in Radiation Oncology, 7(3), p.100890.*

systematic research and critique, to demonstrate its ability to create new inexpensive screening methods with Machine Learning approved by healthcare professionals and domain experts.

This is the second case study in this thesis to predict adverse events in healthcare. Here, we predict breast cancer patients' susceptibility to radiotherapy's acute toxicity side effects. We use our formulated data-driven approach to develop new machine learning models as new inexpensive screening tools to assist clinicians and physicians in triggering conversations during treatment planning about the potential occurrence of early side effects. The new machine learning models predict the occurrence of acute moist desquamation. This case study is a binary classification problem.

The newly developed models are designed to predict two susceptibility class labels, resistant or desquamation negative label for those who were observed not to develop this side effect and susceptible or desquamation positive label for those who were observed to develop acute moist desquamation during their radiotherapy. Radiotherapy has multiple side effects that depend on their severity and time of occurrence during treatment.

6.1 Introduction

Our focus aims to identify patients’ susceptibility to severe complications that can interrupt RT or even a total dose reduction. Such an interruption or reduction can potentially increase the risk of local cancer recurrence. However, the risk of cancer recurrence could be reduced if a patient’s susceptibility to radiation toxicity was better known to allow treatment plans to be personalised.

The latest strategies currently embedded within the treatment planning systems to determine the patient’s risk of radiation toxicity use mechanistic models [361]. Such models are based on a simplified characterisation of the interaction between radiation and biological tissues to explain the underlying mechanisms with explicit algorithms. Unfortunately, these algorithms are based on handcrafted rules with complex exceptions that often fail to predict the actual complications induced by RT.

The investigation of using ML in this field is still new. Recent studies used complex models to predict RT toxicities. One approach used radiomics data (thermal imaging data) on a small sample of patients [362]. This approach limits the large-scale analysis of RT toxicities due to the expense and time required to employ the requisite imaging techniques and the considerable variation between individual patients' normal tissue reaction to RT and resultant toxicities [363]. A different approach utilised hundreds of clinical variables as model inputs raising an issue in interpretability [364] [365].

The REQUITE study provides a comprehensive means of assessing the relationship between the patient's baseline characteristics, medical history, clinical, genomic, dosimetric and radiomic variables and RT range of toxicity outcomes in a large population-based cohort of breast cancer patients [18]. Having such a large dataset could increase the presence of a pattern in the data; without it, machine learning algorithms can't sufficiently learn to produce effective results.

The primary goal of this case study is to identify an inexpensive and clinically valid ML prediction model to predict the occurrence of acute desquamation in the REQUITE breast cancer cohort.

In this case study, we apply the methods scoped in our developed OCTOPUS framework and assess the classification performance of eight machine learning algorithms (Naïve Bayes, Logistic Regression, Artificial Neural Network, Decision Tree, Logistic Model Tree, K-Nearest Neighbour, Support Vector Machine and Random Forest) in predicting the susceptibility to the occurrence of acute desquamation. The Radiation Therapy clinicians from the national collaboration of Radiation Therapy Machine Learning Network manually labelled all records for acute toxicity outcomes based on a CTCAE v4.0 endpoint definition. The endpoint definition is associated with the susceptibility of developing an early RT skin side effect known as acute moist desquamation.

The dataset exhibits class imbalance which could result in the problem of imbalanced learning bias. Therefore, the new models were first built using the Imbalanced Dataset, followed by applying three sampling techniques, Random Under-Sampling (RUS) and Synthetic Minority Over Sampling Technique (SMOTE). In addition, Cost-Sensitive (CS) learning was also applied in our modelling approach.

This study suggests the most suitable models meeting the domain experts' success criteria. The data imbalance characteristic causing the transition in classifier training performance was captured visually by Adaptive Projection Analysis (APA) in the Targeted Projection Pursuit (TPP) [48] and numerically via Information Gain (IG) attribute evaluation and Mean Decrease Impurity (MDI). The development of our new framework, its results from this study and selected models by domain experts in the RTML collaboration at the University of Manchester were first published in 2021 in the Journal of Computers in Biology and Medicine [151].

6.2 Predicting susceptibility to acute desquamation workflow

The deployment of machine learning modelling in this study aims to apply our new data-driven framework to model acute moist desquamation as an adverse event, compare a large number of prediction models' performances and select the best-suited models when applying multiple imbalanced modelling remedies in a clinical setting. The newly developed models effectively tackle a real-world treatment management challenge by predicting acute desquamation, an early-stage RT toxicity.

Early-stage radiation toxicities occur during treatment or within ninety days of exposure to RT. The patient may have skin changes ranging from desquamation (peeling skin) to skin necrosis (death of skin cells) and ulceration. These changes imply that the skin integrity has been broken over the breast or in the inframammary fold. Patients with such toxicities experience irritation, pain and serious fluid build-up under the skin, impacting their Quality of Life (QoL) [366]. Therefore, RT-treated patients' QoL has become an increasingly important research priority [366]. RT reduces the rates of cancer recurrence and increases long-term survival. Hence over 70% of breast cancer patients receive RT during treatment [367]. Typically, the incidence rate of acute desquamation range between 11% to 71% in breast cancer RT patients [366]. The original REQUITE dataset underwent rigorous data preparation and pre-processing by the RTML network specialists, followed by the modelling, evaluation and simplification phases (Figure 6.1).

The imbalanced training dataset (ITD, n=1029, m=123) was used to train eight algorithms to establish the extent of the class imbalance modelling problem. Once verified, two different strategies were used to mitigate the issue. In one strategy, ITD (n=1029, m=123) was modified with sampling techniques, SMOTE (n=1866, m=123), ROS (n=1866, m=123) and RUS (n=192, m=123), and used for training eight ML algorithms (Naïve Bayes, Support Vector Machine, Logistic Regression, Artificial Neural Network, C4.5 Decision Tree, Logistic Model Tree, Random Forest and K-Nearest Neighbour).

At a later strategy, ITD (n=1029, m=123) was used to train three systematically nominated ML algorithms with a cost-sensitive approach inducing multiple misclassification penalty matrices. All models were tested with the same isolated validation data (VD, n=1029, m=123). The models' selected performance metrics were compared after test to identify the model of interest to clinicians and oncologists.

The chosen hero model interpretability was simplified and concluded as a final preclinical-valid model. IG was monitored for all predictor variables at every stage.

Clinical machine learning studies are often criticised for the lack of transparency regarding the methods used to prepare and pre-process their data before modelling. Therefore, to uphold the clinical validity of our final model and the confirmatory nature of this study, the description of the data preparation and pre-processing procedures followed is presented with reasonable details [368].

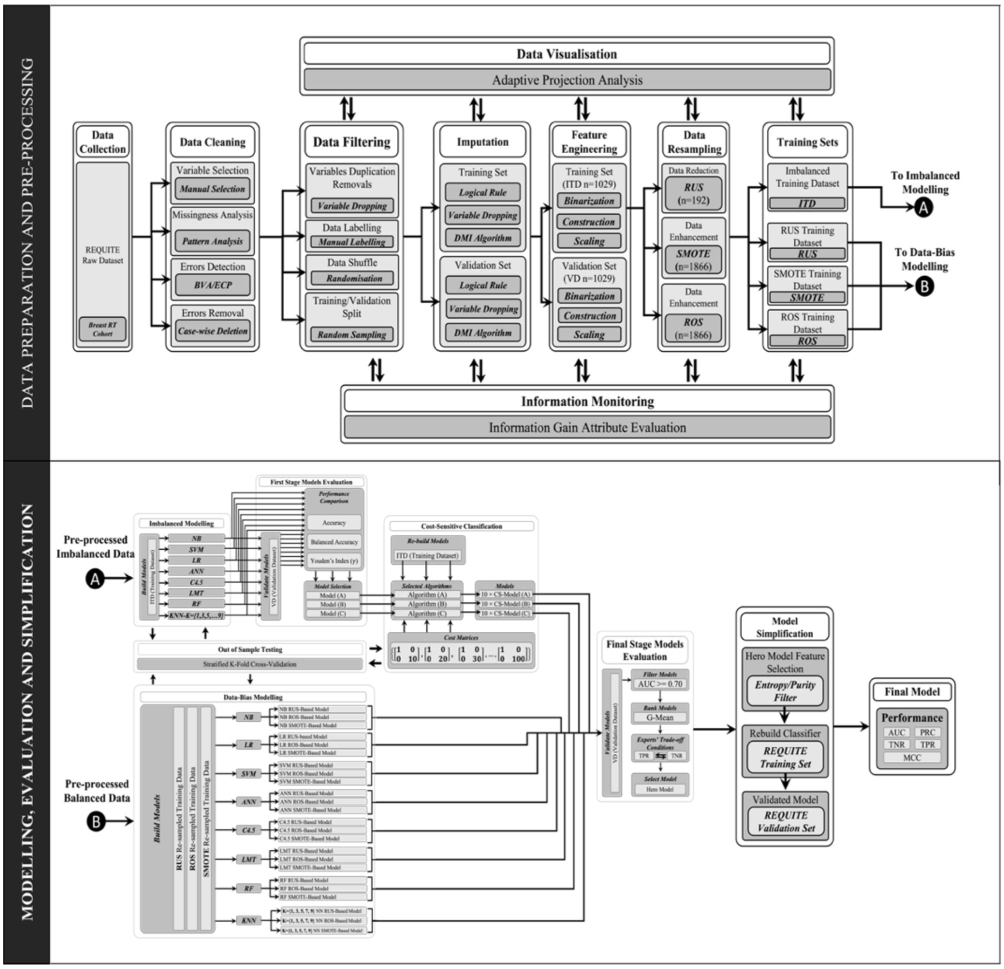


Fig. 6.1 Workflow followed to predict susceptibility to acute RT desquamation.

6.3 Data collection protocol

REQUITE is an international prospective cohort study that recruited cancer patients in 26 hospitals in eight countries. This study uses collected data from patients who underwent breast RT. The multicentre breast cancer patients cohort was recruited prospectively in seven European countries and the US. All patients gave written informed consent [369]. The study was approved by local ethics committees in participating countries and registered at the ISRCTN registry [370] (ISRCTN98496463). The study is a cross-sectional assessment of 2069 patients from the REQUITE international multicentre cohort, aged 23-80 and treated with breast RT between April 2014 and March 2017.

6.4 Experimental design

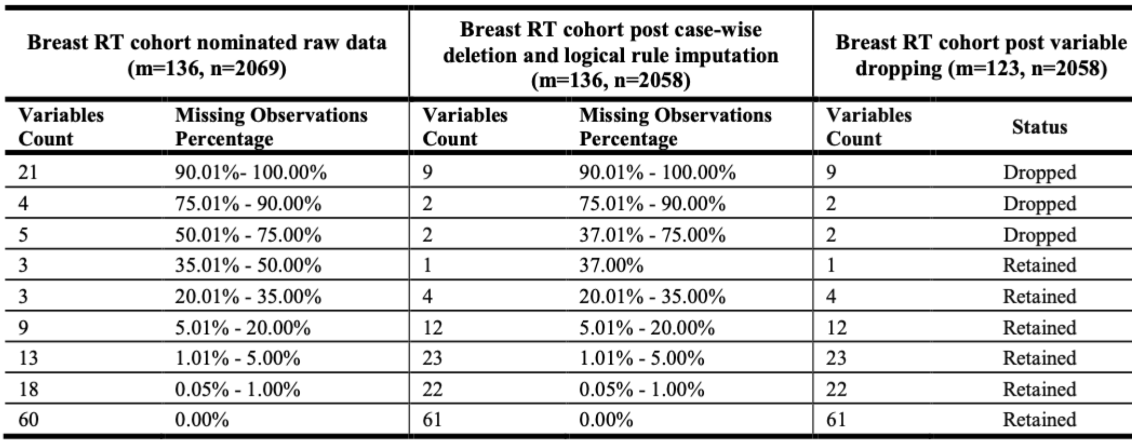
For the perdition of RT susceptibility to acute early side effects, binary-class ML classification models were applied to predict susceptibility to acute desquamation based on the outcome collected at the end of radiation treatment for REQUITE breast cancer patients. The raw REQUITE dataset ($n = 2069$) contained ($m > 300$) variables. The RTML clinicians manually labelled all records for acute desquamation outcome based on the CTCAE v4.0 endpoint definition: grade $1 \geq$ ulceration or grade ≥ 3 erythema. All variables were nominated manually in modelling acute desquamation by clinicians and RT physicists. Only an initial set of $m = 136$ applicable variables and $n = 2058$ (Susceptible patients or $D\text{esq}+ = 192$, Resistant Patients or $D\text{esq}- = 1866$) records remained (Case-wise deletion (n=11 with missing class label). Finally, after the initial analysis, a highly imbalanced dataset (n=2058, m=123) was deemed viable for modelling and evaluation. The descriptive statistics of the REQUITE dataset variables are reported in a previous study [18]. The input variables used in this study are easily obtainable during the treatment planning phase. They consist of baseline characteristics, familial history, breast cancer staging information, chemotherapy regimens, lifestyle attributes, medical conditions, sociodemographic factors, medical operations, treatment history, female-specific factors, psychological health attributes, medications, breast RT dosimetry measurements such as normo-fractionation procedure, and quality of life output. Radiomic data (imaging data) and genomics were not used in this study.

6.5 Pre-processing the training features

Initial exploration of the dataset showed that a number of numerical type variables exhibited a multimodal distribution, such as comb, edge peak and plateau-like distributions. Such distributions can indicate the presence of several patterns of response in the sample. The effect of such distributions and multi-pattern presence in the sample may lead to many outliers or extreme values. Therefore, in the data preparation, a rigorous cycle of Boundary Value Analysis (BVA) and Equivalence Class Partitioning (EPC) techniques [219] [220] were used by domain experts for detecting and correcting or removing corrupt or inaccurate records from the dataset. Also, missingness analysis was performed using our new Multi-Method Imputation (MMI) approach by cross-checking the data with the REQUITE study questionnaire design to ascertain the causes of incomplete records and deduce patterns.

With MMI, a combination of non-statistical and statistical imputation techniques was used, and non-statistical methods were used to reduce uncertainty via logical rule imputation and variable dropping (see Table 6.1). The investigation of missing data patterns [371] assisted in the non-statistical imputation of missing data with logical rule imputation, variable dropping (m=13 with > 37% missing values at random compared to observed values in the remaining variables to avoid introducing correlation bias when statistical imputation techniques are used). Although in MMI, the threshold of missing values to drop a variable is set at 40% for confirmatory analysis, clinicians felt this is still too high to incorporate imputed values at that level in the analysis. The retained dataset for feature engineering and modelling finally had m=123 variables and n=2058 records.

Table 6.1. Percentage of imputed missing data in breast RT cohort variables



Breast RT cohort nominated raw data (m=136, n=2069)		Breast RT cohort post case-wise deletion and logical rule imputation (m=136, n=2058)		Breast RT cohort post variable dropping (m=123, n=2058)	
Variables Count	Missing Observations Percentage	Variables Count	Missing Observations Percentage	Variables Count	Status
21	90.01%- 100.00%	9	90.01% - 100.00%	9	Dropped
4	75.01% - 90.00%	2	75.01% - 90.00%	2	Dropped
5	50.01% - 75.00%	2	37.01% - 75.00%	2	Dropped
3	35.01% - 50.00%	1	37.00%	1	Retained
3	20.01% - 35.00%	4	20.01% - 35.00%	4	Retained
9	5.01% - 20.00%	12	5.01% - 20.00%	12	Retained
13	1.01% - 5.00%	23	1.01% - 5.00%	23	Retained
18	0.05% - 1.00%	22	0.05% - 1.00%	22	Retained
60	0.00%	61	0.00%	61	Retained

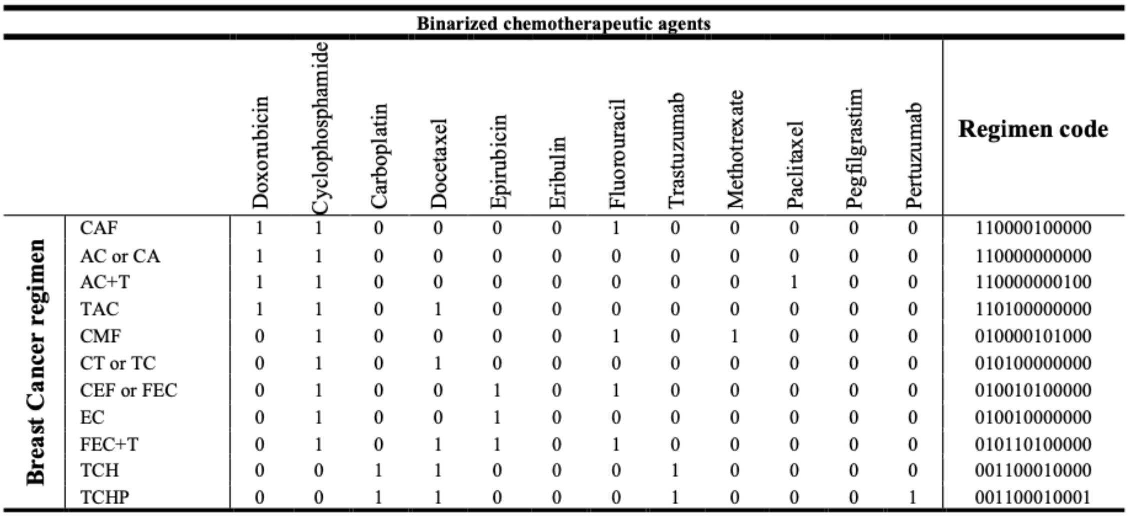
The retained records n=2058 were shuffled with a randomisation algorithm. Following randomisation, a 50:50 training: test split ratio with class stratification was performed. The split formed the raw Imbalanced Training Dataset (raw ITD, n=1029) and the raw test Dataset (raw VD, n=1029). The process was followed by applying a state-of-the-art hybrid Expectation-Maximization (EM)-Decision Tree imputation for each set independently with a Decision-Tree based Missing-Value Imputation (DMI) Algorithm [108] to enhance the best expectations of missing values. Datasets' information levels were monitored in each set pre-imputation (raw(ITD), raw(VD)) and post-imputation (DMI(ITD) and DMI(VD)) with Information Gain Attribute Evaluation [194]. The evaluation of information worth is affected by the number of records; hence, the 50:50 training-test split allows for a fair information bias comparison.

The retained 123 variables for modelling consisted of 106 raw features and sixteen additional engineered features. Breast size measurements are calculated as a single continuous variable by adding bra cup and band sizes to represent 'sister' sizes equal to the same breast volume [372]. For instance, a UK-size 34B bra holds an approximate breast volume equal to 32C, approximately 390 cc.

With feature engineering, sixteen features were constructed. In many patients, the chemotherapy regimens consisted of a combination of cytotoxic agents. In order to account for the vast number of possible chemotherapeutic combinations that patients could be prescribed, the prescriptions were binarised [61] based on their generic chemical names (see Table 6.2).

In addition, the chemotherapy drugs' categorical values were binarised, a format that could be provided to machine learning algorithms to improve prediction performance [373]. The categorical values represent the administered chemo-drug combinations in a chemotherapy regime.

Table 6.2. Illustration examples of binarised chemotherapy regimens



		Binarized chemotherapeutic agents												
		Doxorubicin	Cyclophosphamide	Carboplatin	Docetaxel	Epirubicin	Eribulin	Fluorouracil	Trastuzumab	Methotrexate	Paclitaxel	Pegfilgrastim	Pertuzumab	**Regimen code**
Breast Cancer regimen	CAF	1	1	0	0	0	0	1	0	0	0	0	0	110000100000
	AC or CA	1	1	0	0	0	0	0	0	0	0	0	0	110000000000
	AC+T	1	1	0	0	0	0	0	0	0	1	0	0	110000000100
	TAC	1	1	0	1	0	0	0	0	0	0	0	0	110100000000
	CMF	0	1	0	0	0	0	1	0	1	0	0	0	010000101000
	CT or TC	0	1	0	1	0	0	0	0	0	0	0	0	010100000000
	CEF or FEC	0	1	0	0	1	0	1	0	0	0	0	0	010010100000
	EC	0	1	0	0	1	0	0	0	0	0	0	0	010010000000
	FEC+T	0	1	0	1	1	0	1	0	0	0	0	0	010110100000
	TCH	0	0	1	1	0	0	0	1	0	0	0	0	001100010000
	TCHP	0	0	1	1	0	0	0	1	0	0	0	1	001100010001

Chemotherapy can be neoadjuvant and adjuvant. Neoadjuvant therapy is performed before the primary treatment to help reduce the size of a tumour or kill cancer cells that have spread, generally given before the surgical procedure. Adjuvant therapy is administered after the primary treatment to destroy the remaining cancer cells to prevent a possible cancer recurrence. In many cases, chemotherapy drugs (agents) are administered in combinations, which means the patient receives two or three different medicines simultaneously.

These combinations are known as chemotherapy regimens. Every cancer responds differently to chemotherapy. Standard breast cancer chemotherapy regimens include AT, AC, AC+T, CMF, CEF, CAF, TAC and others [374]. NHS UK published a wide range of chemotherapy side effects that may occur in breast cancer patients, some of whom may have plans to undergo breast RT [375]. Therefore, including chemotherapy attributes in this study was recommended.

To adjust for different RT regimens, the dose was calculated as the biologically effective dose (BED). BED is the product of the number of fractions (n), dose per fraction (d), and a factor determined by the dose and α/β ratio for acute effects (10 Gy), which is used in radiobiology to describe $dd$the slope of the cell survival curve for different irradiated tissues [376]. Three features were constructed by calculating the BED.

$$\boldsymbol{BED = n\,d\,\left(1 + \frac{d}{\alpha/\beta}\right)}$$

Out of all 123 variables, all numeric features (m=62) were normalised with $\mathcal{Z}$-score standardisation to eliminate the impact of larger magnitudes variables when modelling with a scale-sensitive algorithm [377].

6.6 Inspecting irregular training records

Inter Quartile Range (IQR) test was applied per attribute to look for defined values with an Extreme Value Factor (EVF) of 6 times the IQR and an Outlier Factor (OF) of 3 times the IQR.

A large number of outliers and fewer extreme values were detected (see Figures 6.2A and 6.2B). Nevertheless, experts proceeded with caution; thus, these data points were carefully verified, and a decision was made to retain them. These outliers may be attributed to the multimodal distribution in some variables, especially those from device readings.

In normal circumstances, outliers and extreme values are rare. If they are not rare in a dataset, such as this case, they may not be abnormal observations or the data collection methods may have had some integrity issues. The RTML team reviewed the documented data collection design protocols, and no problems were spotted. The abnormal data points verification was based on RT clinicians' empirical review of the REQUITE survey design. No additional dataset was available to contribute to the verification of such points. Another assumption was made that these outliers could benefit the model. If modelling proceeded with removing such a large number of outliers, modelling with the remaining data may underestimate the variance and overpower the bias in the model resulting in potentially under-fitted models.

Moreover, since there are no known external reasons (measurement error or incorrect recording) to influence the elimination of the large count of outliers and extreme values, these points should instead be used to provoke the model. If the model is evaluated as appropriate, then the retention of such outliers would have prevented us from underestimating the variance (in our case). In other cases, if they are rare, a choice may consider them negligible to retain or remove [378].

6.7 Training instances resampling

The REQUITE dataset shows that in a breast radiation treatment, only a small portion of patients suffered from acute desquamation [379], raising a potential class imbalance problem. Class imbalance poses an additional barrier to using ML algorithms. These algorithms usually are optimised using loss functions that attribute the same importance to all samples in the training dataset regardless of its endpoint. Therefore, the trained ML model will include a strong bias towards the majority class. As per our formulated approach, one strategy to tackle class imbalance in the training data is to apply data resampling techniques to $\mathrm{ITD} \equiv \mathrm{DMI}(\mathrm{ITD})$, by which the endpoint response classes of records become equal (see Figure 6.3); Random Under Sampling (RUS) (n=192, $D\mathrm{esq}^{+}$ = 96, $D\mathrm{esq}^{-}$ = 96), Random Over Sampling (ROS) (n = 1866, $D\mathrm{esq}^{+}$ = 933, $D\mathrm{esq}^{-}$ = 933) and Synthetic Minority Oversampling Technique (SMOTE) (n = 1866, $D\mathrm{esq}^{+}$ = 933, $D\mathrm{esq}^{-}$ = 933).

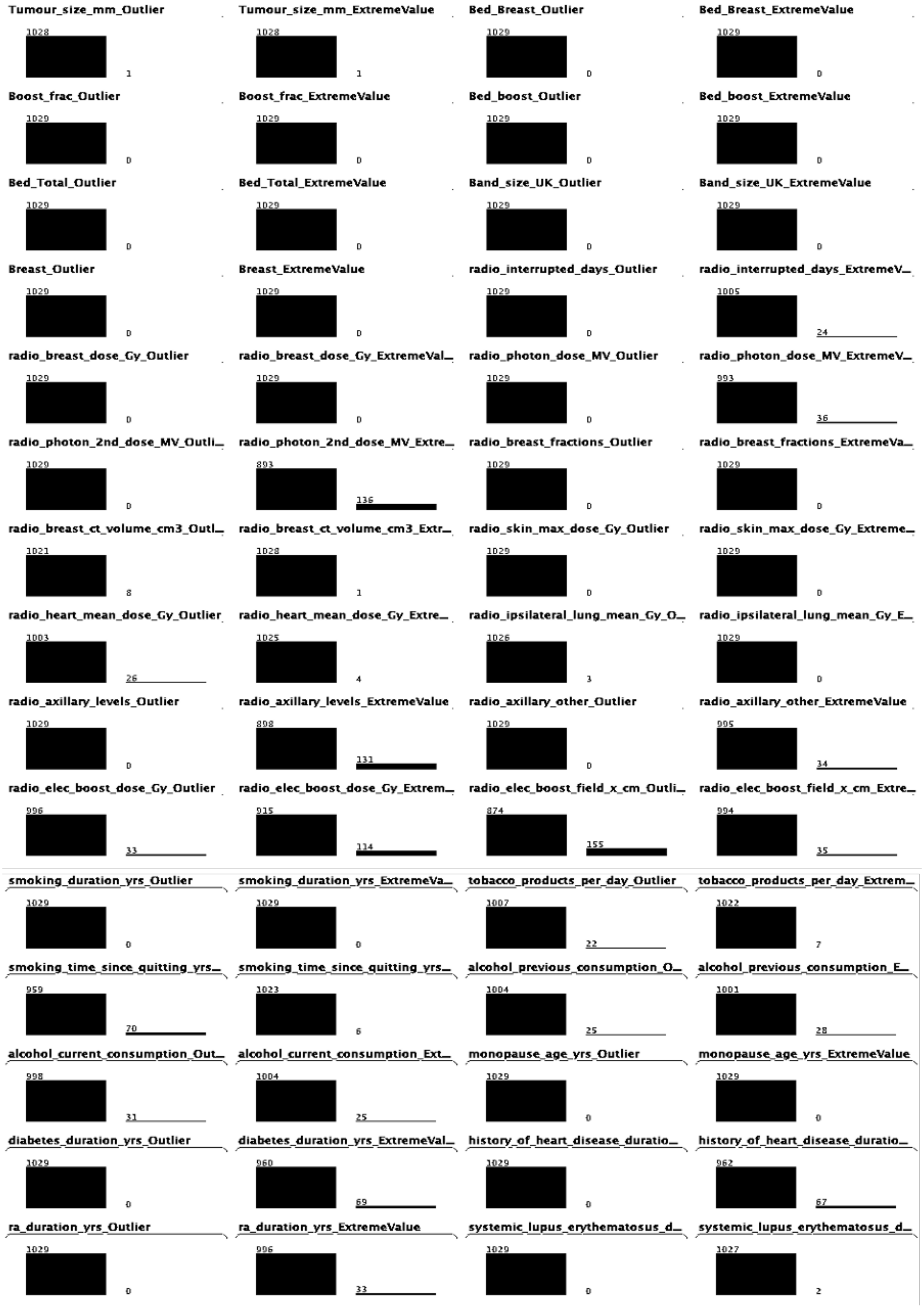


Fig. 6.2A Outliers and extreme values detection in the imbalanced targeted dataset

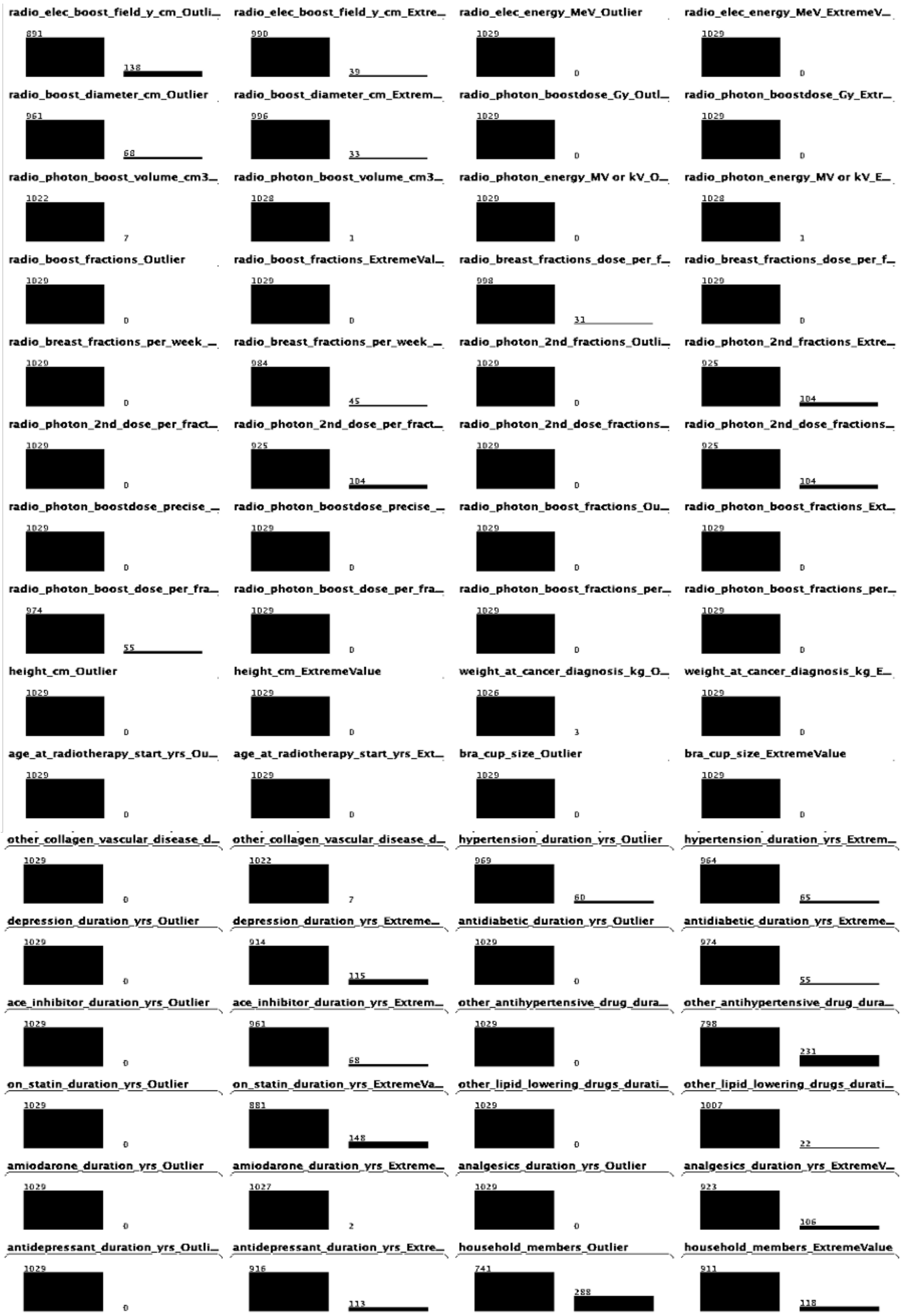


Fig. 6.2B Outliers and extreme values detection in the imbalanced targeted dataset

The effect of such resampling techniques on the training dataset was visualised with a multi-dimensional Adaptive Projection Algorithm (APA) into a 3D point cloud (see Figure 6.4).

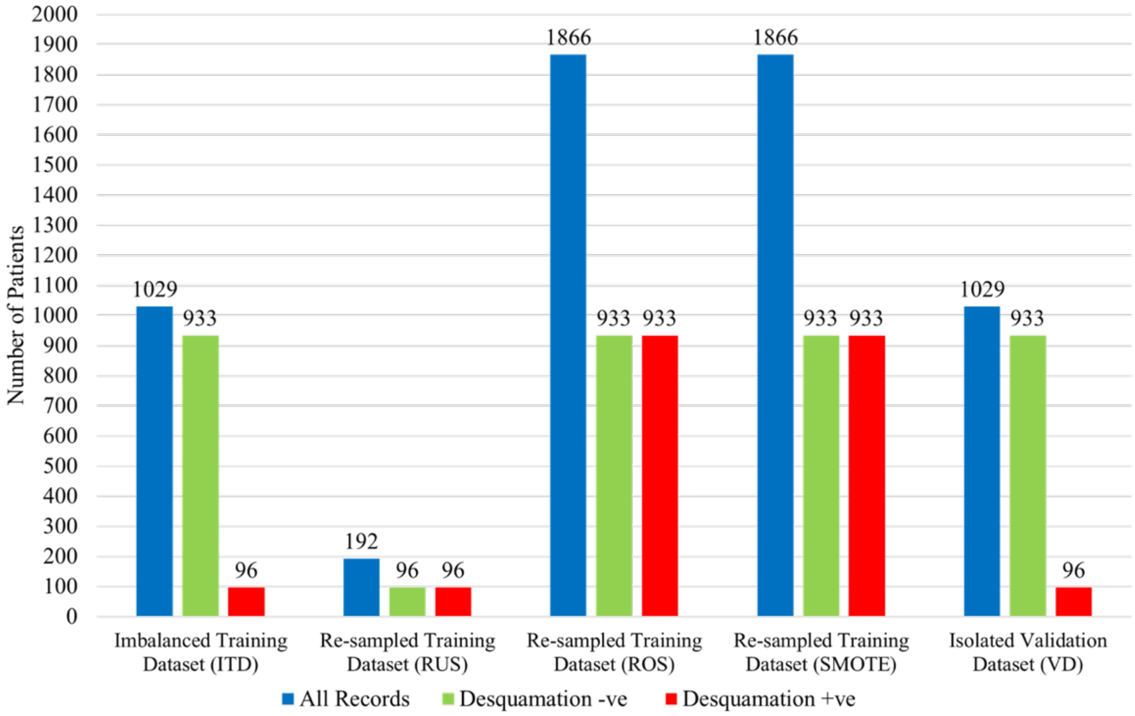


Fig. 6.3 Sample size for ITD, RUS, ROS, SMOTE and VD sets

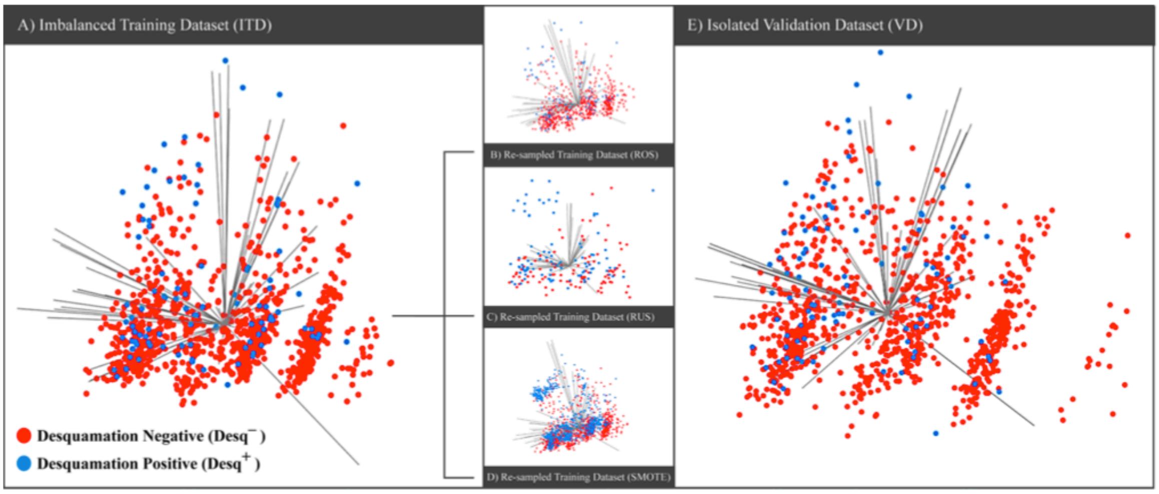


Fig. 6.4 TPP visualisation of ITD, RUS, ROS, SMOTE and VD sets

6.8 Test data characteristics

The ML models were tested on a new group of 1029 records from the REQUITE cohort (based on the original 50:50 class-stratified split without replacement). Like the ITD, the Test (Validation by domain definition) Dataset (VD) was severely imbalanced: the resistant group (*D*esq^{-}) had 933 subjects, and the susceptible class (*D*esq^{+}) had 96 patients (see Figure 6.3).

The TPP (APA) visualisation in Figure 6.4 indicates the classes that can be separated, the attribute combinations primarily associated with each group, the outliers, the sources of error in the classification algorithms, and the existence of clusters in the data. In this case, the APA shows a high degree of overlap of the variable's values between patients with and without desquamation, suggesting that it could be difficult to differentiate these two classes using these variables. Additionally, the visualisation of the ITD highlights the imbalance in the data (Figure 6.4 A) and how resampling techniques achieve the class balance (see Figure 6.4 B, C and D).

The ROS training dataset shows somewhat widely scattered positive class records since the ROS resampling technique randomly duplicated records from the positive class. SMOTE resampling technique has intensified the existing positive class records by generating synthetic prototype records analogous to the positive class records, and these records seem to cluster near the original positive examples.

The RUS visualisation depicts how a balanced dataset may expose divisions within the data more clearly, e.g., desquamation samples on top of the RUS visualisation seem to be easily separable. At the same time, in the ITD, ROS and SMOTE, it is difficult to observe a clear division between classes.

Moreover, the APA analysis shows that the ITD and VD are somewhat similar (consistent in the visual pattern), especially in the majority pattern. This similarity suggests that the randomised data split did not introduce any major bias into either split (training and test sets) and that the training dataset is visually closely representative of the whole data.

6.9 Inspecting irregular test records

Similar to the training ITD data, Inter Quartile Range (IQR) test was calculated for VD's attribute to look for defined values with an Extreme Value Factor (EVF) of 6 times the IQR and an Outlier Factor (OF) of 3 times the IQR. Although, a large number of outliers and fewer extreme values were detected (see Figures 7.5-A and 7.5-B). Despite the original data points being carefully examined by the domain experts to establish their validity, we proceeded with caution; thus, these data points were carefully verified, and a decision was made to retain them.

Again, these outliers may be attributed to the multimodal distribution in some variables, especially those from device readings.

And to reiterate, in normal circumstances, outliers and extreme values are rare in datasets in general. If they are not rare in a dataset, such as this case, they may not be abnormal observations, or the data collection methods may have had some issues.

The RTML team reviewed the documented data collection design protocols, and no problems were spotted. The abnormal data points verification in the test set is based on RT clinicians' empirical review of the REQUITE survey design. No additional dataset was available to contribute to the verification of such points. Another assumption was made that these outliers could test the robustness of the model. Suppose testing proceeded with removing such a large number of outliers. In that case, the remaining data might overestimate the model's performance overlooking an important pattern, making the testing inefficient in evaluating all patterns captured by the model.

Moreover, since no known external reasons (measurement error or incorrect recording) influence the elimination of the outliers and extreme values, these points should be used to test the classifier. If the model is evaluated as appropriate, then the retention of such outliers would have prevented an overestimation of performance. In other cases, if they are rare, a choice may be made to consider them negligible to retain or remove [378].

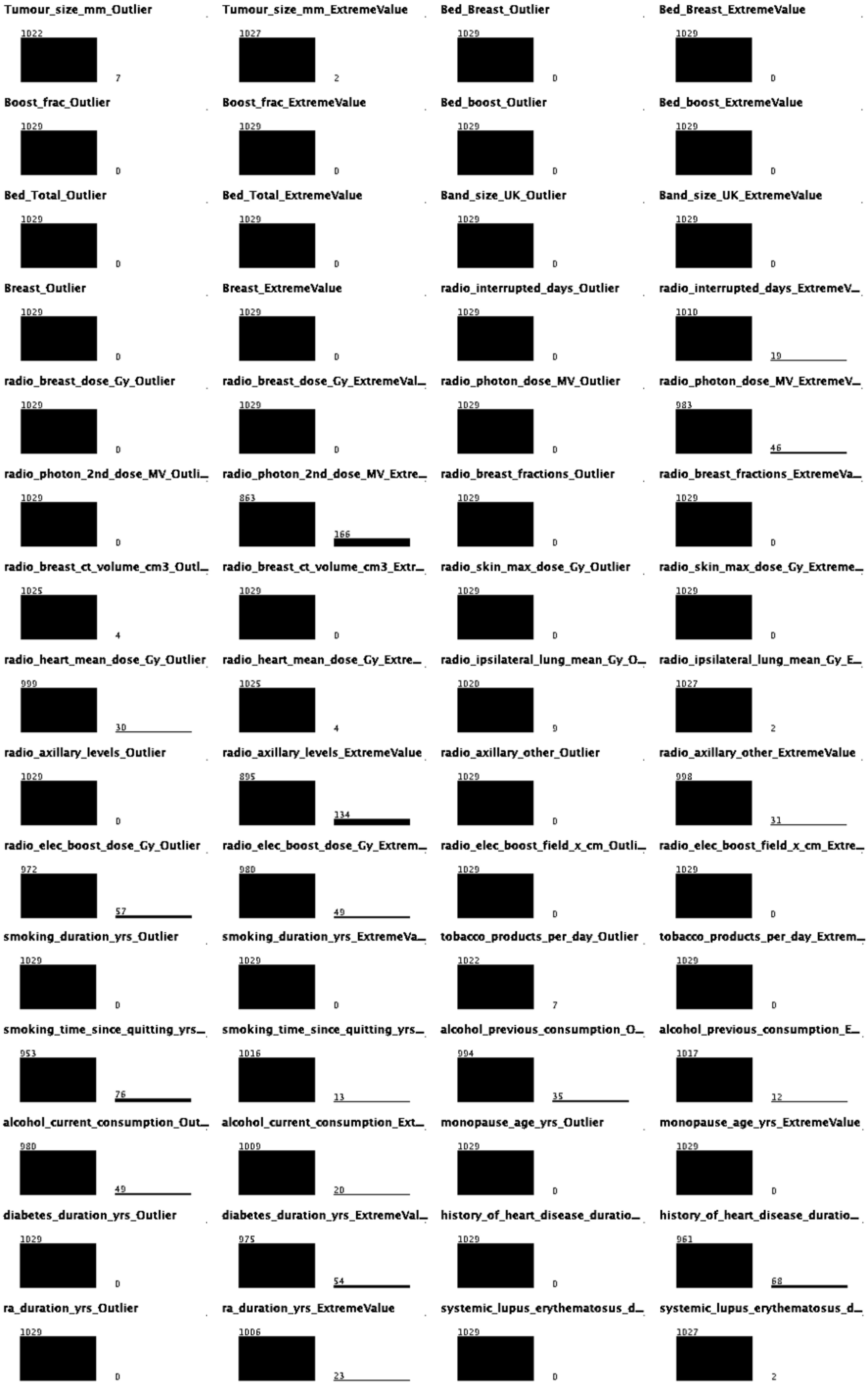


Fig. 6.5A Outliers and extreme values count per variable in the test set (VD)

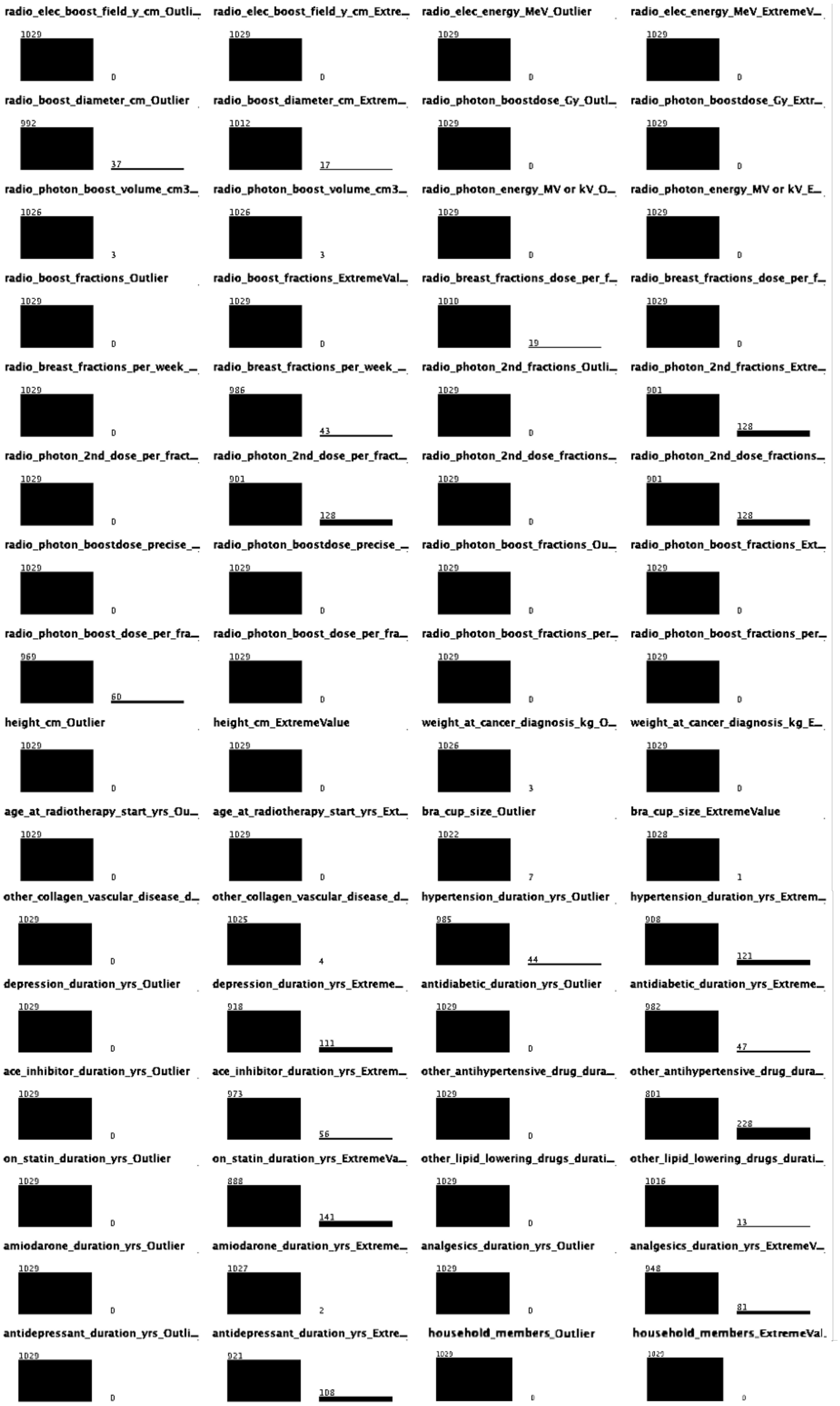


Fig. 6.5B Outliers and extreme values detection in the test set (VD)

6.10 Modelling susceptibility to acute desquamation

In this phase, we model the engineered REQUITE data in three different modes, the imbalanced, the balanced and the Cost-Sensitive modes. The data is modelled with eight different machine learning algorithms, Naïve Bays (NB), Logistic Regression (LR), Artificial Neural Networks (ANN) with Multi-layer Perceptron (MLP) architecture, K Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Trees (C4.5), Random Forest (RF) and Logistic Model Tree (LMT).

6.10.1 Imbalanced data modelling results

A single model was built with ITD and tested with VD for each of the eight ML algorithms. The KNN was an exception, for which five models were constructed with ITD and tested with VD to account for the different values of the K parameter, where K={1,3,5,7,9} from our methodology in the previous chapter. Table 6.3 shows the models' Accuracy, Balanced Accuracy, Youden's Index, AUC score, TPR and TNR performances for all twelve models in training and validation.

The training and validation performance results (Table 6.3) confirm the problem of the class imbalance issue with a severe high accuracy bias towards the desquamation-negative group (majority class) by sacrificing the desquamation-positive records (minority class) as type II errors (FN). In terms of training accuracy, (K=9) NN ranked first, scoring 0.909, while NB, a popular algorithm in medical research, came last with 0.776. Similar behaviour of accuracy performance ranking was observed after the test.

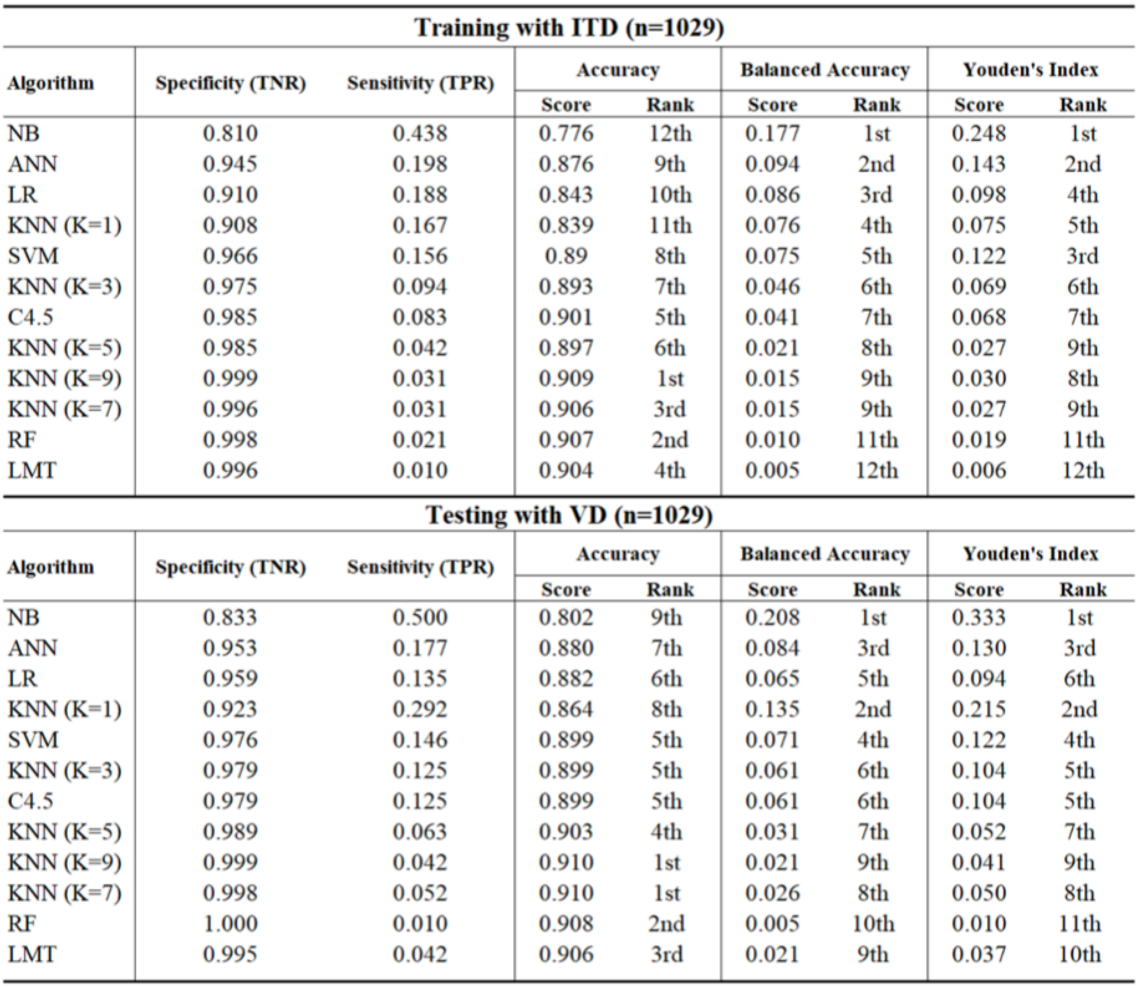
The balanced accuracy metric exposes classifiers that take advantage of the majority class to boost their overall accuracy. Conversely, the lower the balanced accuracy, the least a classifier takes advantage of the distribution of the majority class. Youden's index (γ) evaluates the ability of a classifier to avoid misclassifications in both classes. A higher value of γ indicates a good-performing classifier.

When analysing the training performances in Table 6.3, NB scored the highest in balanced accuracy and Youden's index (ɣ) to be considered the least susceptible classifier to accuracy bias towards the majority class and the best in avoiding misclassification. On the other hand, despite the high accuracy of the LMT model of 0.904, both balanced accuracy and Youden's index (ɣ) metrics agreed to rank it last. The low ranking indicates that the LMT model mainly took advantage of the majority class distribution to boost its accuracy score with a TNR of 0.996. The worst model in misclassification avoidance in both classes proved with the lowest TPR of 0.01. RF, a popular ensemble algorithm in the data science community for its accomplishments, missed the lowest performance on both the balanced accuracy metric and Youden's index (ɣ) and ranked just before LMT with 0.10 and 0.019, respectively, showing its severe bias towards the majority class.

By analysing the training performances of the classifiers with ITD in Table 6.3, the question of class importance in this particular domain problem arises when selecting algorithms for further improvement with a CS strategy. The higher the balanced accuracy and Youden's index (ɣ), the higher degree of discrimination between both classes in the imbalanced setting. In contrast, the lowest measurements on the same two metrics indicate the lowest degree of discrimination of the minority group.

In severe binary imbalanced learning, typically, a cost matrix in a CS approach penalises misclassifications of the minority group members to seek an improved TPR. Selecting the NB algorithm for CS modelling based on its balanced accuracy training performance with ITD may favour the minority group over the majority class. On the one hand, it allows its TNR performance to worsen from the lowest level of 0.810 among all classifiers to produce a higher TPR. On the other hand, selecting the worst-performing algorithm on both balanced accuracy and Youden's index (ɣ) in training, i.e., LMT or RF, for CS modelling, may indicate caring about both classes equally. Any improvement to their TPR may decrease the highest level of TNR from 0.995 and 1.000, respectively. The previous assumption can be valid if all learners in this study are to show the same depth of improvement to (TPR) and deterioration of (TNR) when presented with the exact cost (penalty) combinations in an explicit cost matrix penalising misclassification in the desquamation-positive minority group.

Table 6.3. Imbalanced ML models' training and test performances



Training with ITD (n=1029)

Algorithm	Specificity (TNR)	Sensitivity (TPR)	Accuracy		Balanced Accuracy		Youden's Index	
			Score	Rank	Score	Rank	Score	Rank
NB	0.810	0.438	0.776	12th	0.177	1st	0.248	1st
ANN	0.945	0.198	0.876	9th	0.094	2nd	0.143	2nd
LR	0.910	0.188	0.843	10th	0.086	3rd	0.098	4th
KNN (K=1)	0.908	0.167	0.839	11th	0.076	4th	0.075	5th
SVM	0.966	0.156	0.89	8th	0.075	5th	0.122	3rd
KNN (K=3)	0.975	0.094	0.893	7th	0.046	6th	0.069	6th
C4.5	0.985	0.083	0.901	5th	0.041	7th	0.068	7th
KNN (K=5)	0.985	0.042	0.897	6th	0.021	8th	0.027	9th
KNN (K=9)	0.999	0.031	0.909	1st	0.015	9th	0.030	8th
KNN (K=7)	0.996	0.031	0.906	3rd	0.015	9th	0.027	9th
RF	0.998	0.021	0.907	2nd	0.010	11th	0.019	11th
LMT	0.996	0.010	0.904	4th	0.005	12th	0.006	12th

Testing with VD (n=1029)

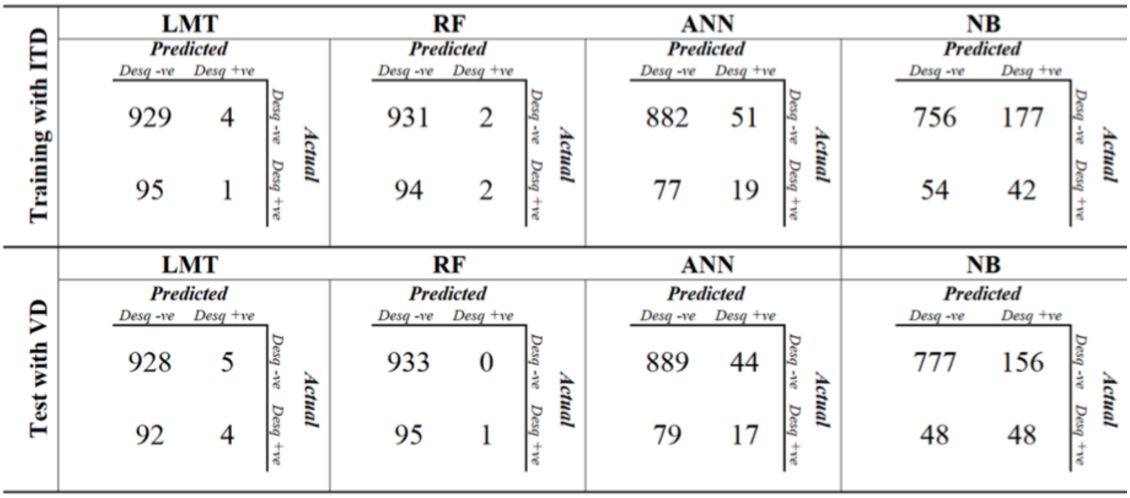
Algorithm	Specificity (TNR)	Sensitivity (TPR)	Accuracy		Balanced Accuracy		Youden's Index	
			Score	Rank	Score	Rank	Score	Rank
NB	0.833	0.500	0.802	9th	0.208	1st	0.333	1st
ANN	0.953	0.177	0.880	7th	0.084	3rd	0.130	3rd
LR	0.959	0.135	0.882	6th	0.065	5th	0.094	6th
KNN (K=1)	0.923	0.292	0.864	8th	0.135	2nd	0.215	2nd
SVM	0.976	0.146	0.899	5th	0.071	4th	0.122	4th
KNN (K=3)	0.979	0.125	0.899	5th	0.061	6th	0.104	5th
C4.5	0.979	0.125	0.899	5th	0.061	6th	0.104	5th
KNN (K=5)	0.989	0.063	0.903	4th	0.031	7th	0.052	7th
KNN (K=9)	0.999	0.042	0.910	1st	0.021	9th	0.041	9th
KNN (K=7)	0.998	0.052	0.910	1st	0.026	8th	0.050	8th
RF	1.000	0.010	0.908	2nd	0.005	10th	0.010	11th
LMT	0.995	0.042	0.906	3rd	0.021	9th	0.037	10th

The RT potential benefits must be weighed against the possibility of causing damage to healthy tissue. The final aim is to maximise curative response while minimising the probability of complications [8]. Hence, the RTML domain experts noted that favouring the minority group over the majority class could prevent patients from benefiting from the treatment and being shifted to other alternatives. However, caring about both groups equally may lead to increased false negatives (FN) as more patients are likely to develop acute desquamation due to undergoing radiotherapy, which in turn compromises patients' QoL and runs the risk of local cancer recurrence in the event of RT interruption.

Experts confirmed that the sensitivity achieved was insufficient for all ITD models in training and test without mitigating the class imbalance problem, ranging from 0.01 to 0.44 in training and from 0.04 to 0.5 in test for LMT and NB, respectively. Hence all ITD models are considered not effective at predicting acute desquamation.

Therefore, we decided to examine both scenarios by seeking an improvement with CS classification for the top two performing classifiers, NB and ANN, and the bottom two, RF and LMT, in terms of their balanced accuracy and Youden's index (γ) scores in training. Finally, the TPR-TNR trade-off evaluation occurs when comparing all tested models having applied both strategies, CS classification and resampling, to mitigate the imbalanced learning issue. The confusion matrices for the four selected models in Table 6.4 describe the numeric count of correctly classified patients, FP (type I) and FN (type II) errors misclassifications.

Table 6.4. Training and test confusion matrices of the imbalanced models



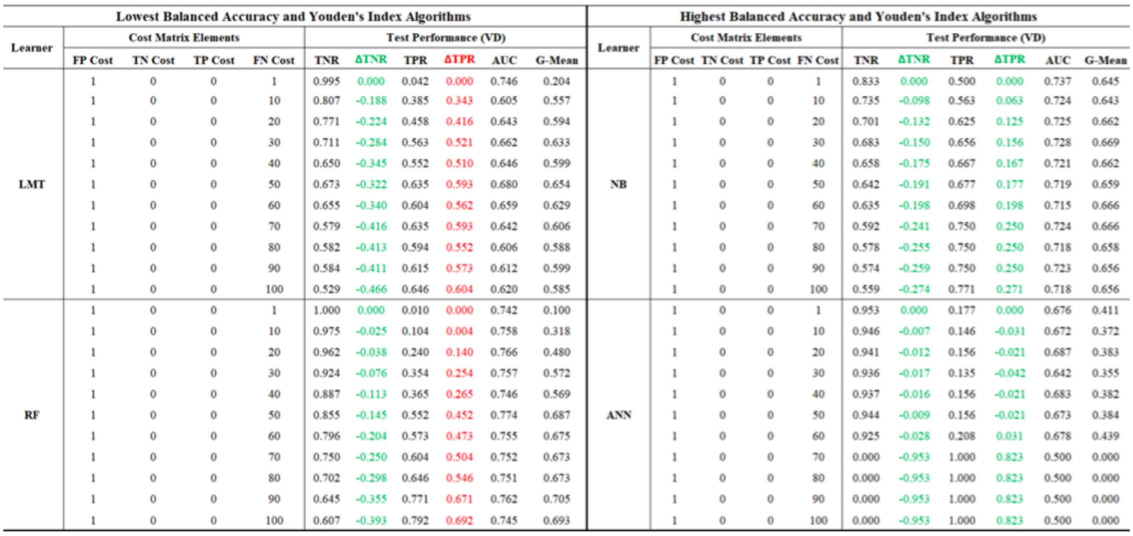
	LMT			RF			ANN			NB		
	Predicted			Predicted			Predicted			Predicted		
	Desq -ve	Desq +ve	Actual	Desq -ve	Desq +ve	Actual	Desq -ve	Desq +ve	Actual	Desq -ve	Desq +ve	Actual
Training with ITD	929	4	Desq -ve	931	2	Desq -ve	882	51	Desq -ve	756	177	Desq -ve
	95	1	Desq +ve	94	2	Desq +ve	77	19	Desq +ve	54	42	Desq +ve
	LMT			RF			ANN			NB		
	Predicted			Predicted			Predicted			Predicted		
	Desq -ve	Desq +ve	Actual	Desq -ve	Desq +ve	Actual	Desq -ve	Desq +ve	Actual	Desq -ve	Desq +ve	Actual
Test with VD	928	5	Desq -ve	933	0	Desq -ve	889	44	Desq -ve	777	156	Desq -ve
	92	4	Desq +ve	95	1	Desq +ve	79	17	Desq +ve	48	48	Desq +ve

6.10.2 Cost-sensitive modelling results

The expected improvement to the four selected algorithms NB, ANN, RF and LMT with CS classification is achieved with an incremental inverse-class distribution cost matrix to penalise the classifier for the misclassification of FN records.

The incremental penalty is expected to skew the correct classification toward the positive group, as no further improvements are required for the negative class. Forty models were built with ITD accompanied by a defined cost matrix and then tested with VD. To evaluate such an improvement, four metrics measurements in test were reported: the AUC-ROC, G-mean, TPR and TNR (See Table 6.5). It is sufficient to report and analyse only the test results with VD than the training with ITD for all CS models to allow for a fair comparison later with other models built with resampling techniques. Resample models used training datasets of different sizes (samples of ITD).

Table 6.5. Cost-Sensitive test performance on the selected ITD algorithms



Lowest Balanced Accuracy and Youden's Index Algorithms											Highest Balanced Accuracy and Youden's Index Algorithms										
Learner	Cost Matrix Elements				Test Performance (VD)						Learner	Cost Matrix Elements				Test Performance (VD)					
	FP Cost	TN Cost	TP Cost	FN Cost	TNR	ΔTNR	TPR	ΔTPR	AUC	G-Mean		FP Cost	TN Cost	TP Cost	FN Cost	TNR	ΔTNR	TPR	ΔTPR	AUC	G-Mean
LMT	1	0	0	1	0.995	0.000	0.042	0.000	0.746	0.204	NB	1	0	0	1	0.833	0.000	0.500	0.000	0.737	0.645
	1	0	0	10	0.807	-0.188	0.385	0.343	0.605	0.557		1	0	0	10	0.735	-0.098	0.563	0.063	0.724	0.643
	1	0	0	20	0.771	-0.224	0.458	0.416	0.643	0.594		1	0	0	20	0.701	-0.132	0.625	0.125	0.725	0.662
	1	0	0	30	0.711	-0.284	0.563	0.521	0.662	0.633		1	0	0	30	0.683	-0.150	0.656	0.156	0.728	0.669
	1	0	0	40	0.650	-0.345	0.552	0.510	0.646	0.599		1	0	0	40	0.658	-0.175	0.667	0.167	0.721	0.662
	1	0	0	50	0.673	-0.322	0.635	0.593	0.680	0.654		1	0	0	50	0.642	-0.191	0.677	0.177	0.719	0.659
	1	0	0	60	0.655	-0.340	0.604	0.562	0.659	0.629		1	0	0	60	0.635	-0.198	0.698	0.198	0.715	0.666
	1	0	0	70	0.579	-0.416	0.635	0.593	0.642	0.606		1	0	0	70	0.592	-0.241	0.750	0.250	0.724	0.666
	1	0	0	80	0.582	-0.413	0.594	0.552	0.606	0.588		1	0	0	80	0.578	-0.255	0.750	0.250	0.718	0.658
	1	0	0	90	0.584	-0.411	0.615	0.573	0.612	0.599		1	0	0	90	0.574	-0.259	0.750	0.250	0.723	0.656
	1	0	0	100	0.529	-0.466	0.646	0.604	0.620	0.585		1	0	0	100	0.559	-0.274	0.771	0.271	0.718	0.656
RF	1	0	0	1	1.000	0.000	0.010	0.000	0.742	0.100	ANN	1	0	0	1	0.953	0.000	0.177	0.000	0.676	0.411
	1	0	0	10	0.975	-0.025	0.104	0.004	0.758	0.318		1	0	0	10	0.946	-0.007	0.146	-0.031	0.672	0.372
	1	0	0	20	0.962	-0.038	0.240	0.140	0.766	0.480		1	0	0	20	0.941	-0.012	0.156	-0.021	0.687	0.383
	1	0	0	30	0.924	-0.076	0.354	0.254	0.757	0.572		1	0	0	30	0.936	-0.017	0.135	-0.042	0.642	0.355
	1	0	0	40	0.887	-0.113	0.365	0.265	0.746	0.569		1	0	0	40	0.937	-0.016	0.156	-0.021	0.683	0.382
	1	0	0	50	0.855	-0.145	0.552	0.452	0.774	0.687		1	0	0	50	0.944	-0.009	0.156	-0.021	0.673	0.384
	1	0	0	60	0.796	-0.204	0.573	0.473	0.755	0.675		1	0	0	60	0.925	-0.028	0.208	0.031	0.678	0.439
	1	0	0	70	0.750	-0.250	0.604	0.504	0.752	0.673		1	0	0	70	0.000	-0.953	1.000	0.823	0.500	0.000
	1	0	0	80	0.702	-0.298	0.646	0.546	0.751	0.673		1	0	0	80	0.000	-0.953	1.000	0.823	0.500	0.000
	1	0	0	90	0.645	-0.355	0.771	0.671	0.762	0.705		1	0	0	90	0.000	-0.953	1.000	0.823	0.500	0.000
	1	0	0	100	0.607	-0.393	0.792	0.692	0.745	0.693		1	0	0	100	0.000	-0.953	1.000	0.823	0.500	0.000

From Table 6.5, the CS classification showed a consistent deterioration of TNR for all models across all four algorithms. ANN was impacted by the highest level of TNR deterioration (ΔTNR = – 0.953) when compared to its original ITD test result at an FN penalty of 10, CS-ANN TNR deterioration was preceded by the LMT model at an FN cost of 100 with a ΔTNR of – 0.466.

For NB, the maximum TNR loss was – 0.274, and for RF was – 0.393.

A consistent TPR improvement is also observed when examining the TPR for LMT, RF and NB CS models. Initially, ANN showed a slight loss until an FN cost of 60, where gains started to show. Then, ANN achieved the most TPR improvement with ΔTPR = 0.823 at FN penalties of 70, 80, 90 and 100. ANN's massive improvement resulted in a total misclassification of all the desquamation-negative patients (majority class). The TPR gains as a result of CS classifications were in the range of 0.343 to 0.604 in LMT models, 0.004 to 0.692 in RF models, and NB models showed gains between 0.063 and 0.271. The impact of incremental FN penalty in the cost matrix on each of the four selected classifiers can be observed in Figure 6.6. The shift in classifier attention is quantified by computing the absolute change in TNR and TPR for each model after applying a specific FN penalty.

Figure 6.6 – A shows that the change rate for TPR was greater in LMT and RF models at every FN penalty. However, NB models showed almost a similar rate of change in classifier TNR and TPR. ANN maintained a similar behaviour to NB for the initial six steps of incremental FN cost, and then a constant massive change rate for both TNR and TPR occurred.

Figure 6.6 – B shows that the shift of CS classification with incremental FN costs is linear on both TNR and TPR. However, the impact varied among different classifiers for the same FN penalty values.

For FN penalties (Figure 6.6 – B) from 10 to 60, the change in TPR performance for LMT was 0.521 for FN cost of 60, followed by RF with a change rate of 0.473 and 0.198 in the NB case. Nevertheless, the ANN classifier ranked last, showing strong resistance to budge with FN penalties; its change fluctuated lightly between 0.021 and 0.042.

For the same range of FN penalties, 10 to 60, the TNR change also showed a direct linear rise. LMT had a steep |ΔTNR| elevating higher than all other classifiers ranging from 0.0340 to 0.188. NB presented a less elevated absolute TNR change and very close to the absolute change in its TPR with FN penalties {10,20,30, 40,50,60}, with RF not far behind at the FN penalty of 60. ANN maintained its resistance to change, with FN penalties showing a slight change compared to its ITD model with an FN cost of 1.

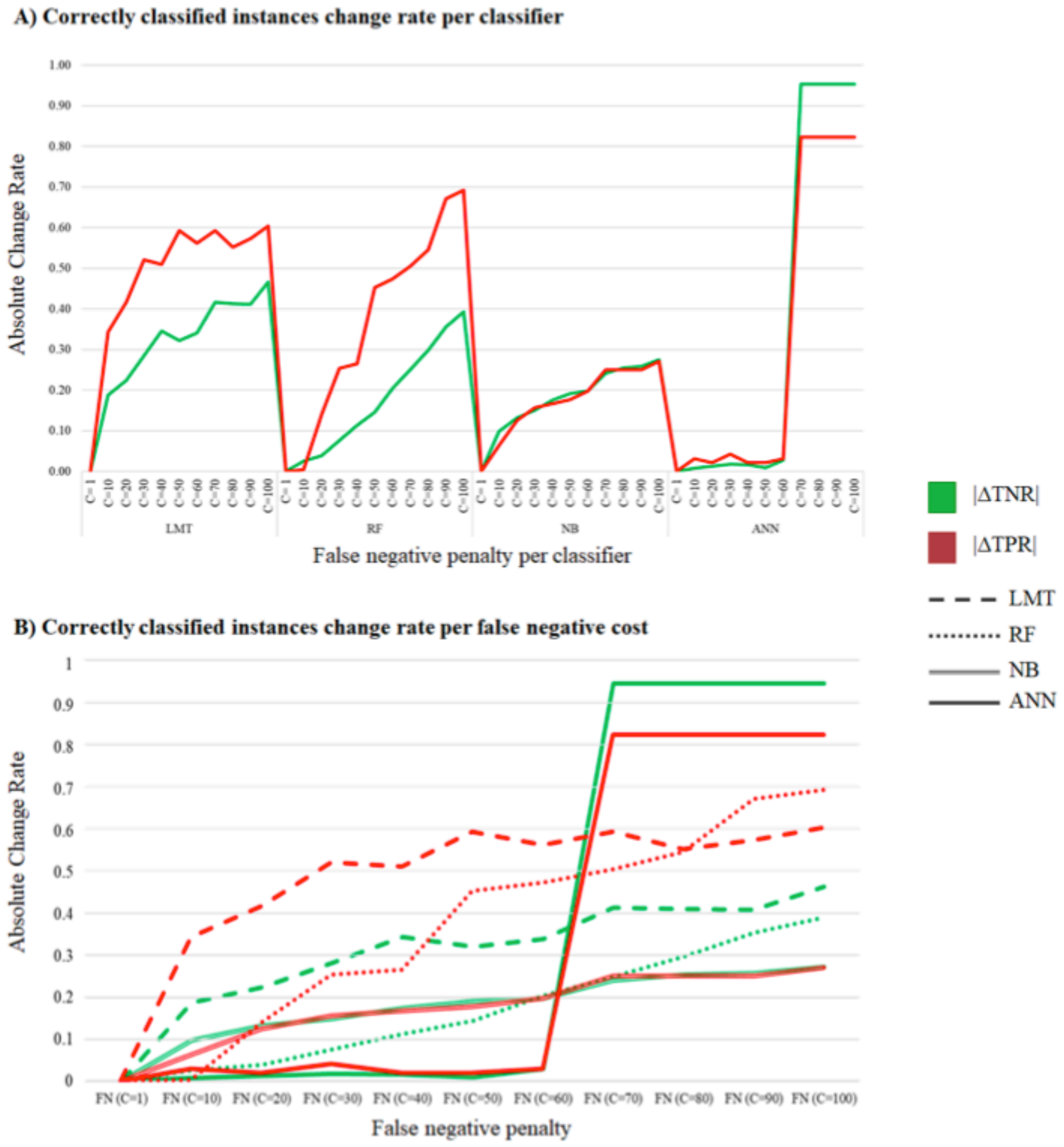


Fig. 6.6 Absolute change in TNR and TPR test performance per FN penalty in LMT, RF, NB and ANN models

For FN penalties from 70 to 100 (Figure 6.6 – B), a sudden step-change in ANN classifier TNR and TPR occurred with |ΔTNR| = 0.953 and |ΔTPR|

= 0.823 across all FN penalties {70, 80, 90, 100}. This sharp constant rise in ANN's |ΔTNR| and |ΔTPR| compared to its TNR and TPR at each penalty indicates a catastrophic impact of completely overfitting the negative class with a TPR of 1.000 and fully underfitting the majority group with a TNR of 0.000. At the FN penalty of 80, the absolute TPR change in the RF classifier overtook its opponent in the prior LMT models, and both maintained a larger change above NB but below ANN models.

The average CS impact on the absolute change in TPR and TNR for both NB and ANN models was very close at all FN penalties. The average |ΔTPR| was 0.191 and 0.346 compared to the |ΔTNR| average of 0.197 and 0.390, respectively. LMT and RF average |ΔTPR| was 0.527 and 0.400 compared to an average |ΔTNR| of 0.341 and 0.190, respectively.

Figure 6.7 shows the AUC, G-Mean, TPR and TNR performance combinations for LMT, RF, NB and ANN for all incremental FN penalties. Figures 6.7 – A, 6.7 – B, 6.7 – C and 6.7 – D demonstrate the AUC-ROC vulnerability to the class imbalance problem by achieving a reasonably good score > 0.70 despite the models' poor power of discrimination towards the minority positive class [380], in the case of LMT, RF and ANN at an FN cost equal to an FP of 1 in the ITD models. When applying incremental penalties to FN misclassifications, the AUC-ROC performance continues to retain its score for all CS models within a margin of 8% in the case of LMT, 3% for RF and 2% for NB. ANN initially tries to retain its AUC performance within a margin of 3% until its sudden drop to its minimum of 0.50 for all FN costs above 60, at which the ANN classifier loses its ability to classify all patients in the majority group.

The G-Mean score is proven to be more robust than the AUC – ROC when assessing the ability of classifiers to avoid overfitting and underfitting the classes; the greater the G-Mean, the better. When examining the G-Mean evaluations for LMT, RF and ANN ITD Models (FN penalty = 1), the

G-Mean evaluations were small, 0.204, 0.100 and 0.411, respectively, indicating poor classification performance.

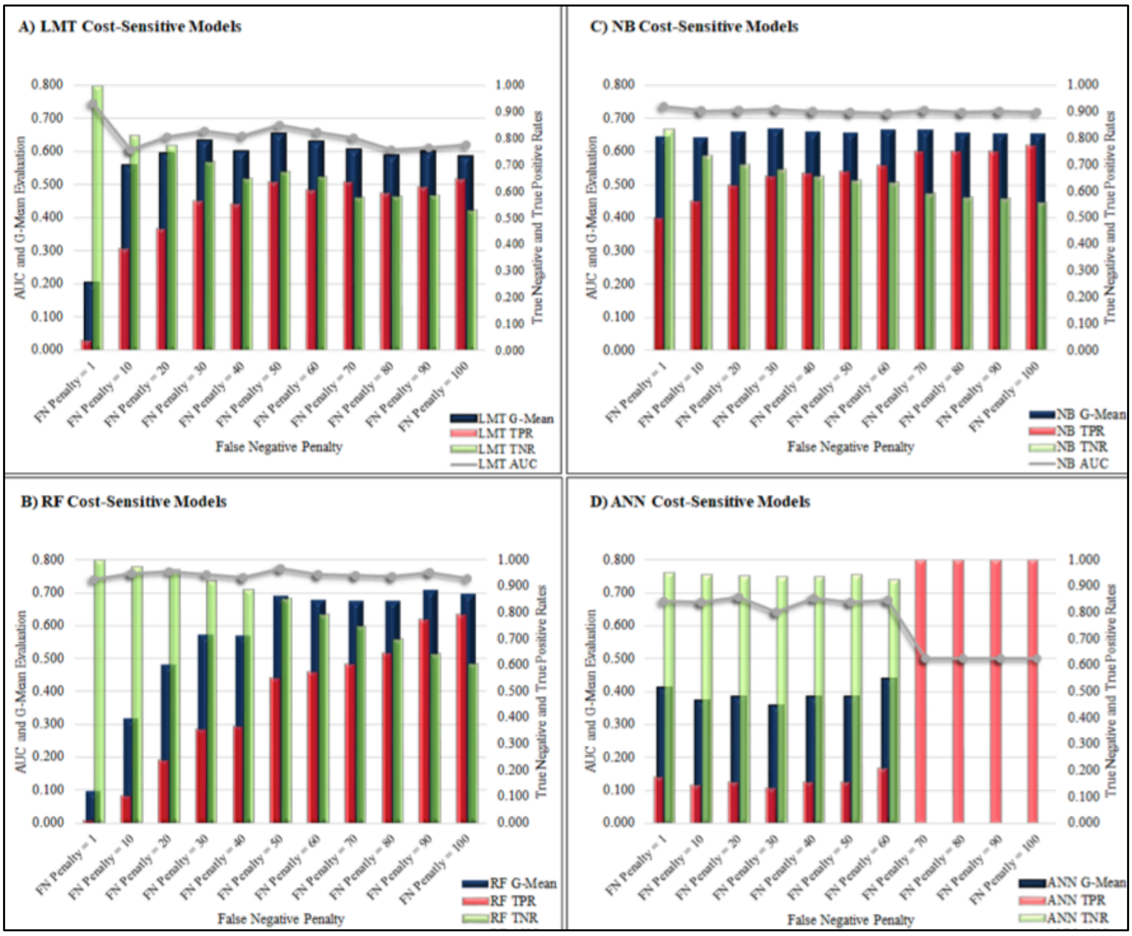


Fig. 6.7 TNR, TPR, G-Mean and AUC computations per FN penalty in LMT, RF, NB and ANN models

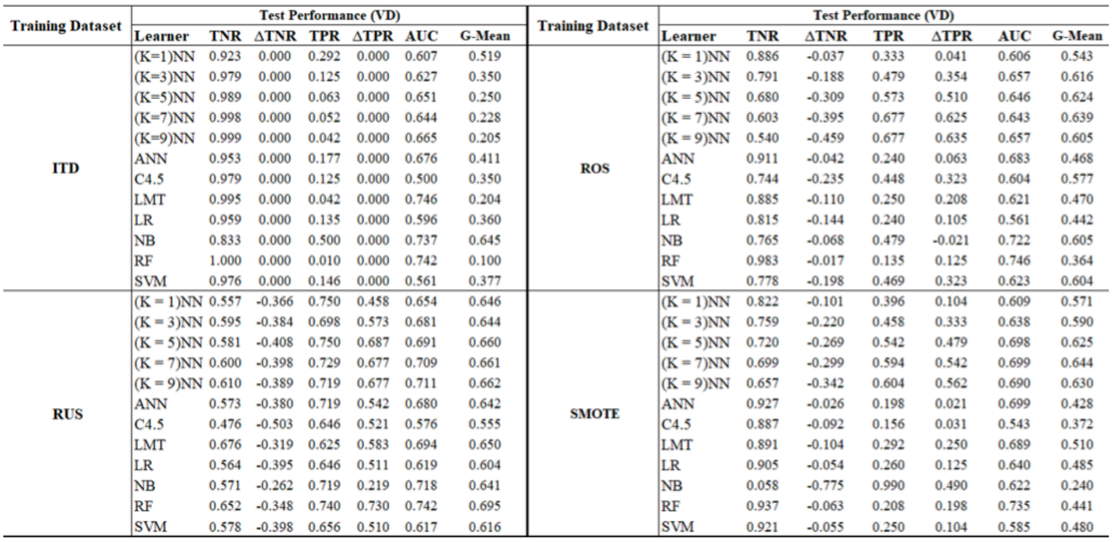
In Figure 6.7 – D, the ANN's G-Mean values dropped to zero when the model completely overfitted the positive class and misclassified all the negative class labels. NB (Figure 6.7 – C) presented a greater G-Mean for its ITD model of 0.645, indicating better discrimination between both classes at an FN cost of 1. NB maintained a consistent G-Mean with a minimal change margin of 2% across all CS models ranging between 0.663 and 0.669.

Examining the G-Mean for all CS models in Figure 6.7 and Table 6.5 shows that CS-RF and CS-NB models reserved the top ten ranks in the G-Mean evaluation. The top five places were for RF CS models at FN costs {50, 60, 70, 80, 90, 100}, the bottom five ranks were occupied by NB CS models at FN costs {20, 30, 40, 60, 70}.

6.10.3 Resampled data modelling results

Table 6.6 shows the TNR, TPR, TN change rate (ΔTNR), TP change rate (ΔTPR), G-Mean and AUC test performances of resampling techniques RUS, ROS and SMOTE for RF, LMT, NB, C4.5, ANN, KNN, SVM and LR classifiers. Furthermore, by analysing the effect of resampling techniques on both TNR and TPR in Figure 6.8, it is clear that the resampling techniques improved the TPR across all classifiers while the TNR deteriorated across all classifiers for all resampling techniques from the original ITD-based state.

Table 6.6. Models test performances with data resampling strategy



Training Dataset	Test Performance (VD)							Training Dataset	Test Performance (VD)						
	Learner	TNR	ΔTNR	TPR	ΔTPR	AUC	G-Mean		Learner	TNR	ΔTNR	TPR	ΔTPR	AUC	G-Mean
ITD	(K=1)NN	0.923	0.000	0.292	0.000	0.607	0.519	ROS	(K = 1)NN	0.886	-0.037	0.333	0.041	0.606	0.543
	(K=3)NN	0.979	0.000	0.125	0.000	0.627	0.350		(K = 3)NN	0.791	-0.188	0.479	0.354	0.657	0.616
	(K=5)NN	0.989	0.000	0.063	0.000	0.651	0.250		(K = 5)NN	0.680	-0.309	0.573	0.510	0.646	0.624
	(K=7)NN	0.998	0.000	0.052	0.000	0.644	0.228		(K = 7)NN	0.603	-0.395	0.677	0.625	0.643	0.639
	(K=9)NN	0.999	0.000	0.042	0.000	0.665	0.205		(K = 9)NN	0.540	-0.459	0.677	0.635	0.657	0.605
	ANN	0.953	0.000	0.177	0.000	0.676	0.411		ANN	0.911	-0.042	0.240	0.063	0.683	0.468
	C4.5	0.979	0.000	0.125	0.000	0.500	0.350		C4.5	0.744	-0.235	0.448	0.323	0.604	0.577
	LMT	0.995	0.000	0.042	0.000	0.746	0.204		LMT	0.885	-0.110	0.250	0.208	0.621	0.470
	LR	0.959	0.000	0.135	0.000	0.596	0.360		LR	0.815	-0.144	0.240	0.105	0.561	0.442
	NB	0.833	0.000	0.500	0.000	0.737	0.645		NB	0.765	-0.068	0.479	-0.021	0.722	0.605
	RF	1.000	0.000	0.010	0.000	0.742	0.100		RF	0.983	-0.017	0.135	0.125	0.746	0.364
	SVM	0.976	0.000	0.146	0.000	0.561	0.377		SVM	0.778	-0.198	0.469	0.323	0.623	0.604
RUS	(K = 1)NN	0.557	-0.366	0.750	0.458	0.654	0.646	SMOTE	(K = 1)NN	0.822	-0.101	0.396	0.104	0.609	0.571
	(K = 3)NN	0.595	-0.384	0.698	0.573	0.681	0.644		(K = 3)NN	0.759	-0.220	0.458	0.333	0.638	0.590
	(K = 5)NN	0.581	-0.408	0.750	0.687	0.691	0.660		(K = 5)NN	0.720	-0.269	0.542	0.479	0.698	0.625
	(K = 7)NN	0.600	-0.398	0.729	0.677	0.709	0.661		(K = 7)NN	0.699	-0.299	0.594	0.542	0.699	0.644
	(K = 9)NN	0.610	-0.389	0.719	0.677	0.711	0.662		(K = 9)NN	0.657	-0.342	0.604	0.562	0.690	0.630
	ANN	0.573	-0.380	0.719	0.542	0.680	0.642		ANN	0.927	-0.026	0.198	0.021	0.699	0.428
	C4.5	0.476	-0.503	0.646	0.521	0.576	0.555		C4.5	0.887	-0.092	0.156	0.031	0.543	0.372
	LMT	0.676	-0.319	0.625	0.583	0.694	0.650		LMT	0.891	-0.104	0.292	0.250	0.689	0.510
	LR	0.564	-0.395	0.646	0.511	0.619	0.604		LR	0.905	-0.054	0.260	0.125	0.640	0.485
	NB	0.571	-0.262	0.719	0.219	0.718	0.641		NB	0.058	-0.775	0.990	0.490	0.622	0.240
	RF	0.652	-0.348	0.740	0.730	0.742	0.695		RF	0.937	-0.063	0.208	0.198	0.735	0.441
	SVM	0.578	-0.398	0.656	0.510	0.617	0.616		SVM	0.921	-0.055	0.250	0.104	0.585	0.480

Figure 6.9 shows the depth of impact (absolute change in TPR and TNR) of the resampling techniques. In RUS-based models, the TPR change was greater than TNR across almost all classifiers except for NB. The largest (TPR, TNR) change is observed in the RF model (0.730, 0.348). In the ROS-based models, the impact of resampling was greater on TNR for all models but LR and NB models; however, the depth of effect (TPR, TNR) is small (0.105, 0.144) and (0.021, 0.068), respectively. SMOTE-based models also show that TPR was impacted higher than TNR except in ANN, C4.5 and NB.

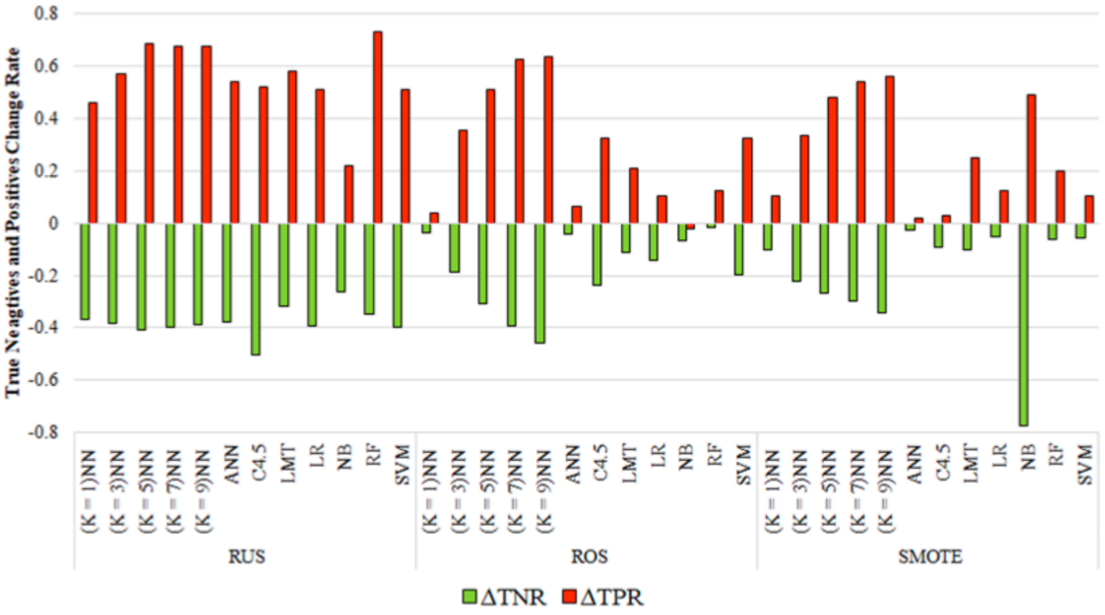


Fig. 6.8 TN and TP change rates in test per classifier in RUS, ROS and SMOTE models

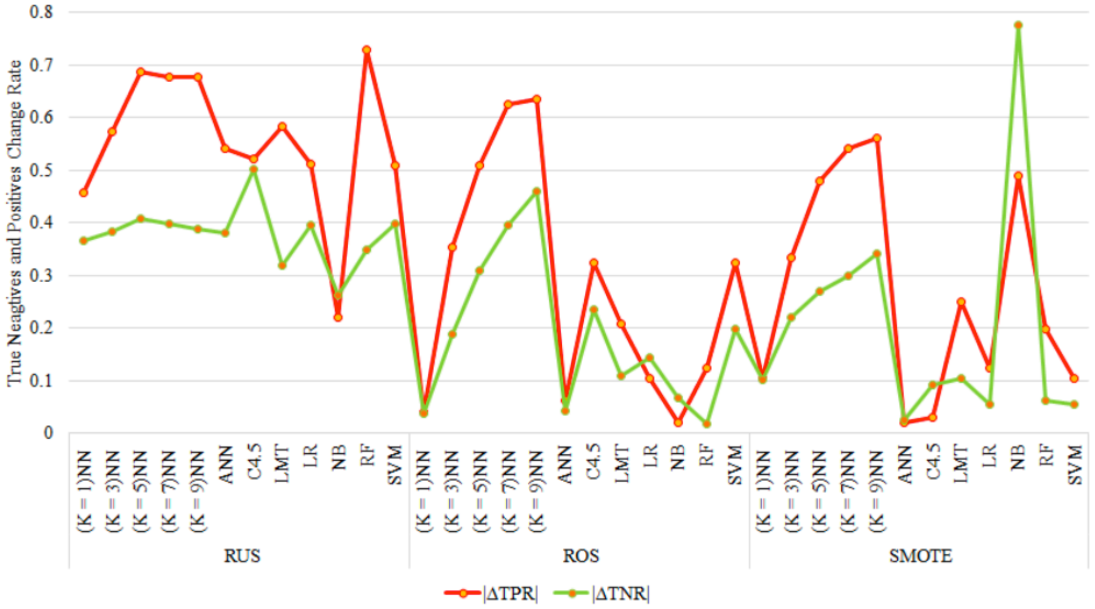


Fig. 6.9 Absolute TN and TP change rates in test per classifier in resampled data models

Figure 6.10 shows the evaluation of the G-Mean and AUC-ROC in relation to the balance between TPR and TNR. In RUS-based models, it is observed that the TPR is overtaking the TNR in all models. Larger G-Mean values indicate that the classifier is not overfitting or underfitting any of the classes. The evaluation of the G-Mean and AUC-ROC are harmonised across all RUS-based models (Figure 6.10 – A). The lowest G-Mean and AUC-ROC measurements are observed for the C4.5 model at 0.555 and 0.576, respectively. The highest G-mean evaluation was 0.695, achieved by the RF model, with the highest AUC of 0.742.

Unlike the RUS-based models, the ROS models (Figure 6.10 – B) experienced a frequent disagreement between the AUC-ROC and the G-mean scores. While the G-mean score was small, indicating there is a large bias of accuracy towards one of the classes in the case of (K=1)NN, ANN, LMT, LR, NB and RF, the AUC-ROC seems to have shown a deceiving high evaluation for such models, for instance, 0.746 for RF and 0.722 for NB.

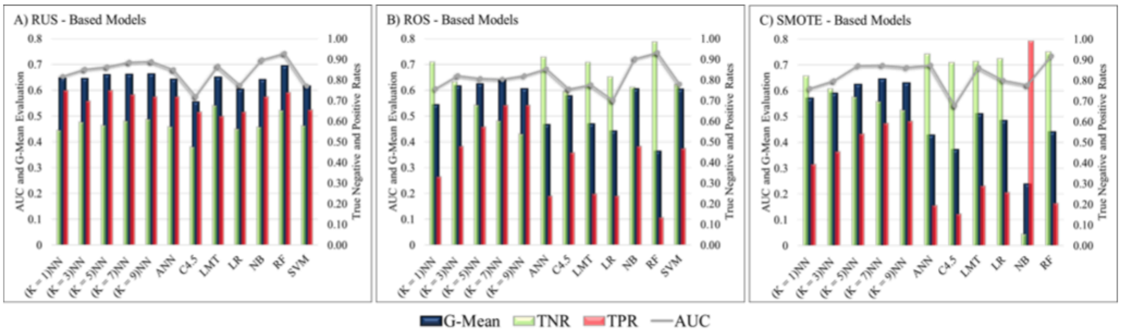


Fig. 6.10 TNR, TPR, G-Mean and AUC test performances for the resampled data models

In SMOTE-based (Figure 6.10 – C) models, the AUC-ROC again can show misleading high evaluations for models with inflated class accuracy in either class, specifically in the cases of ANN, LMT, LR, NB and RF. For example, RF achieved a good AUC-ROC score of 0.746 with a poor TPR of 0.198 and an excessive TNR of 0.937. However, examining the G-Mean for all SMOTE models cuts through the deception of the inflated AUC-ROC scores; therefore, the RF G-Mean score is 0.441, which is relatively low. A similar case is observed in the ANN SMOTE-based model; the AUC-ROC is 0.699, while the G-Mean is 0.428.

6.11 Attributes information gain analysis

The information Gain (IG) of each variable was also computed. The IG is the expected reduction of entropy when partitioning the data for a given variable. Entropy is related to how likely we are to predict the class labels of samples, i.e., when data has high entropy, it is difficult to predict the class label of an example, and when the entropy is low, the opposite is verified. So, IG provides a measure of how much the prediction of the class labels of samples would improve if the dataset was split using just one feature. IG was used to monitor any bias that occurs in either training or test datasets. Entropy and purity could vary due to data pre-processing techniques such as imputation and resampling with different numbers of records. The more plausible the conclusive pattern of IG among datasets, the less bias is introduced in modelling. By looking at both ITD and VD datasets in Figure 6.11, it is notable that most of their features preserved close purity and entropy levels before and after imputation.

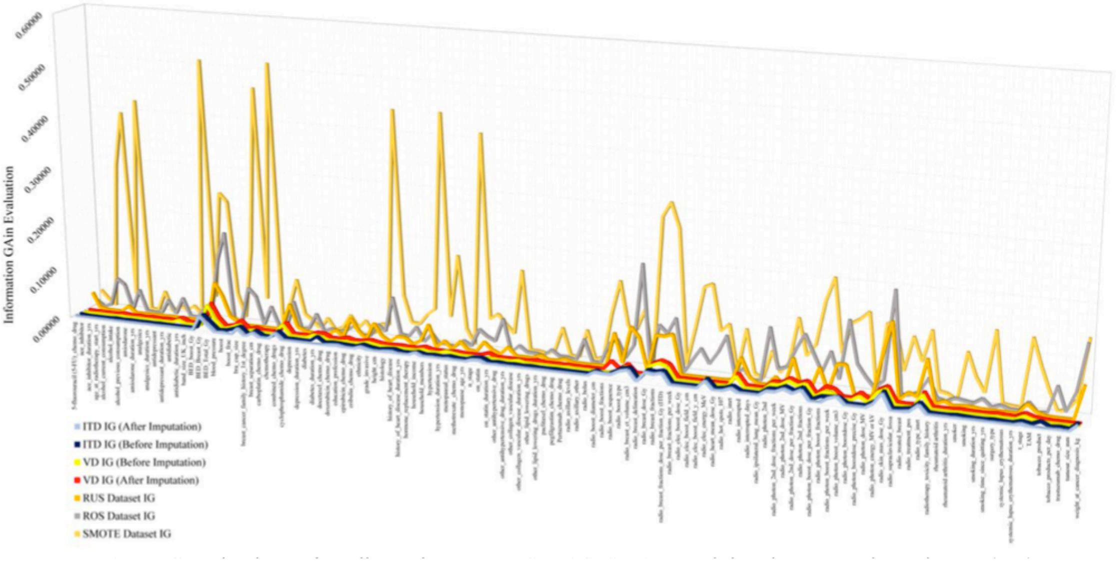


Fig. 6.11 IG evaluations of ITD, RUS, ROS, SMOTE training and test datasets (VD) features

Features that showed dominance in IG evaluation before the DMI imputation have also maintained power after the DMI imputation. Note that the imputation of ITD and VD separately removes the opportunity for both datasets to share the same statistical parameter setting used by the imputation algorithm. This execution makes the training and test datasets independent from each other.

6.12 Models' evaluation and selection

In clinical trials, the AUC-ROC metric is known to preserve the discriminant validity in treatment comparisons in balanced data [381] or where a suitable compensating method is applied to overcome the class imbalance. Hence clinicians rely on such a measure as a critical evaluator in judging the performance of a prediction model. However, in our previous results, we demonstrated that a model's AUC-ROC score, in some instances, could be deceiving. Therefore, additional metrics such as the G-Mean was nominated to reveal such cases and provide a less biased assessment. In other cases, choosing a single model as a Hero model becomes challenging when different models are deemed suitable. Hence, domain experts should set an additional success criterion to define an acceptable level of TPR-TNR trade-off.

To select the best-performing models out of 76 models built with either of the used imbalanced learning in this paper to predict the occurrence of acute desquamation, clinicians called for filtering out all models with an AUC-ROC score below 0.700. Models with good discriminatory powers between the positive and the negative groups must have a minimum AUC-ROC of 0.700. Twenty-seven prediction models with AUC-ROC ≥ 0.700 remained (Figure 6.12). Seven models with a high AUC-ROC associated with a low G-Mean indicated overfitting the majority negative class, and underfitting the minority positive class were also dropped out.

Based on all models' validation TPR and TNR evaluations and the clinicians' trade-off between TPR and TNR in Figure 6.13, RTML experts agreed on two trade-off conditions that all models compete towards, based on lower and upper threshold values of 0.630 and 0.700, respectively. These conditions are (TPR ≥ 0.630 & TNR ≥ 0.700) and (TNR ≥ 0.630 & TPR ≥ 0.700). Three models met both conditions. They are CS-RF(FN:FP=90:1, TNR=0.645, TPR=0.771, AUC=0.762, G-Mean=0.705), RUS-RF(TNR=0.652, TPR=0.740, AUC=0.742, G-Mean=0.695), CS-RF(FN:FP=80:1, and TNR=0.702, TPR=0.646, AUC=0.751, G-Mean=0.673).

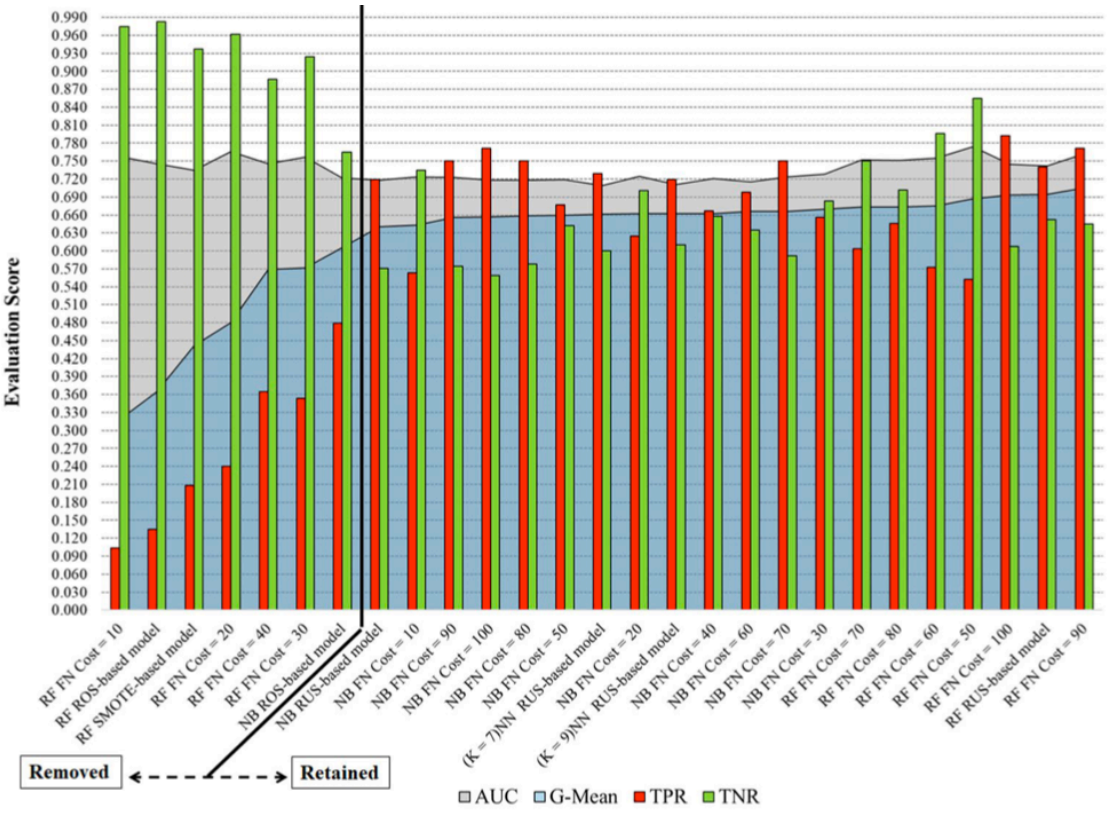


Fig. 6.12 Filtered models and their associated TNR, TPR, G-Mean and AUC-ROC test performances

The confusion matrices for the compliant three tested models are found in Table 6.7. Maximising TPs is essential; therefore, specialists' consensus concluded that the best-performing model (Hero Model) was CS-RF(FN:FP = 90:1) for exceeding all other models' sensitivity, AUC and G-Mean performances while maintaining a competitive specificity. The calculated balanced accuracy and Youden's index for the hero model were 0.249 and 0.416, respectively. It is found that these values were also the highest among all models in this case study.

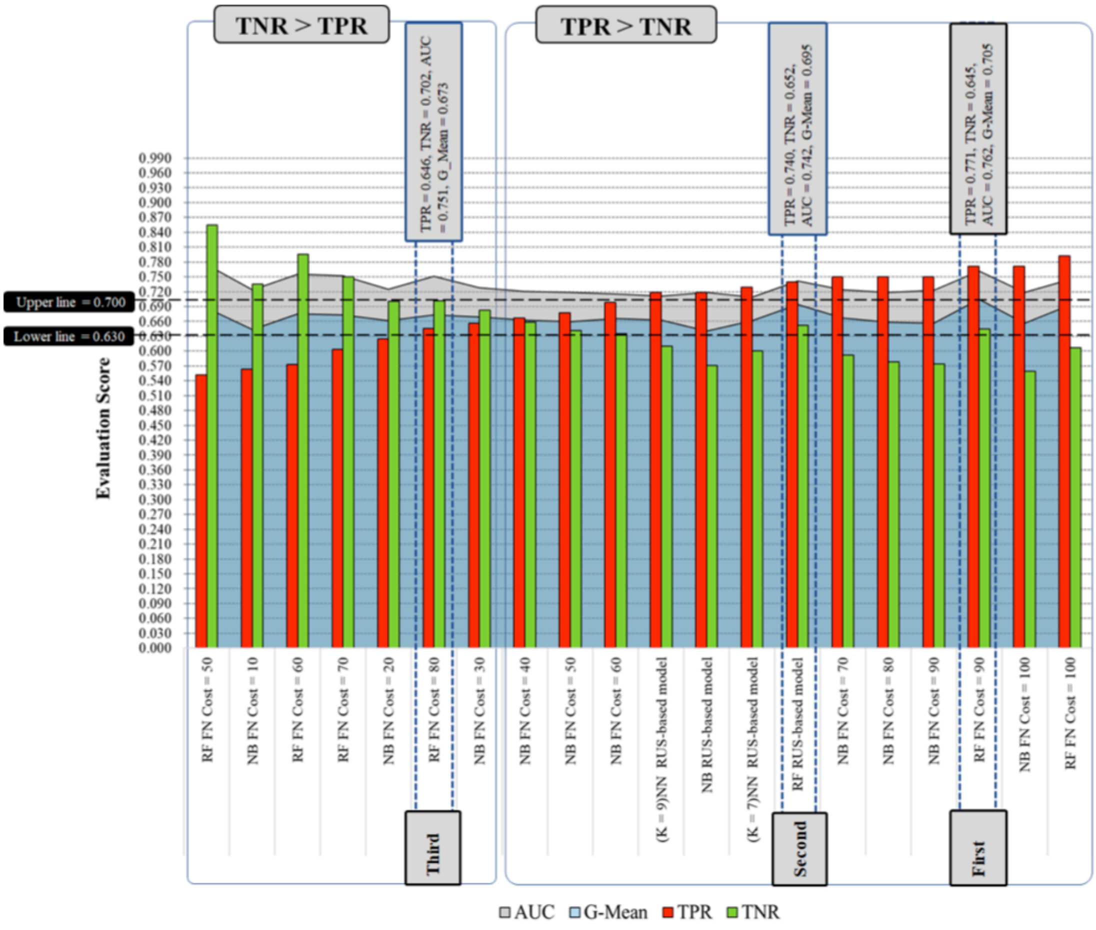
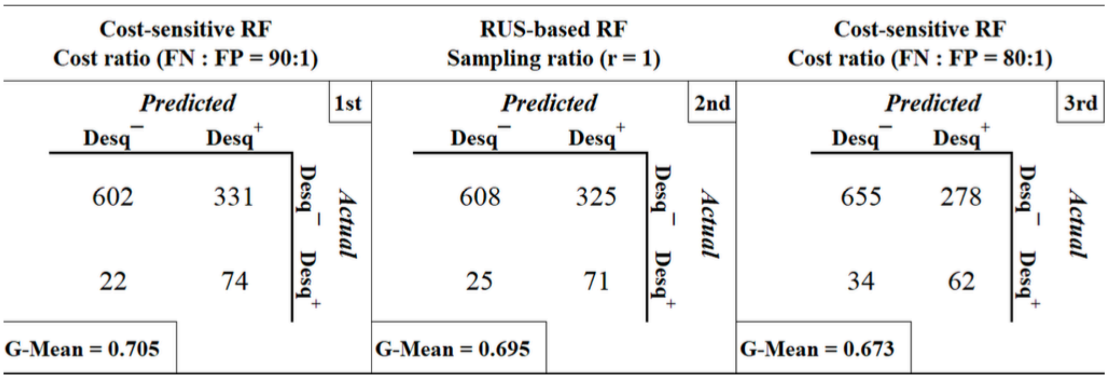


Fig. 6.13 True Positive Rate (TPR) and True Negative Rate (TNR) trade-off threshold lines for all tested models with VD. FN prediction costs refer to Penalty values in the explicit cost-sensitive models. While FP predictions costs are kept at a value of 1, both TP and TN prediction costs always remain at the value of zero

In OCTOPUS Framework, from our observations, clinicians tend to sacrifice deep interpretability for a reliable model that passes their evaluation. Therefore, following the selection of the hero model, under the clinicians' request, we removed unimportant features (features that the model did not use), seeking further performance improvement compared to modelling all available features, which partially may simplify the model. Such a request is not in line with recommendations in the OCTOPUS framework. In our literature reviews, we presented the key debates for including all applicable available features and the benefits of allowing better monitoring of future changes within the variables, not just the instances if a model is deployed in production.

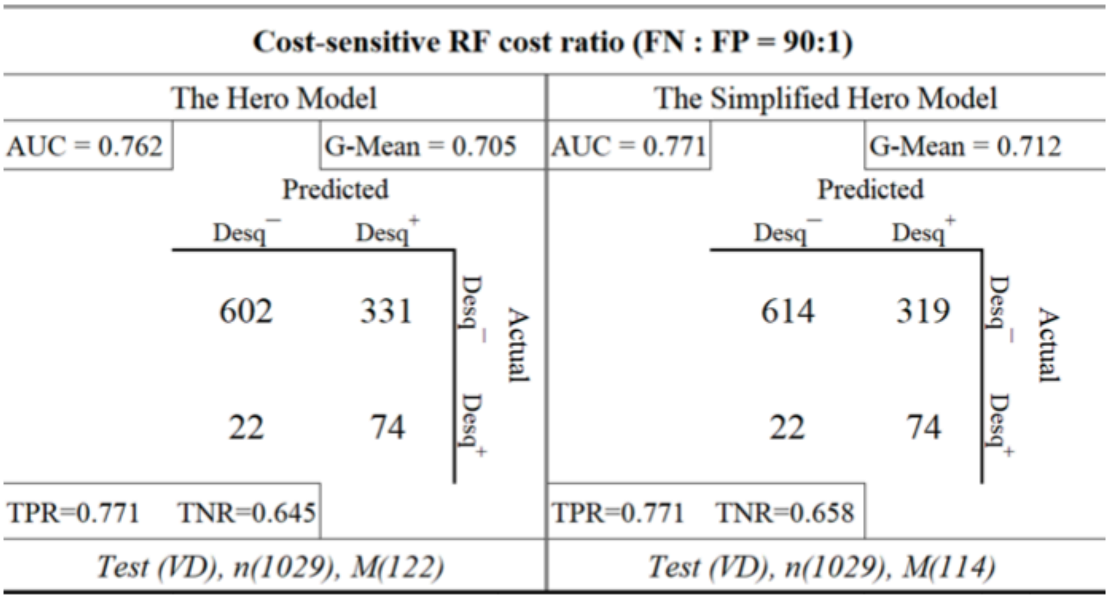
Table 6.7. Performance ranking of the compliant three models on VD



	Cost-sensitive RF Cost ratio (FN : FP = 90:1)		RUS-based RF Sampling ratio (r = 1)		Cost-sensitive RF Cost ratio (FN : FP = 80:1)	
		1st		2nd		3rd
	Predicted		*Predicted*		*Predicted*	
Actual	Desq$^-$	Desq$^+$	Desq$^-$	Desq$^+$	Desq$^-$	Desq$^+$
Desq$^-$	602	331	608	325	655	278
Desq$^+$	22	74	25	71	34	62
	G-Mean = 0.705		G-Mean = 0.695		G-Mean = 0.673	

Upon domain experts' request, Feature importance in RF was calculated with Mean Decrease Impurity [382]. Only eight features were estimated to have zero importance in the cost-sensitive RF model CS-RF(FN:FP = 90:1). These features were removed, and the model was rebuilt and tested. As a result, the new model performance slightly improved its specificity (The resistant group "Desq–") to 0.658. The AUC and the G-Mean improved by less than 1% to 0.771 and 0.712, respectively, while its sensitivity (the susceptible group "class of interest") had no improvement and remained unchanged. Feature importance is described in supplementary material tables B and C. The final model's performance is described in Table 6.8.

Table 6.8. Simplified hero model's performances on VD



Cost-sensitive RF cost ratio (FN : FP = 90:1)					
The Hero Model			The Simplified Hero Model		
AUC = 0.762		G-Mean = 0.705	AUC = 0.771		G-Mean = 0.712
	Predicted			Predicted	
Actual	Desq$^-$	Desq$^+$	Actual	Desq$^-$	Desq$^+$
Desq$^-$	602	331	Desq$^-$	614	319
Desq$^+$	22	74	Desq$^+$	22	74
TPR=0.771	TNR=0.645		TPR=0.771	TNR=0.658	
Test (VD), n(1029), M(122)			*Test (VD), n(1029), M(114)*		

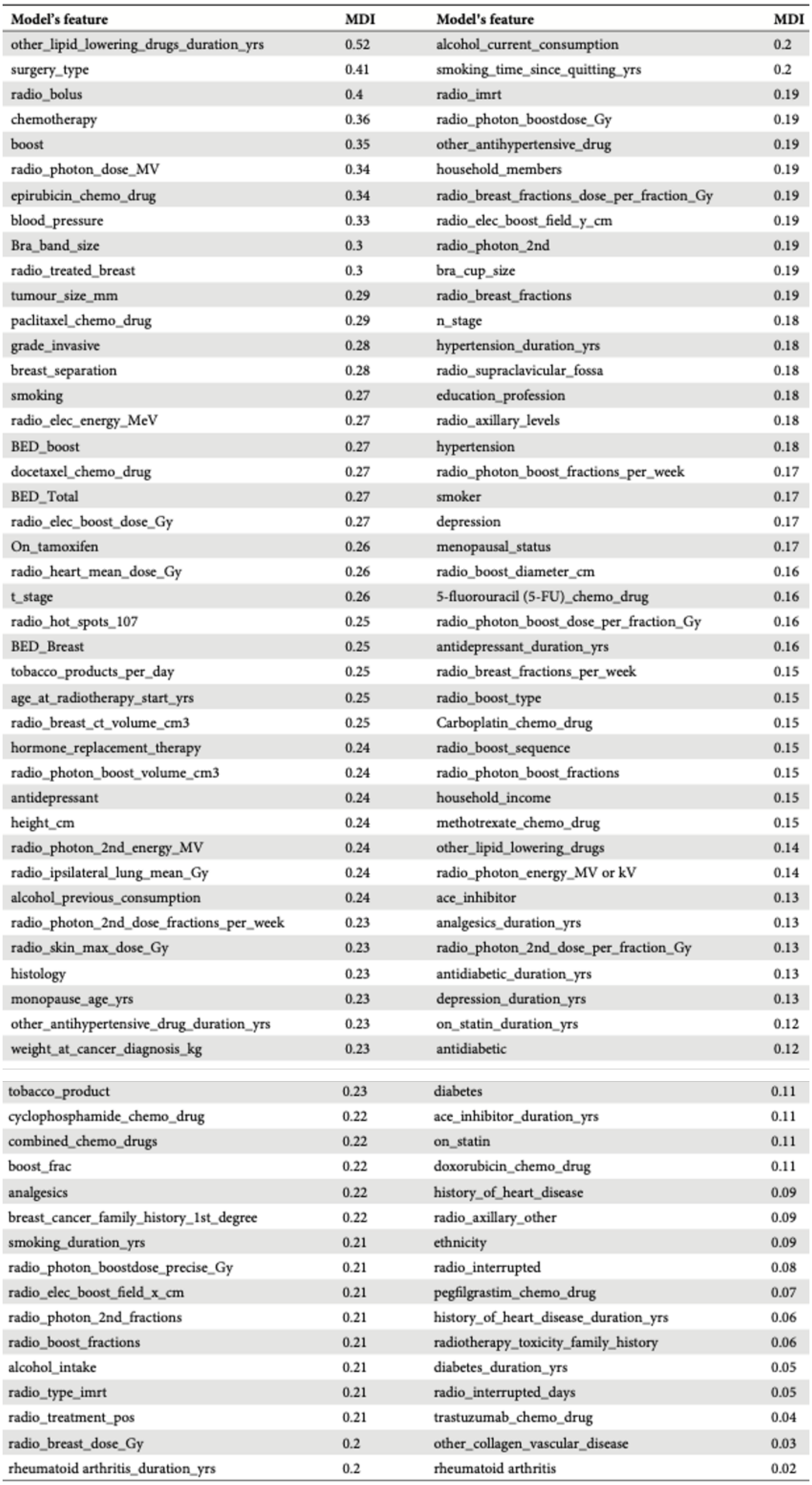
Two additional metrics were also calculated for the simplified model, Matthew Correlation Coefficient (MCC) and the Area Under the Precision-Recall Curve (AU-PRC) score; their values are 0.251 and 0.902, respectively. MCC describes the correlation coefficient between the observed and predicted classifications. An MCC of 0.251 shows that the model's predictions are not random and lean towards strong predictions. The AU-PRC of the final model was 0.902, which indicates a good detection of positive outcomes and a strong prediction performance overall for the model. These two metrics could not have interpreted the domain success any better than it was done already.

6.13 Feature importance analysis

The hero CS-RF (90:1) model had m = 122 features. Eight features were estimated to have zero importance, including features about the presence/absence of systemic lupus erythematosus and other collagen vascular diseases and the use of pertuzumab, eribulin, and amiodarone therapy. In the final step, these features were removed, and the model was rebuilt and revalidated in the VD. Table 6.9 lists the features included in the final simplified CS-RF classifier by order of importance. In descending order, the top 10 features were duration of other lipid-lowering drug use, type of surgery (wide local excision vs quadrantectomy), use of radiation therapy bolus, use of chemotherapy, use of boost, radiation therapy photon dose (MV), use of epirubicin therapy, hypertension, bra band size, and side of radiation therapy.

Given that our hero model has somewhat fewer features and had a multicentre patient sample with diverse radiation treatment regimens, it is promising that its AUC-ROC of 0.77 in the VD is similar to the range of AUC-ROCs reported in the abstract by Reddy et al. [364] our initial models for acute desquamation included 122 features.

Table 6.9. Ranked features' importance in the simplified CS-RF model



Model's feature	MDI	Model's feature	MDI
other_lipid_lowering_drugs_duration_yrs	0.52	alcohol_current_consumption	0.2
surgery_type	0.41	smoking_time_since_quitting_yrs	0.2
radio_bolus	0.4	radio_imrt	0.19
chemotherapy	0.36	radio_photon_boostdose_Gy	0.19
boost	0.35	other_antihypertensive_drug	0.19
radio_photon_dose_MV	0.34	household_members	0.19
epirubicin_chemo_drug	0.34	radio_breast_fractions_dose_per_fraction_Gy	0.19
blood_pressure	0.33	radio_elec_boost_field_y_cm	0.19
Bra_band_size	0.3	radio_photon_2nd	0.19
radio_treated_breast	0.3	bra_cup_size	0.19
tumour_size_mm	0.29	radio_breast_fractions	0.19
paclitaxel_chemo_drug	0.29	n_stage	0.18
grade_invasive	0.28	hypertension_duration_yrs	0.18
breast_separation	0.28	radio_supraclavicular_fossa	0.18
smoking	0.27	education_profession	0.18
radio_elec_energy_MeV	0.27	radio_axillary_levels	0.18
BED_boost	0.27	hypertension	0.18
docetaxel_chemo_drug	0.27	radio_photon_boost_fractions_per_week	0.17
BED_Total	0.27	smoker	0.17
radio_elec_boost_dose_Gy	0.27	depression	0.17
On_tamoxifen	0.26	menopausal_status	0.17
radio_heart_mean_dose_Gy	0.26	radio_boost_diameter_cm	0.16
t_stage	0.26	5-fluorouracil (5-FU)_chemo_drug	0.16
radio_hot_spots_107	0.25	radio_photon_boost_dose_per_fraction_Gy	0.16
BED_Breast	0.25	antidepressant_duration_yrs	0.16
tobacco_products_per_day	0.25	radio_breast_fractions_per_week	0.15
age_at_radiotherapy_start_yrs	0.25	radio_boost_type	0.15
radio_breast_ct_volume_cm3	0.25	Carboplatin_chemo_drug	0.15
hormone_replacement_therapy	0.24	radio_boost_sequence	0.15
radio_photon_boost_volume_cm3	0.24	radio_photon_boost_fractions	0.15
antidepressant	0.24	household_income	0.15
height_cm	0.24	methotrexate_chemo_drug	0.15
radio_photon_2nd_energy_MV	0.24	other_lipid_lowering_drugs	0.14
radio_ipsilateral_lung_mean_Gy	0.24	radio_photon_energy_MV or kV	0.14
alcohol_previous_consumption	0.24	ace_inhibitor	0.13
radio_photon_2nd_dose_fractions_per_week	0.23	analgesics_duration_yrs	0.13
radio_skin_max_dose_Gy	0.23	radio_photon_2nd_dose_per_fraction_Gy	0.13
histology	0.23	antidiabetic_duration_yrs	0.13
monopause_age_yrs	0.23	depression_duration_yrs	0.13
other_antihypertensive_drug_duration_yrs	0.23	on_statin_duration_yrs	0.12
weight_at_cancer_diagnosis_kg	0.23	antidiabetic	0.12
tobacco_product	0.23	diabetes	0.11
cyclophosphamide_chemo_drug	0.22	ace_inhibitor_duration_yrs	0.11
combined_chemo_drugs	0.22	on_statin	0.11
boost_frac	0.22	doxorubicin_chemo_drug	0.11
analgesics	0.22	history_of_heart_disease	0.09
breast_cancer_family_history_1st_degree	0.22	radio_axillary_other	0.09
smoking_duration_yrs	0.21	ethnicity	0.09
radio_photon_boostdose_precise_Gy	0.21	radio_interrupted	0.08
radio_elec_boost_field_x_cm	0.21	pegfilgrastim_chemo_drug	0.07
radio_photon_2nd_fractions	0.21	history_of_heart_disease_duration_yrs	0.06
radio_boost_fractions	0.21	radiotherapy_toxicity_family_history	0.06
alcohol_intake	0.21	diabetes_duration_yrs	0.05
radio_type_imrt	0.21	radio_interrupted_days	0.05
radio_treatment_pos	0.21	trastuzumab_chemo_drug	0.04
radio_breast_dose_Gy	0.2	other_collagen_vascular_disease	0.03
rheumatoid arthritis_duration_yrs	0.2	rheumatoid arthritis	0.02

Information gain (IG) represents the amount of information gained about a random variable or signal from observing another random variable. After the randomised and stratified training/validation data split, a few variables in the VD had a different IG to discriminate between the positive and negative cases compared with the ITD.

However, Zero IG does not negate the feature's worth, as this depends on the machine learning algorithm used. Any given feature could climb up the ranking in terms of IG if additional observations were added to the same data set. Hence, we included all 122 features in the modelling process.

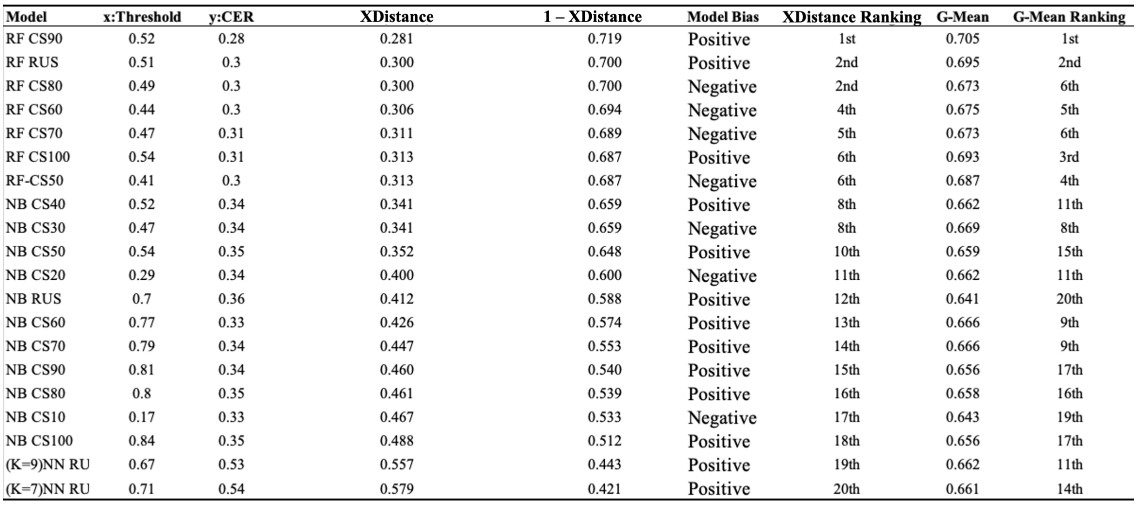
The ten most important features in the final hero model included some that might be expected to predict breast radiation toxicity, such as the use of radiation therapy bolus, chemotherapy, boost, radiation therapy dose, and bra size. Interestingly, the most important feature (use of lipid-lowering drugs) is not usually included in parametric statistical models for radiation toxicity, although HMG-CoA reductase inhibitors (statins) have previously been proposed as radioprotective agents [383]. Yet, unlike traditional statistical probability modelling, feature importance should only be interpreted within the context of the ML prediction model but not outside.

6.14 X-Distance for performance ranking

Table 6.10 ranks the final twenty models, which previously competed to meet two threshold lines' values set by the domain experts in Figure 6.13. the two threshold lines were set since the G-Mean had very light variations that could no longer discriminate with high precision the winning models.

Table 6.10 shows the sensitivity threshold for the crossover error rate (CER) for all the twenty models in Figure 6.13. The XDistance measure is calculated using the Euclidean distance from the no-bias point (0.5,0). The models with a sensitivity threshold greater than 0.5 are denoted as positive-biased models (TPR>TNR), those whose sensitivity thresholds are less than 0.5 are denoted as negative-biased models (TNR>TPR), the model's threshold of 0.5 is a balanced-model (±) (TPR=TNR). For comparison, the G-Mean was also ranked for the same models. The models are ranked based on their XDistance measurements and their G-Mean evaluations.

Table 6.10. XDistance and G-Mean evaluations for the top twenty models



Model	x:Threshold	y:CER	XDistance	1 – XDistance	Model Bias	XDistance Ranking	G-Mean	G-Mean Ranking
RF CS90	0.52	0.28	0.281	0.719	Positive	1st	0.705	1st
RF RUS	0.51	0.3	0.300	0.700	Positive	2nd	0.695	2nd
RF CS80	0.49	0.3	0.300	0.700	Negative	2nd	0.673	6th
RF CS60	0.44	0.3	0.306	0.694	Negative	4th	0.675	5th
RF CS70	0.47	0.31	0.311	0.689	Negative	5th	0.673	6th
RF CS100	0.54	0.31	0.313	0.687	Positive	6th	0.693	3rd
RF-CS50	0.41	0.3	0.313	0.687	Negative	6th	0.687	4th
NB CS40	0.52	0.34	0.341	0.659	Positive	8th	0.662	11th
NB CS30	0.47	0.34	0.341	0.659	Negative	8th	0.669	8th
NB CS50	0.54	0.35	0.352	0.648	Positive	10th	0.659	15th
NB CS20	0.29	0.34	0.400	0.600	Negative	11th	0.662	11th
NB RUS	0.7	0.36	0.412	0.588	Positive	12th	0.641	20th
NB CS60	0.77	0.33	0.426	0.574	Positive	13th	0.666	9th
NB CS70	0.79	0.34	0.447	0.553	Positive	14th	0.666	9th
NB CS90	0.81	0.34	0.460	0.540	Positive	15th	0.656	17th
NB CS80	0.8	0.35	0.461	0.539	Positive	16th	0.658	16th
NB CS10	0.17	0.33	0.467	0.533	Negative	17th	0.643	19th
NB CS100	0.84	0.35	0.488	0.512	Positive	18th	0.656	17th
(K=9)NN RU	0.67	0.53	0.557	0.443	Positive	19th	0.662	11th
(K=7)NN RU	0.71	0.54	0.579	0.421	Positive	20th	0.661	14th

We observe that the shorter the XDistance, the better the model, and the higher the G-Mean, the better the performance. To make both the XDistance and G-Mean comparable, we adjust the ranking scale by calculating $1-XDistance$, which means the higher $1-XDistance$ the better the model (where 0 and1 are not the maximum and minimum XDistance, respectively).

From the ranking in Table 6.10, we observed that the top-performing models selected by the domain experts occupy the top three XDistance ranks, and so are the lowest three CER rates. At the same time, the G-Mean evaluation only agrees with the top two ranked models. The G-Mean ranked the RF CS80 third selected model in the sixth place. The CER maintained an ascending ranking order of the models in parallel to the XDistance ranking for the top 10 models. This plausible ranking could be ought to the dependence between both measures (The CER is a component in measuring the XDistance).

The behaviour of XDistance and the G-Mean can be observed in

Figure 6.14. In Figure 6.14-A, we observe a minor slip in the G-Mean performance for the models with higher sensitivity thresholds. This shows that the G-Mean is very lightly sensitive toward the model's accuracy bias between TPR and TNR. On the other hand, unlike the G-Mean, the new XDistance evolution metric is very sensitive towards increased threshold values (Figure 6.14-B); the XDistance evaluation drops for models with a higher imbalance between TPR and TNR. This explains the agreement between the nominated top three models by the domain experts and those ranked the highest by the XDistance measure. From Figure 6.14-B, we can see that the top model (RF-FN Cost =90) shows less imbalance between TNR and TPR, and it was ranked 1st by the G-Mean; the same model shifted with the least XDistance evaluation.

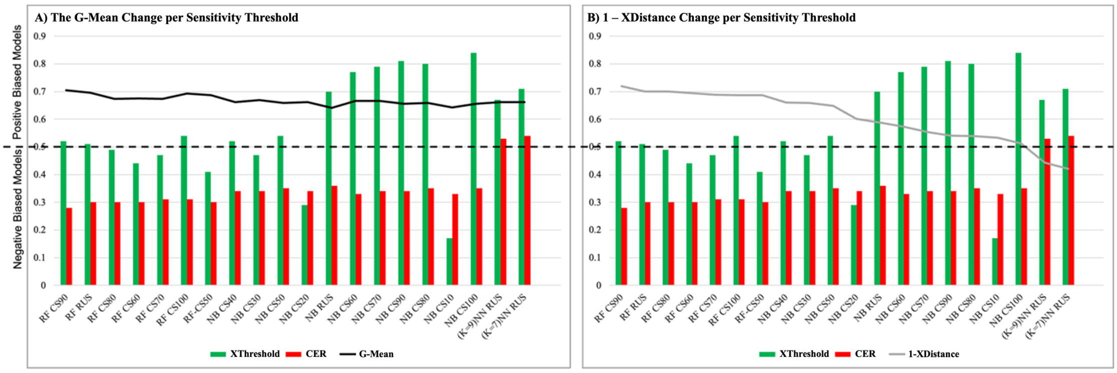


Fig. 6.14 Comparing (1 – XDistance) and G-Mean models performances

The XDistance via the threshold indicates the models' imbalance between TPR and TNR. Therefore, the RF CS90 model with XDistance of +0.272 indicates a better performing positive-biased model with higher sensitivity towards the positive class (class of interest) when compared to the RF CS80 negative-biased model with XDistance of –0.300 that is more specific. Based on the models' validation TP-TN bias, the models in Table 6.10 can be presented in Figures 6.15 and 6.16, where the indicative sign is used to express the direction of the models' TP-TN bias.

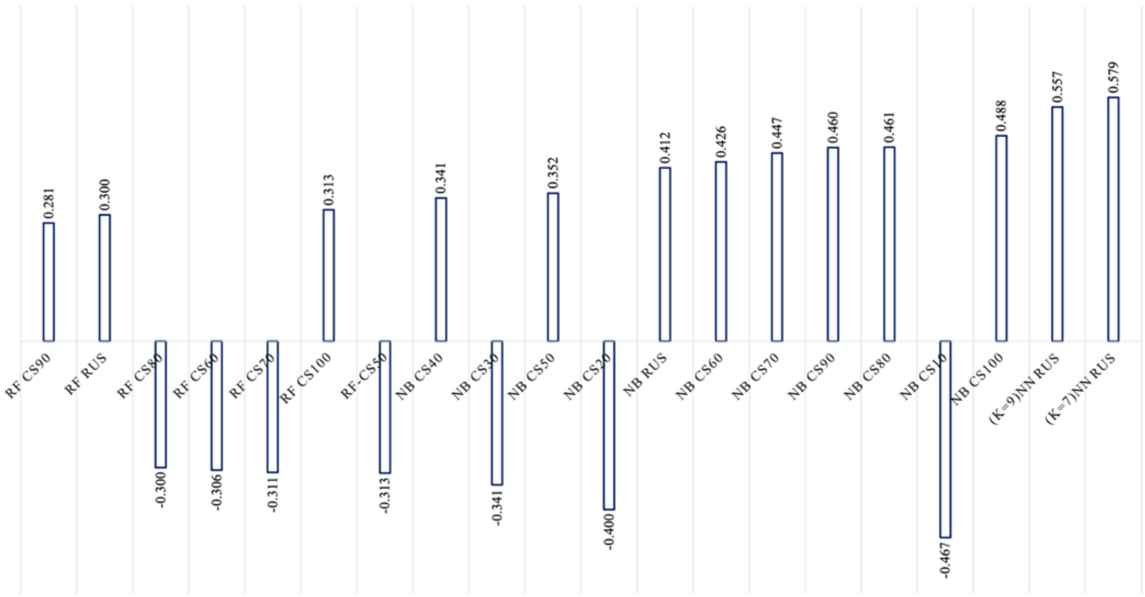


Fig. 6.15 XDistance indicative sign of models' TP-TN imbalance

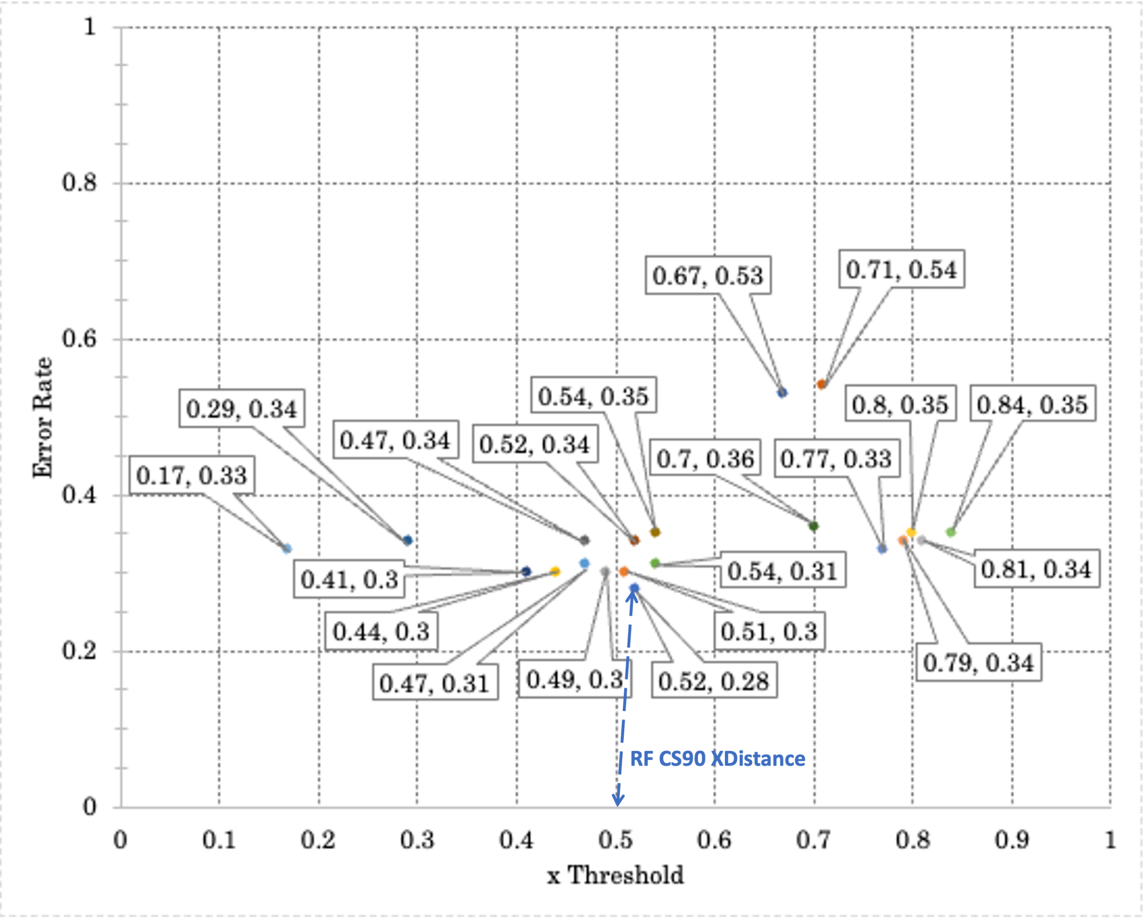


Fig. 6.16 Graphical XMatrix with CER coordinates for the filtered Twenty models

6.15 RT acute desquamation prediction case study discussion

The overall goal of this case study was to build new machine learning models to predict radiation therapy acute toxicity desquamation in breast cancer patients from the REQUITE cohort. The built models are to classify these subjects into two categories, susceptible to occurrence or non-occurrence (resistant).

The ability to predict and classify this variable using simple clinical routinely collected data will significantly impact the identification of subjects likely to avoid QoL deterioration during radiation therapy. The models tested here input features that include baseline characteristics, familial data, breast cancer staging records, chemotherapy-regimen drugs, lifestyle observations, medical conditions, sociodemographic factors, medical operations, treatment history, female-specific factors, mental and behavioural disorders, medications, quality of life and breast RT procedure measurements such as normo-fractionation procedure. The features also included reported RT toxicities risk factors which were previously demonstrated to correlate with acute desquamation significantly. Imaging and genomic risk factors were excluded [384].

Our models initially used 122 input features (attributes) to predict a binary acute desquamation endpoint. The models were built with eight ML algorithms, NB, LR, ANN, SVM, KNN, C4.5, LMT and RF; each has a different learning scheme. In addition, a purity-based ranking technique, IG, was calculated to evaluate the information worth of each input feature independently in relation to the class attribute.

When observing IG evaluation after the randomised and stratified training/test data split, it was noted that a few variables in the test dataset (VD) contained a different worth of information as compared to the training set (ITD). Therefore, a way to interpret the calculated IG values is the possible presence of associations between each feature and the class labels in each training dataset.

This purity measure differs from correlation association and is not utilised as a feature selection in this study. Observed IG evaluation also showed that some variables in the VD set contained more information than the ITD. In ITD, it was observed that “radio_skin_max_dose_Gy”, “BED_Breast_Gy”, “radio_breast_fractions_dose_per_fraction_Gy”, “radio_breast_ct_volume_cm3” and “radio_photon_2nd_fractions” dominated the top five ranks in purity values in relation to the class variable (acute desquamation endpoint). After balancing the two classes with RUS resampling technique, "radio_skin_max_dose_Gy" still reserved the highest IG evaluation, and "radio_breast_fractions_dose_per_fraction_Gy" slipped to sixth place while "BED_Breast_Gy" remained in the top five; other new predictors soared to the top five IG ranks: those are "radio_type_imrt", "radio_boost_type" and "radio_photon_energy_MV or kV". In the oversampled dataset (ROS), similar to ITD, “radio_breast_ct_volume_cm3” and “radio_skin_max_dose_Gy” were in the top five places, while three new predictors joined the top five ranks - “BED_Total_Gy”, “weight_at_cancer_diagnosis_kg” and “radio_photon_boost_volume_cm3”. Unlike all training sets, in SMOTE synthetic oversampled dataset, five new predictors occupied the top five ranks, those being "breast_separation_cm", "band_size_UK_inch", "bra_cup_size", "household_members" and "height_cm".

This information theory investigation into the models' features based on domain experts' advice adds a layer of details to the observed correlations in previous studies by describing the strength of each feature to discriminate between the positive and negative classes [385 – 391].

Furthermore, when considering the ITD, RUS, ROS and SMOTE datasets, some variables showed no purity towards the class: ITD had 42 predictors with zero IG, RUS had 59 predictor variables (the highest), and

ROS and SMOTE had the least predictors with zero IG of 11 and 12 respectively. Zero IG does not negate the potential relevance of these predictors in the predictive models, as they may climb up the ranking if additional records are added to the same dataset. They simply mean that based on purity and entropy in these training datasets, they do not help distinguish between both class labels at the endpoint. Some ML models may still calculate otherwise and utilise them in building predictive models depending on the learning mechanism. Thus, all 122 predictors were included in the modelling process.

Again, as the previous case studies witnessed in machine learning modelling, tackling the imbalanced class learning problem noticeably impacts the performance of standard parametric and non-parametric ML algorithms. Also, the classification modelling performance in the training phase is severely affected by class separability. Thus, training the standard ML algorithms with highly imbalanced classes without adjusting the training instances results in an accuracy bias towards the majority class.

In this case study, we tackled that bias by applying two approaches. In one tactic, resampling techniques (RUS, ROS and SMOTE) were used to adjust the class imbalance in the classification training phase at the dataset level, which amplified the IG in many input features. The other approach (a cost-sensitive tactic) awarded incremental higher weights to the records in the minority class while maintaining unchanged levels of information in the input features.

It was observed that the cost-sensitive approach achieved the highest ranks in the models' evaluation. It remains unclear whether other remedies for imbalanced data classifications, such as Ensembles Learning (implemented at the algorithmic level), could result in better performances [392]. However, the advantages of resampling techniques evaluated here include simplicity and transportability. Nevertheless, they are limited by the amount of IG manipulation since their application results in biased predictions toward the minority class.

The excessive use of such techniques could result in overfitting, as seen in the ROS and SMOTE models. In this study, the original REQUITE cohort dataset was highly imbalanced. Traditional ML algorithms were sensitive to higher information gains. They tended to produce superb performance results in training for ROS and SMOTE datasets, but when testing the models, the overall model performance often dropped below the training phase performance. Unlike resampling techniques, cost-sensitive classification is proven complex to determine the exact penalty for minority records misclassification. Also, as observed in the results, the complexity dramatically increases since the attention (depth of impact) to the minority records of different ML classifiers of various learning schemes is shifted differently for the same misclassification penalty when building predictive models. Adding to the mixture of complexity, a good choice of evaluation metrics becomes crucial.

As previously described in the literature review chapter, Chapter 3, some metrics, despite how popular they are in a research area, i.e., Accuracy and AUC-ROC, produced deceiving good measurement evaluations. Therefore, more imbalanced modelling-focussed metrics were chosen, such as Balanced Accuracy, Youden's Index, the G-Mean and AU-PRC.

This study showed that applying the correct level of resampling without overly disrupting the original data information in the RUS-based method with the slight manipulation of IG levels, together with the desired choice of performance metrics, produced a good prediction solution [303]. Furthermore, the RF-RUS model competed with further developed models with cost modifications in the case of cost-sensitive classification. Three of the 89 models reported in this study satisfied the trade-off threshold conditions (Tables 6.6 and 6.7).

However, one "hero" model gathered experts' interest in this specific domain problem: a cost-sensitive RF model with an FN:FP misclassification penalty ratio of 90:1. Nevertheless, the effect of the classifier's learning scheme becomes highly noticeable in imbalanced datasets when the minority classes prediction accuracies (TPR) are compared. The results also showed that improving the ITD models TPR with CS-classification does not massively impact the positive group by putting the majority group at a higher disadvantage of deteriorating its TNR, i.e., the NB case. It is observed that some algorithms are highly resistant to higher misclassification costs to improve their original TPR in the imbalanced learning setting, i.e., in the case of ANN.

In the resampled models' results analysis, the learning scheme's impact decreased with the class imbalance severity in the datasets compared to balanced datasets. In addition, classifiers behaved differently for the same cost matrix in cost-sensitive classification when trained on the same dataset.

Our "hero" model was further simplified by discarding eight features. According to RF model-based feature selection method Mean Decrease Impurity (MDI), these features were deemed unimportant of zero value. The "hero" classifier is rebuilt with the remaining 114 features. The performance of the "hero" model continued to show a slight improvement in TNR. The MDI feature selection is biased towards preferring variables with more categories [393]. This bias is not a problem in our study since MDI was only used to optimise (simplify) a model with known performance. However, suppose the dataset contains two (or more) correlated features from the model's point of view. In that case, any of these correlated features can be used as a top predictor without preferring one over the others. Once one of them is used, the importance of the others is significantly reduced since the impurity they can eliminate is already removed by the first selected feature. Therefore, they will have lower reported importance. This reduction of importance is not an issue when we want to use this feature selection technique to simplify the model since it is desired to remove mostly unimportant features.

Nevertheless, it can provide a misleading perception that one of the variables is a strong predictor when interpreting the model. In contrast, the others in the same group are unimportant, while in fact, they are very closely associated with the response endpoint (Figure 6.11 and Table 6.9). The misinterpretation of unimportant features' removals is somewhat reduced thanks to random feature selection at each node in Random Forests. However, the generalised effect within the averaged model is not entirely eliminated. The difficulty of interpreting the ranking of associated variables is not Random Forest specific; it applies to most model-based feature selection methods.

Like most biomedical case studies, when biochemical tests are performance assessed, the data obtained is heavily skewed (imbalanced) in our study. Typical disease prevalence is in the range of ~10% for those with the disease, and ~90% do not have that disease. It is common to use the AUC-ROC curve to evaluate the clinical performance validity of a biochemical test. The AUC-ROC curve is a graphical representation of the trade-off between TPR and FPR for every possible cut-off for a test or a combination of tests. The AUC- ROC gives an idea about the benefit of using the test in question.

However, the highly imbalanced datasets tend to provide a much better ROC curve; therefore, visual interpretation and comparisons of AUC-ROC for ML models trained with imbalanced datasets can be misleading [322], as observed in all ITD-based models in Table 6.5. Therefore, additional performance metrics are required to provide a more accurate representation of the model's validity. The TPR and TNR are used less frequently than ROC curves, but as we examined the models, assessing additional performance metrics is proven to be a better choice for imbalanced datasets.

Setting a graphical TPR-TNR trade-off threshold that maximises correct classifications gains and minimises misclassification losses indicates the class importance in the domain experts' view. It allows for a more pragmatic final model selection.

Currently, mechanistic models are embedded within the treatment planning systems to predict RT complications; these are Lyman–Kutcher–Burman models [394][395]. These models allow for effective biological optimisation of the delivered radiation dose among competing treatment strategies; however, the handmade exceptions in their algorithms mean that they often fail to predict the actual side effects induced by RT.

In PubMed/Medline database, the current available studies indicate only two are viable [8][362][364] that produced clinically valid ML models for detecting acute side effects of breast RT. In one study [362], models were built based on detecting the body-surface temperature increase. Thermal images of the irradiated breast were taken from a small population of 90 patients at four consecutive time points. The caveat for this approach remains to be the large-scale analysis of RT toxicities at the expense of time required to obtain the imaging data and accounting for the considerable variation between individual patients' normal tissue reaction to RT and the resultant toxicities. The other is a comparative study [364] that trained a group of ML algorithms on a large population of 2277 patients from 5 clinical centres. And it achieved a good AUC-ROC performance. However, the prediction models are complex, using more than 300 input variables. Using such a large number of variables makes it hard to follow and interpret the model's output. The final and recent study by Rattay et al. [396] attempted to create simpler toxicity prediction models that excluded dosimetry and radiomic data. Unfortunately, the model did not clinically validate in the

REQUITE cohort.

Unlike the previous studies, our study accounts for fewer and easy-to-obtain variables during the RT treatment planning phase, incorporating the largest cohort among other studies. As a result, our models are considered encouraging. In addition, data-driven studies often lack reporting on the data pre-processing techniques involved in building their ML models. By reporting the full methodology designed and delivered by an interdisciplinary team of experts, we provide further clarity and contribute further to the research society.

Thus, our formulated approach equips researchers with a new pragmatic domain-driven approach highlighting concerns when applying data imbalance strategies and assessing multiple models for similar real-world clinical problems.

6.16 RT acute desquamation prediction case study conclusion

Our study shows that applying our new OCTOPUS framework to model phenotype and clinical variables datasets has the potential to offer a fast and inexpensive solution to predict acute toxicities for breast cancer RT patients as an adverse event in healthcare.

This potential was examined by aligning the classification task to predict specific adverse skin effects based on Common Terminology Criteria for Adverse Events. This study's selection of a binary-class prediction task is strategic to include patients classed within severe, life-threatening and death criteria. It identifies patients at higher risk of developing acute desquamation conditions and are more likely to benefit from treatment plans to be personalised and trigger discussions about treatment risks and benefits with patients. The process of training various ML algorithms with 10-Fold Cross-Validation and validating the models with a class-stratified isolated group of patients of the same size as the training data makes this study encouraging for follow-up research and validation on external cohorts.

If validated, then there is potential to proceed with medical screening research.

Our final model has the potential to aid researchers in further understanding the relationships between the used features and moist desquamation side effects.

Before being embedded into applications, the model requires further improvements in line with the optimised radiation dose output obtained from the current mechanistic models. The final model may also be improved by utilising the treatment dosimetry measurements obtained from the current treatment planning system to predict acute desquamation accurately. In addition, decisions obtained from the legacy systems and the new models are recorded and compared.

This domain problem is the first to use only clinical features at a CTCAE >3 setting to predict acute toxicities with ML. This study has the largest number of patients in modelling and validation, among other known studies. This study could be used as a benchmark for future studies to compare its results to other research from the same domain. Nevertheless, further analyses will be followed where additional methods to improve the outcomes will be investigated.

Limitations of this and many other ML papers used in radiation oncology are the number of variables used compared to routine practice, the different toxicity scales and the grades for an acute skin reaction and ulceration that define the class endpoint.

Real-world applicability is also reduced due to unrealistic datasets. However, the volume and variety of data routinely collected on patients will only increase over time. Indeed, many of the variables currently collected in routine practice are not fully utilised. For example, past medical history, drug history and family history form a large number of binary variables in the REQUITE dataset but, at present, are often recorded as free text on the first encounter between patient and oncologist. Regardless, similar models using more limited datasets should be developed and tested before an ML approach to predict RT toxicities can move beyond the research setting into clinical practice. Despite a good amount of research in ML methods for toxicity assessment, this is the first effort to summarize the field's current state and produce competing prediction models to the best of our knowledge.

Further limitations arise; despite the rigorous error detection in the data preprocessing phase, we cannot exclude errors from manual recording during data collection. According to the REQUITE study protocol, patients were assessed at the start and end of treatment and annually thereafter. This may have missed cases of acute desquamation as acute radiation toxicity is known to peak up to 2 weeks after the end of treatment. Although we incorporated differences in radiation therapy techniques by including all available recorded treatment parameters in the analysis, this may not fully account for variability in treatment plans between participating centres or treating physicians. Similarly, variable transformation or feature engineering (e.g., calculating the BED and binarisation of chemotherapy drugs) could have led to the creation of a new feature that is less powerful and suppresses important information inferred by its raw components.

In modelling the radiation therapy dose variable, alternatives such as a categorical variable divided by type of radiation therapy regimen could have been used (e.g., hypo- vs standard fractionation). Variable aggregation could have led to model overfitting due to misleading combined features and may show false significance or insignificance in the analysis.

Although the resampling techniques used in this study have advantages in their simplicity and transportability, other remedies to address imbalanced data, such as ensemble learning (which is implemented at the algorithmic level), could be used to improve model performance.

Cost-sensitive learning was selected to penalise false negatives.

However, its application depends on the clinical situation. For example, suppose a model was designed to allocate patients to a toxicity-lowering radiation therapy regimen that might affect tumour control. In that case, FPs may need a higher cost than FNs. This study used the impurity-based ranking mean decrease impurity filter to simplify the final model with a known performance, but it is important to keep in mind that feature selection based on impurity reduction is generally biased toward preferring variables with more categories.

The top-performing model could classify patients with acceptable performance in the validation cohort (AUC-ROC = 0.77). However, before they can be used in clinical practice, further optimisation of ML prediction models, including genomic markers, is required, and the models should be validated in external cohorts.

Using the Crossover Error Rate (CER) and the new XDistance measure demonstrates an example of the true original interdisciplinary nature of our research. Both metrics have achieved the level of discrimination among models using the success criteria. Both nominated the exact top three performing models selected by the experts, while the G-Mean managed to rank the top two. The CER shows that there is still room for exploring performance metrics for machine learning in different domains. Our research venture with XDistance showed a possible graphical metric that offers somewhat a TP-TN-balanced model assessment. Like all other performance metrics, the CER and XDistance have their limitations. First, we only applied these two metrics to twenty classifiers from a few varieties of learners. Therefore, further research is required to establish whether other classifiers' crossover error curves share similar behaviour.

We only applied XDistance to a binary classification problem, and it is still unknown if the new metrics can offer a reliable assessment in multi-class classifications. In addition, currently, we have no automatic way to determine the actual crossover between the FN and FP curves other than visually. Perhaps different image processing techniques could extract the crossover coordinates from many graphs. We also acknowledge that the XDistance and the CER may not be better than other metrics at a larger scale of models at interpreting different success criteria.

Finally, we assume that the XDistance has no defined maximum and minimum range. It follows our empirical assumption that real-world classification performance can rarely offer a perfect classification of all classes. Thus, there should always be a crossover point to measure the XDistance, but it becomes hard to generalise our assumption without knowing the behaviour of all algorithms' curves.

In the unlikely perfect classification scenario that occurred, the crossover points would not be formed due to FPR and FNR equal to zero, and the XDistance becomes undefined. Some may argue that the CER in that scenario is zero; others may say the crossing happens at ∞; hence, the XDistance is also zero indicating perfect classification or 1 or ∞, which seems to be a new paradox since $0 \neq 1 \neq \infty$. Similarly, another hypothetical scenario creates an inverted classifier where FPR and FNR are equal to 1. Therefore, here we open the door for enthusiastic new researchers to research in this area further.

Chapter 7

Discussion and Conclusion

This chapter covers

- *Contributions to knowledge*
- *Limitations*
- *Implications for practice*
- *Future work*

The overall aim of our research is to create a data-driven framework, OCTOPUS, to develop predictive new models that can be used as new potential tools for identifying individuals' susceptibility to adverse events and complications in healthcare applications. Predicting the occurrences of adverse events is the first step of preventive healthcare measures. Therefore, in this thesis, we defined susceptibility to adverse events and complications into two classification problems, a binary-class problem that includes resistant and susceptible subjects and a three-class problem which includes resistant, susceptible, and hyper-susceptible individuals. The latter can also be mapped to a qualitative low, moderate, and high-risk scale.

This thesis established that the definition of susceptibility to adverse events differs among various domains. In each domain, the susceptibility boundaries can also be hard and soft depending on the domain experts' definition of the endpoint of the predictive task. However, an efficient prediction of patients within the given definitions is needed for making effective recommendations of interventions, i.e., lifestyle modifications and treatment planning.

The correct classification of subjects for the right interventions to avoid disease and illness is a type of preventive healthcare. The result improves individuals' quality of life (QoL), and also prevents the potential occurrence of future deterioration of their health and QoL following a particular course of intervention.

In this thesis, we researched and discussed three impact scenarios in the case of the misidentification of individuals' susceptibility status. Two could result in an increased cost due to unnecessary interventions (i.e., misclassification of a healthy subject in the visceral fat-associated case study) or potentially an unnecessary change in treatment course or planning (misclassification of a resistant patient in the radiotherapy project). However, these scenarios may be managed at a further cost with additional clinical tests to make informed decisions.

The third scenario, from the advanced radiotherapy toxicity case study, could result in a significant negative impact on the patient impairing their QoL in the short term, long term, permanently or even increasing the patient mortality risk. Although experts' budgets may manage additional costs, there is no acceptable compensation for a permeant impairment to QoL or loss of life. From that point of view, the correct classification of susceptible patients outweighs the correct identification of resistant individuals. Both case studies justify the need for our research.

The scale of susceptibility status (endpoint definition) in both case studies was set with hard boundaries. These boundaries were set by life scientists based on documented observations in health sciences research in the case of susceptibility to VAT-associated diseases. Also, they are based on clinically observed outcomes by physicians in hospitals in the case of susceptibility to radiotherapy early toxicities.

Both case studies are life-critical in principle and practice. For the dangerous visceral fat levels study, the current traditional methods used to screen VAT amounts, such as MRI scans, are proven very costly. The high cost may deter patients from screening due to a lack of affordability. Thus, individuals progress silently to develop a chronic disease in the short or long term.

In that instance, such cost savings highly outweigh the instant benefit of the early identification of susceptible individuals. Nevertheless, that expense falls behind on the long-term scale compared to the cost of managing the positive diagnoses of chronic visceral fat-associated diseases, including diabetes, cardiovascular diseases and cancers.

The breast cancer radiation therapy (RT) case carries out a much higher urgency to identify susceptible patients. The Current methods used in advanced radiation therapy planning to predict side effects, such as mechanistic models embedded within the treatment planning systems, only allow for effective biological optimisation of the delivered radiation dose among competing treatment strategies. The exceptions in these models are handmade, which often fail to predict the side effects induced by RT. Therefore, the benefit of instant correct prediction of susceptibility to RT toxicity, in this case, outweighs the cost.

Therefore, the cost-benefit win in the RT case is due to the nature of acute toxicities, whose effects and outcomes are often observed within a short window. This window usually runs from a few days up to nine weeks but with a likely lasting impact of a lifetime.

7.1 Contributions to knowledge

Our new octopus framework yielded new original models for preventive healthcare. It was formulated by examining various concepts and techniques from multiple fields, including machine learning, life sciences, clinical sciences, psychology, cyber security, test engineering and nuclear physics.

As presented in Chapter 1, the contribution of knowledge in this thesis was achieved in two parts; the primary objectives include developing the Octopus framework and utilising the new framework to develop new predictive models for preventive healthcare, assessed by domain experts. And the secondary objective includes proposing potential new solutions or mitigation methods to achieve an acceptable modelling performance.

7.1.1 Octopus data-driven framework

We met the primary objectives A and B, in Chapter 1 of this thesis, by introducing Octopus, our newly formulated data-driven framework, in Chapter 4 to predict susceptibility to adverse events in preventive healthcare is promising. It was applied in Chapters 5 and 6 to confirm its reliability. With carefully scoped methods and considerations while allowing for trivial enhancements, it has the potential to provide faster means of early screening and reduce the cost of the lack of intervention.

To formulate the Octopus framework, we investigated the potential impact of various methods, techniques, tactics, and strategies before integrating any of them into modelling patients' susceptibility to adverse events. The new framework is a modest contribution to data science and healthcare communities. It includes heuristic-systematic recommendations and the scope of methods for successful modelling meeting specific areas of concern in health care. These methods and recommendations proposed mitigations in data quality assessment, cleaning, preprocessing, engineering, modelling and performance evaluation.

Examples of such recommendations can be seen early in Chapter 4, starting with conceptualising the susceptibility problem in healthcare and the VISPAQ quality assessment of data sources. To enhance data quality, we adopted complex techniques for the first time from other domains, such as Boundary Value Analysis (BVA), from test engineering to find common statistically hidden irregular values. We also shared our considerations for machine learning-driven measurement scale topology.

In addition, we designed our new Multi-Method Imputation process (MMI) to assist healthcare data analysts in minimising the spread and impact of missing data on their analysis. Based on learning schemes, the Octopus framework scoped only a range of data preprocessing techniques and eight machine-learning classification algorithms out of hundreds of available methods in the field. It recommended further suitable adjustments that could benefit ordinal class and cost-sensitive classifications.

The logic behind selecting the framework's methods is made clear. All the scoped methods emerged from following a heuristic-systematic research narrative to filter accumulated research presented in the literature.

Finally, OCTOPUS was evaluated by its application to develop predictive models for preventive healthcare in two case studies. In each case study, the methodology demonstrated the structural flow of using a specific collection of methods to model the defined healthcare endpoints.

7.1.2 New predictive models for preventive healthcare

In Chapters 5 and 6, we modelled real-world preventive healthcare problems. The new models were approved by data scientists, life scientists, healthcare professionals, clinicians, surgeons and physicians from the UK National Health Service (NHS). Therefore, our primary research objectives C and D in Chapter 1 were also met.

Using the OCTOPUS framework, new machine-learning models were built to predict susceptibility to adverse events that met the domain experts' desired success criteria. The new successful models are regarded as new potential inexpensive predictive tools by the domain experts that will require further validation. The models also helped further understand the potential predictive role of newly engineered features in both studies, the Physical Activity Index (PAI) in visceral fat-associated diseases case, and the chemotherapy agents and the RT Biological Effective Dose (BED) in the advanced RT toxicity prediction case study.

7.1.3 New modelling enhancement methods

In the Octopus framework, Chapter 4, we combined state-of-the-art data visualisation techniques with traditional unsupervised machine learning and information theory to formulate a new sampling strategy to tackle imbalanced learning from imbalanced datasets. We introduced a Minority Pattern Reconstruction (MPR) resampling method favouring hyper-susceptible subjects (the minority group at the highest risk). Additionally, we proposed using its hybrid with two other sampling techniques, SMOTE NC and the conventional Random Under-Sampling (RUS), seeking possible performance improvements.

Octopus framework provided a detailed description and critique of various classification evaluation metrics. It closely examined their effectiveness and interpretations, specifically in imbalanced learning. Therefore, within the Octopus framework, we developed a new potentially balanced metric, XDistance, influenced by the ultimate fact that bringing someone from the dead is the same as erasing someone's memory. Both scenarios are impossible, at least for normal organic beings. Therefore, we considered a breach of confidentiality in cybersecurity equivalent to a patient's death in healthcare. The impact of both scenarios represents a point of no return. Thus, we derived XDistance from the Cross Error Rate (CER). The CER is a performance metric designed to evaluate biometric systems in Cybersecurity. XDistance successfully ranked performances of susceptibility classifiers. Hence, XDistance, a new empirical graphical performance metric, is introduced for imbalanced learning.

Therefore, the secondary research objectives are met by developing a new sampling strategy and a performance ranking metric.

7.2 Limitations

When creating a new framework, there is a large number of methods and considerations to take into account. Most of the recommendations and considerations are based on empirical results published in research, and no one approach can fit all problems. This is seen in the literature review chapters.

Therefore, the traditional belief that a single algorithm cannot fit all is correct in line with the No Free Lunch Theorem. However, our research shows that a careful choice of a pool of methods and algorithms in an end-to-end modelling workflow potentially yields a robust framework that can be suitable to solve multiple problems in a single area—susceptibility to adverse events in preventive healthcare in our case. But that does not rule out potential limitations, including the curse of iterations driven by the wide variety of parameters and strategies governed by the various possibilities of problem definitions. Some limitations identified in our work are:

A) Susceptibility is one problem; however, it has multiple interpretations in each domain. Different interpretations of the term susceptibility still exist, even in a single discipline such as healthcare. This variety limits the type of analysis. And the varied success criteria definitions vary the evaluation metrics and adjustments to the chosen methods used for each problem's output.

B) We are in a new era when clinicians have heard about the great potential of machine learning in clinical decision support systems. Trust can be further established with domain experts' full participation in all data mining project phases. This leads to interpretability limitations, the creation of opaque models, by including a large number of features while handling the consequences of such modelling. Despite their opaque nature, they are accepted by domain experts. Despite domain experts' acceptability, using many features questions the generalisation-interpretability trade-off. The domain experts in our case studies seemed to favour finding a reliable model to the interpretability of the model itself. This may have been driven by the significant involvement of domain experts in all phases of the projects, including setting the data collection protocols, examining the quality, data cleaning, data engineering, formulation of a methodology, model selection and evaluation.

C) There is no definitive mitigation to deal with missing data in data preparation. Still, depending on the problem at hand, with a satisfactory amount of literature reviews, researchers could form their own process to deal with missing data. We formulated a new multi-method imputation (MMI) process for handling missing data in confirmatory health analysis. However, recording each step in the flow is time-consuming.

D) One must take care when deciding on the topology used to define the features measurement scales, which is also governed by the type of machine learning tool used for modelling. The measurement scales are still an open field of research, and we formulated our own topology of measurement scales to suit our analysis and limited by the modelling tool. Initially, our choice was driven by avoiding creating rules outside the boundaries of normal values. In our framework, the formulation of such rules was not examined in the context of rule-based models. The lack of such an examination is driven by favouring prediction generalisation to interpretability.

E) In imbalanced learning, one must be careful when applying data resampling techniques. The effect of their application can increase the data purity resulting in higher information gain evaluations in the unit of divergence bits. Our experiments showed that our selected machine learning algorithms are susceptible to overfitting with the artificial increase in Information Gain (IG) evaluation. Increased IG could sacrifice the classifier bias for more variance, producing overly optimistic models. This was specifically seen in oversampling methods, ROS and some cases of SMOTE. Our framework is limited by not specifying a definitive recommendation for over-sampling thresholds to prevent overfitting.

F) In modelling, we observed that Random Under-Sampling (RUS) outperformed other Over-Sampling Techniques, such as Random Over Sampling (ROS) and Synthetic Minority Over Sampling Technique (SMOTE). These methods change the features' IG levels when examined in units of divergence bits. However, when developing the new MPR resampling technique, we presented the information leakage by calculating each resampled set's normalised features' information gain (IG). RUS showed minimal features' IG leakage and shrinkage, preserving similar normalised IG levels compared to the original imbalanced dataset. Perhaps its superior performance may be attributed to safeguarding the natural noise-to-signal ratio within the classes. This preservation of information may shape the classifiers to achieve better bias-variance trade-offs, benefitting the minority region of the instance space.

The reconstruction of a data set with the new MPR resampling technique and its hybrids favouring the minority class maintained similar normalised IG levels to the original imbalanced set's features. Thus, it highly improved imbalanced learning. MPR hybrid sampling outperforms direct RUS application in modelling the VAT prediction in females. However, the MPR technique was only applied to a three-class problem, and the opportunity did not arise for its application to problems with different counts of class labels.

G) Noisy data can be data that is corrupted by errors due to improper data collection procedures. This is different to genuine class overlapping. Although our research showed that reputable data sources would likely have better data collection protocols than others, we provided a VISPAQ six-point process to check the data source before committing to a case study. Data errors may not appear abnormal when measured with the traditional statistical outliers' detection methods like IQR, hence the use of BVA and ECP. However, using both techniques can be time-consuming.

H) A collection of standard data preprocessing techniques may be sufficient to engineer a dataset enough to produce promising results. Octopus framework gives recommendations to the sequences of applying some methods. However, some of these recommendations are always limited by the characteristics of the data and the learning algorithm.

I) Unlike multinominal classification, cost-sensitive modelling is easy to implement in binary classification. However, each machine learning scheme is affected differently for the same cost matrix. This observed behaviour was mitigated using an incremental inverse class distribution approach in steps equal to multiples of the minority-to-majority class ratio in the Radiotherapy Toxicity prediction project. This means that if a single derived cost for misclassification has ever been calculated, it may not be sufficient to improve classifier performance. Instead, this cost might be more suitable for selecting the model that best fits the success criteria. When cost-sensitive classification was applied from Octopus framework, it was implemented with an incremented inverse class distribution cost matrix. One limitation to mention, in the matrices, the false negative penalty has a lower limit but no stopping unified upper limit. Our observations were made over 8 algorithms; the upper limit penalty per algorithm in our framework is defined by reaching the performance of the inverted base learner. This upper limit could be different among algorithms.

J) The XDistance and the CER metrics are limited to the model selection criteria provided by the experts in the radiotherapy study. We are not sure how the FN-FP curves will behave for different algorithms. We do not know if the XDistance could be used to interpret other success criteria in other models. Also, it is unknown how other distance calculations rank the models' balance estimates compared to the Euclidean distance.

K) We had the opportunity to analyse the use of newly engineered variables in our case studies, such as the Physical Activity Index (PAI) and Biological Effective Dose (BED). However, the PAI was analysed independently outside the models' context. BED was only analysed in the context of the chosen hero model. Different machine learning models utilise features differently; thus, it is not possible to draw a solid unified conclusion about their influence on the predictions in their fields.

7.3 Implications for practice

Our research mainly focussed on designing and developing a data-driven framework to predict patients' susceptibility to adverse health events using the UK Biobank and the European study REQUITE. We showed original research contributions in multiple areas. For the interdisciplinary community, our framework and models may seem a giant leap forward from an application point of view in susceptibility modelling. Still, the same contributions may seem a small step to others from the computing society scratching the surface of deeper computational issues. Some of the implications in practice can be:

A) Our data-driven framework developed new promising models. Selected models fulfilled the domain expert's success criteria with the potential of being used as new tools for screening only if they pass external validation.

B) The developed models require input values from a large number of features. These features are collected routinely and stored electronically. Therefore, due to the large number of features, it is impractical to feed their values manually. The inputs to the models must be automated from electronic health records repositories if such models are considered for subjects' screening.

C) The unknown performance of our framework and models if applied to other susceptibility endpoints and different adverse effects. Moreover, further restrictions are linked to the new models themselves; even if the models were validated on external cohorts, the models might become unstable due to concept drift after some time or population change.

D) Applying BVA and ECP requires inspecting all features while accounting for their logical relationships. Therefore, it is only advised for critical data cleaning tasks, especially if the data was manually collected. Also, the larger number of features, the higher the complexity of applying both methods.

E) Despite the rigorous error and irregularities detection in the data preparation phase, we cannot exclude errors due to manual data collection recording.

F) We used datasets from the UK Biobank and REQUITE consortium; the sample count was in the thousands, and the characteristics of our framework are not known for samples of larger dimensions (i.e., hundreds of thousands or in millions) should these become available. Perhaps, active learning could handle classification tasks better on larger datasets.

G) Our new MPR sampling strategy was applied to a multiclass imbalance problem in one case study, where the hyper-susceptible subjects were the minority and the group of main interest, followed by the susceptible individuals (moderates), also had a certain imbalance ratio. Whether our sampling strategy would work in other multiclass imbalanced learning problems is unknown.

H) The scope of evaluation metrics is highly dependent on the interpretation of the defined success criteria. In some cases, a selection of multiple performance metrics may be required to interpret a given success criterion. In addition, performance metrics from other domains, such as the Crossover Error Rate (CER), could also be applicable. Therefore, the developed XDistance metric seems promising in ranking the models' performances based on their FN-FP balance. However, a computational error could occur in severely imbalanced models. Severe imbalance prevents the formation of crossover error curves, leading to calculation errors of CER and XDistance metric. Still, this error can be used to indicate severely imbalanced models.

I) In the new framework, we always examined information gain differences between the training and test sets. Thus, we made their size proportionally almost equal. Nevertheless, how our models will perform on different training-test split ratios is unknown.

J) The susceptibility to visceral fat diseases prediction models used the summary form of physical activity features, the Physical Activity Index (PAI). We do not know the impact of modelling the susceptibility to visceral fat endpoints using the raw PAI components (granular features).

7.4 Future Work

The limitations listed above represent the best starting point for further research. Applying our framework to other datasets with different susceptibility endpoint definitions tests further its effectiveness and applicability. The models require additional validations on external cohorts before being considered for medical screening research. The following additional future work emerges:

A) Further research can also be carried out on the current case studies in this thesis by predicting susceptibility to other radiotherapy side effects by varying endpoint definitions in the REQUITE data on the scale of CTCAE v4.0.

B) Also, on the visceral fat case study, further work can be carried out by altering the scale of the MRI visceral fat amounts to predict susceptibility to a specific associate disease. Also, future work could explore the effect of diet-related features in the visceral fat prediction case study.

C) Our framework conceptualises susceptibility labels' definitions into three categories, resistant, susceptible and Hyper-susceptible. However, the CTCAE v4.0 and MRI-VAT scales offer an opportunity to model the suitability to adverse events aligned to the risk scale from the British Standard Guide for Occupational Health and Safety Management System (BS8800) [397]. The standard specifies five risk levels driven by the likelihood of occurrence. These levels are *Trivial, Tolerable, Moderate, Substantial and Intolerable* (see Table 7.1).

Table 7.1. BS8800 standard guide for occupational health and safety risk matrix

		Severity		
		Slightly Harmful	**Harmful**	**Extremely Harmful**
Likelihood	**Highly Unlikely**	*Trivial Risk*	*Tolerable Risk*	*Moderate Risk*
	Unlikely	*Tolerable Risk*	*Moderate Risk*	*Substantial Risk*
	Likely	*Moderate Risk*	*Substantial Risk*	*Intolerable Risk*

Table 7.1 is a risk matrix from the BS8800 [397]. It shows a more informed classification of risk based on statistical likelihood. The likelihood ratios were not given for our projects, and the prediction requirement was label focused. If these figures are made available to define new susceptibility endpoints, further modelling tasks can be carried out to predict multiple adverse events in healthcare.

D) Our framework is designed to predict patients' susceptibility endpoints with hard-set boundaries. This setup is per design requirements. Although our models met the success criteria, we are unsure if soft boundaries, such as predicting the degree of susceptibility (possible fuzziness of membership), could offer a better solution.

list of tables

list of figures

references & bibliography

[1] Garg, A. and Mago, V., 2021. Role of machine learning in medical research: A survey. *Computer science review*, *40*, p.100370.

[2] Aldraimli, M., Nazyrova, N., Djumanov, A., Sobirov, I. and Chaussalet, T.J., 2020, October. A Comparative Machine Learning Modelling Approach for Patients' Mortality Prediction in Hospital Intensive Care Unit. In *The International Symposium on Bioinformatics and Biomedicine* (pp. 16-31). Cham: Springer International Publishing.

[3] Elish, M.C., 2018, October. The stakes of uncertainty: developing and integrating machine learning in clinical care. In *Ethnographic Praxis in Industry Conference Proceedings* (Vol. 2018, No. 1, pp. 364-380).

[4] Critical Data, M.I.T., 2016. *Secondary analysis of electronic health records* (p. 427). Springer Nature.

[5] Raghupathi, V. and Raghupathi, W., 2017, February. Preventive healthcare: A neural network analysis of behavioural habits and chronic diseases. In *Healthcare* (Vol. 5, No. 1, p. 8). MDPI.

[6] Clarke, K., Carlsen, A.K. and Grunfeld, R. (2017) *Preventive care saves lives*, *Campbell County Health*. Available at: https://www.cchwyo.org/news/2017/november/preventive-care-saves-lives/ (Accessed: 22 May 2022).

[7] Yu, C.S., Lin, Y.J., Lin, C.H., Lin, S.Y., Wu, J.L. and Chang, S.S., 2020. Development of an online health care assessment for preventive medicine: a machine learning approach. *Journal of medical Internet research*, *22*(6), p.e18585.

[8] Becker, D.M., 1988. History of preventive medicine. *Prevention in Clinical Practice*, pp.13-21.

[9] Park, C., Awadalla, A., Kohno, T. and Patel, S., 2021. Reliable and trustworthy machine learning for health using dataset shift detection. *Advances in Neural Information Processing Systems*, *34*, pp.3043-3056.

[10] Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. and Celi, L.A., 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, *2*(9), pp.e489-e492.

[11] Piccini, N., Hameed, H., Rasool, R., Mukhtar, T. and Shrivas, T., 2019. 101 ML Algorithms. [online] Data Science Blog | AI, ML, big data analytics. Available at: <https://blog.datasciencedojo.com/ machine-learning-algorithms/> [Accessed 10 November 2019].

[12] Agyapong, K.B., Hayfron-Acquah, J.B. and Asante, M., 2016. An overview of data mining models (Descriptive and predictive). International Journal of Software & Hardware Research in Engineering, 4(5), pp.53-60.

[13] Delen, D. and Demirkan, H., 2013. Data, information and analytics as services. Decision Support Systems, 359- 363.

[14] Lechevalier, D., Narayanan, A. and Rachuri, S., 2014, October. Towards a domain-specific framework for predictive analytics in manufacturing. In 2014 IEEE International Conference on Big Data (Big Data) (pp. 987-995). IEEE.

[15] Hse.gov.uk. 1999. *Managing risks and risk assessment at work – Overview -HSE.* [online] Available at: <https://www.hse.gov.uk/ simple-health-safety/risk/index.htm> [Accessed 8 December 2020]

[16] Parkin, R.T. and Balbus, J.M., 2000. Variations in concepts of "susceptibility" in risk assessment. *Risk Analysis*, *20*(5), pp.603-612.

[17] Stanford, S., Iriondo, R. and Shukla, P., 2021. Best public datasets for machine learning, data science, sentiment analysis, computer vision, NLP…. [online] Medium. Available at: <https://medium.com/towards-artificial-intelligence/best-datasets-for-machine-learning-data-science-computer-vision-nlp-ai-c9541058cf4f> [Accessed 12 February 2021].

[18] Seibold, P., Webb, A., Aguado-Barrera, M.E., Azria, D., Bourgier, C., Brengues, M., Briers, E., Bultijnck, R., Calvo-Crespo, P., Carballo, A. and Choudhury, A., 2019. REQUITE: a prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer. Radiotherapy and Oncology, 138, pp.59-67.

[19] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. and Liu, B., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. Plos med, 12(3), p.e1001779

[20] Lee, C.H. and Yoon, H.J., 2017. Medical big data: promise and challenges. Kidney research and clinical practice, 36(1), p.3.

[21] Floca, R., 2014. Challenges of open data in medical research. In Opening Science (pp. 297-307). Springer, Cham.

[22] Lee, K., Weiskopf, N. and Pathak, J., 2017. A framework for data quality assessment in clinical research datasets. In AMIA Annual Symposium Proceedings (Vol. 2017, p. 1080). American Medical Informatics Association.

[23] Smerek, M.M., 2015. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0).

[24] Arne, G.N., 2009. Breast cancer risk assessments to barrier contraception exposure. A new approach. Makedonska Akademija na

Naukite i Umetnostite Oddelenie Za Bioloshki i Meditsinski Nauki Prilozi, 30(1), pp.217-232.

[25] Tyrer, J., Duffy, S.W. and Cuzick, J., 2004. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine, 23(7), pp.1111-1130.

[26] Barlow, W.E., White, E., Ballard-Barbash, R., Vacek, P.M., Titus-Ernstoff, L., Carney, P.A., Tice, J.A., Buist, D.S., Geller, B.M., Rosenberg, R. and Yankaskas, B.C., 2006. Prospective breast cancer risk prediction model for women undergoing screening mammography. Journal of the National Cancer Institute, 98(17), pp.1204-1214.

[27] Boyle, P., Mezzetti, M., La Vecchia, C., Franceschi, S., Decarli, A. and Robertson, C., 2004. Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. European journal of cancer prevention, 13(3), pp.183-191.

[28] Chen, J., Pee, D., Ayyagari, R., Graubard, B., Schairer, C., Byrne, C., Benichou, J. and Gail, M.H., 2006. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. Journal of the National Cancer Institute, 98(17), pp.1215-1226.

[29] Colditz, GA and Rosner, B., 2000. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. American journal of epidemiology, 152(10), pp.950-964.

[30] Costantino, J.P., Gail, M.H., Pee, D., Anderson, S., Redmond, C.K., Benichou, J. and Wieand, H.S., 1999. Validation studies for models projecting the risk of invasive and total breast cancer incidence. Journal of the National Cancer Institute, 91(18), pp.1541-1548.

[31] Tice, J.A., Cummings, S.R., Smith-Bindman, R., Ichikawa, L., Barlow, W.E. and Kerlikowske, K., 2008. Using clinical factors and mammographic breast density to estimate breast cancer risk:

development and validation of a new predictive model. Annals of internal medicine, 148(5), pp.337-347.

[32] Stevens, S.S., 1946. On the theory of scales of measurement.

[33] Velleman, PF and Wilkinson, L., 1993. Nominal, ordinal, interval, and ratio typologies are misleading. The American Statistician, 47(1), pp.65-72.

[34] Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. (pp. 10 & 504). Springer Science & Business Media.

[35] Rumsey, D.J., 2016. Statistics for dummies. (pp. 95-107) John Wiley & Sons.

[36] Pathak, M., 2020. Handling Categorical Data in Python. [online] Data Camp. Available at: <https://www.datacamp.com/community/tutorials/categorical-data> [Accessed 27 July 2020].

[37] Luce, R.D., 1997. Quantification and symmetry: Commentary on Michell, Quantitative science and the definition of measurement in psychology. British journal of Psychology, 88(3), pp.395-398.

[38] Joshi, A., Kale, S., Chandel, S. and Pal, D.K., 2015. Likert scale: Explored and explained. Current Journal of Applied Science and Technology, pp.396-403.

[39] Inouye, D.I., Yang, E., Allen, G.I. and Ravikumar, P., 2017. A review of multivariate distributions for count data derived from the Poisson distribution. Wiley Interdisciplinary Reviews: Computational Statistics, 9(3), p.e1398.

[40] Mosteller, F. and Tukey, JW, 1977. Data analysis and regression: a second course in statistics.

[41] Chrisman, N.R., 1998. Rethinking levels of measurement for cartography. Cartography and Geographic Information Systems, 25(4), pp.231-242.

[42] Heckert, N.A., Filliben, J.J., Croarkin, C.M., Hembree, B., Guthrie, W.F., Tobias, P. and Prinz, J., 2002. Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods.

[43] Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), pp.559-572.

[44] i2tutorials. 2019. *What are the Pros and cons of the PCA? | i2tutorials*. [online] Available at: https://www.i2tutorials.com /what-are-the-pros-and-cons-of-the-pca/ [Accessed 28 November 2021].

[45] Shlens, J., 2014. A tutorial on independent component analysis. *arXiv preprint arXiv:1404.2986*.

[46] Kruskal, J.B., 1969, January. Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". In *Statistical computation* (pp. 427-440). Academic Press.

[47] Friedman, J.H. and Tukey, J.W., 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, *100*(9), pp.881-890

[48] Faith, J., 2007, July. Targeted projection pursuit for interactive exploration of high-dimensional data sets. In *2007 11th International Conference Information Visualization (IV'07)* (pp. 286-292). IEEE.

[49] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 2000. CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9, p.13.

[50] De Jonge, E. and Van Der Loo, M., 2013. An introduction to data cleaning with R. Heerlen: Statistics Netherlands.

[51] Misra, S., Li, H. and He, J., 2019. Machine learning for subsurface characterisation. (pp Pages 129-155). Gulf Professional Publishing.

[52] Weiss, G.M. and Provost, F., 2001. The effect of class distribution on classifier learning: an empirical study.

[53] D. Pyle, S. Editor, and D. D. Cerra, Data Preparation for Data Mining. Morgan Kaufmann Publishers, 1999.

[54] Burdack, J., Horst, F., Giesselbach, S., Hassan, I., Daffner, S. and Schöllhorn, W.I., 2020. Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning. Frontiers in bioengineering and biotechnology, 8, p.260.

[55] Chandrasekar, P., Qian, K., Shahriar, H. and Bhattacharya, P., 2017, July. Improving the prediction accuracy of decision tree mining with data pre-processing. In 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC) (Vol. 2, pp. 481-484). IEEE.

[56] Srivastava, S., 2014. Weka: a tool for data pre-processing, classification, ensemble, clustering and association rule mining. International Journal of Computer Applications, 88(10).

[57] Weka.sourceforge.io.2020.weka.filters.unsupervised.attribute (weka-dev 3.9.5 API). [online] Available at:<https://weka. sourceforge.io/doc.dev/weka/filters/unsupervised/attribute/package-summary.html> [Accessed 20 September 2020].

[58] Weka.sourceforge.io.2020.weka.filters.supervised.instance (weka-dev 3.9.5 API). [online] Available at:< https://weka. sourceforge.io/doc.dev/weka/filters/package-summary.html> [Accessed 18 September 2020].

[59] Suresh, K.P., 2011. An overview of randomisation techniques: an unbiased assessment of outcome in clinical research. Journal of human reproductive sciences, 4(1), p.8.

[60] Weka.sourceforge.io.2019. RemoveUseless (weka-dev 3.9.5 API). [online] Available at: <https://weka.sourceforge.io/doc.dev/weka/ filters/unsupervised/ attribute/ RemoveUseless.html> [Accessed 1 June 2019].

[61] DeepAI. 2020. Binarization. [online] Available at: <https://deepai.org/machine-learning-glossary-and-terms/binarization> [Accessed 9 April 2020].

[62] C. P. Nick Dingwall, "Are categorical variables getting lost in your random forests?," 2016. [Online]. Available: https://roamanalytics.com/2016/10/28/are-categorical-variables-getting-lost-in-your-random-forests/. [Accessed: 30-Jan-2019]

[63] Soley-Bori, M., 2013. Dealing with missing data: Key assumptions and methods for applied analysis. Boston University, 23.

[64] Biobank.ndph.ox.ac.uk. 2017. UK-Biobank : Resource catalogue. [online] Available at: <https://biobank.ndph.ox.ac.uk/showcase/ docs.cgi?id=0> [Accessed 6 January 2018].

[65] Azuaje, F., 2011. Witten, I, Frank, E: Data mining: Practical machine learning tools and techniques 3rd edition (pp. 58-60).

[66] Allison, P.D., 2001. Missing data (p. 4). Sage publications.

[67] Humphries, M., 2013. Missing Data & How to Deal: An overview of missing data. Population Research Center. University of Texas. Recuperado (pp.39-41).

[68] Schafer, J.L. and Graham, J.W., 2002. Missing data: our view of the state of the art. Psychological methods, 7(2), p.147.

[69] Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In Annals of economic and social measurement, volume 5, number 4 (pp. 475-492). NBER.

[70] Koné, S., Bonfoh, B., Dao, D., Koné, I. and Fink, G., 2019. Heckman-type selection models to obtain unbiased estimates with missing measures outcome: theoretical considerations and an application to missing birth weight data. BMC medical research methodology, 19(1), pp.1-13.

[71] Yan, T. and Curtin, R., 2010. The relation between unit nonresponse and item nonresponse: A response continuum perspective. International Journal of Public Opinion Research, 22(4), pp.535-551.

[72] Groves, R.M., 2004. Survey errors and survey costs (Vol. 536). John Wiley & Sons.

[73] Groves, R.M., Cialdini, R.B. and Couper, M.P., 1992. Understanding the decision to participate in a survey. Public opinion quarterly, 56(4), pp.475-495.

[74] Pérez-Duarte, S., Sánchez-Muñoz, C. and Törmälehto, V.M., 2010, May. Re-weighting to reduce unit non-response bias in household wealth surveys: a cross-country comparative perspective illustrated by a case study. In *European Conference on Quality in Official Statistics*.

[75] Lynn, P., 2005. [online] Restore.ac.uk. Available at: <https://www.restore.ac.uk/Longitudinal/surveynetwork/documents/EC9692005Weighting.pdf> [Accessed 20 February 2018].

[76] Little, RJ and Rubin, D.B., 2019. Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

[77] Van Buuren, S., 2018. Flexible imputation of missing data. CRC press. Available at: <https://stefvanbuuren.name/fimd/> [Accessed 12 Septemeber 2018].

[78] Gelman, A. and Hill, J., 2006. Data analysis using regression and multilevel/hierarchical models. (pp. 531-533). Cambridge university press.

[79] Critical Data, MIT, 2016. Secondary analysis of electronic health records (pp.146-153). Springer Nature.

[80] Gleason, T.C. and Staelin, R., 1975. A proposal for handling missing data. Psychometrika, 40(2), pp.229-252.

[81] Quinlan, J.R., 1986. Induction of decision trees. Machine learning, 1(1), pp.81-106.

[82] Ghahramani, Z. and Jordan, M.I., 1994. Supervised learning from incomplete data via an EM approach. In Advances in neural information processing systems (pp. 120-127).

[83] Smola, A.J., Vishwanathan, SVN and Hofmann, T., 2005, March. Kernel methods for missing variables. In AISTATS (Vol. 261, p. 261).

[84] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[85] Jakobsen, J.C., Gluud, C., Winkel, P., Lange, T. and Wetterslev, J., 2014. The thresholds for statistical and clinical significance–a five-step procedure for evaluation of intervention effects in randomised clinical trials. BMC medical research methodology, 14(1), pp.1-12.

[86] Groenwold, R.H., Moons, K.G. and Vandenbroucke, J.P., 2014. Randomised trials with missing outcome data: how to analyse and what to report. Cmaj, 186(15), pp.1153-1157.

[87] Sterner, J. A., White, I. R., Carlin, J. B., Spratt, M. and Royston, P., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. British medical journal (International ed.), 339(7713), pp.157-160.

[88] Dziura, J.D., Post, L.A., Zhao, Q., Fu, Z. and Peduzzi, P., 2013. Strategies for dealing with missing data in clinical trials: from design to analysis. The Yale journal of biology and medicine, 86(3), p.343.

[89] Clark, T.G. and Altman, D.G., 2003. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. Journal of clinical epidemiology, 56(1), pp.28-37.

[90] SAS/STAT User's Guide, 14.1. NC 27513-2414: SAS Institute Inc, 2015.

[91] Soley-Bori, M., 2013. Dealing with missing data: Key assumptions and methods for applied analysis. Boston University, 23.

[92] Ahmed K, et al., 2009. Applying Missing Data Imputation Methods to HOS Household Income Data. Prepared by the National Committee

for Quality Assurance (NCQA) for the Centers for Medicare and Medicaid Services.

[93] Flyer, P. and Hirman, J., 2009. Missing data in confirmatory clinical trials. Journal of biopharmaceutical statistics, 19(6), pp.969-979.

[94] Medium. 2020. Power of Missing data imputation in R. [online] Available at: <https://medium.com/analytics-vidhya/power-of-missing-data-imputation-in-r-78b31aa2029d> [Accessed 6 October 2020].

[95] Scheffer, J., 2002. Dealing with missing data, Research Letters in the Information and Mathematical Sciences, 3, 153-160

[96] Sulis, I. and Porcu, M., 2008. Assessing the effectiveness of a stochastic regression imputation method for ordered categorical data. Workimg Paper, 4.

[97] Hilbe, J.M., 2009. Logistic regression models. CRC press.

[98] Enders, C.K., 2010. Applied missing data analysis. (pp. 46-49) Guilford press.

[99] Zhang, S., 2012. Nearest neighbour selection for iteratively kNN imputation. Journal of Systems and Software, 85(11), pp.2541-2552.

[100] Hassanat, A.B., Abbadi, M.A., Altarawneh, G.A. and Alhasanat, A.A., 2014. Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. arXiv preprint arXiv:1409.0919.

[101] R. J. Rossi, Mathematical Statistics: An Introduction to Likelihood Based Inference. New York: John Wiley & Sons, 2018.

[102] Chan, F., 2016. Probability Distributions and their Mass/Density Functions. [online] Fong Chun Chan's Blog. Available at: <https://tinyheero.github.io/2016/03/17/prob-distr. html> [Accessed 14 November 2017].

[103] D. J. C. MacKay, Information theory, inference, and learning algorithms. Cambridge University Press, 2003.

[104] J. Kuha, “AIC and BIC,” Sociol. Methods Res., vol. 33, no. 2, pp. 188–229, Nov. 2004.

[105] The Prevention and Treatment of Missing Data in Clinical Trials. Washington, D.C.: National Academies Press, 2010.

[106] Russell, S. and Norvig, P., 2009. Artificial Intelligence: A Modern Approach (3rd edition.). Essex CM20 2JE: Pearson Education Limited, p.816.

[107] Gupta, M.R. and Chen, Y., 2011. Theory and Use of the EM Algorithm. Foundations and Trends® in Signal Processing, 4(3), pp.223-296.

[108] Rahman, G. and Islam, Z., 2011, December. A decision tree-based missing value imputation technique for data pre-processing. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 41-50).

[109] Little, R.J. and Rubin, D.B., 2002. Missing data in experiments. Statistical analysis with missing data, pp.24-40.

[110] Rubin, D.B., 2004. Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley & Sons.

[111] Grace-Martin, K., 2019. Multiple Imputation in a Nutshell - The Analysis Factor. [online] The Analysis Factor. Available at:<https://www.theanalysisfactor.com/ multiple-imputation-in-a-nutshell/> [Accessed 3 November 2019].

[112] Jakobsen, J., Gluud, C., Wetterslev, J. and Winkel, P., 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology, 17(162).

[113] Dong, Y. and Peng, C.Y.J., 2013. Principled missing data methods for researchers. SpringerPlus, 2(1), pp.1-17.

[114] Scheffer, J., 2002. Dealing with missing data. Research Letters in the Information and Mathematical Sciences, 3, 153–160.

[115] Weiss, G.M. and Provost, F., 2001. The effect of class distribution on classifier learning.

[116] Collins, L.M., Schafer, J.L. and Kam, C.M., 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological methods, 6(4), p.330.

[117] McMullen, N. and Ochoa, D., 2016. Non-Monotonic Transformations of Random Variables.

[118] Patro, S. and Sahu, K.K., 2015. Normalization: A pre-processing stage. arXiv preprint arXiv:1503.06462.

[119] Kumar, D.A. and Venugopalan, S.R., 2017. The Effect of Normalization on Intrusion Detection Classifiers (Naïve Bayes and J48). International Journal on Future Revolution in Computer Science & Communication Engineering, 3(7), pp.60-64.

[120] Sarker, I.H., Alqahtani, H., Alsolami, F., Khan, A.I., Abushark, Y.B. and Siddiqui, M.K., 2020. Context pre-modeling: an empirical analysis for classification-based user-centric context-aware predictive modeling. Journal of Big Data, 7(1), pp.1-23.

[121] Dernoncourt, F., 2016. Which algorithms need feature scaling, beside from SVM? [online] Cross Validated. Available at: <https://stats.stackexchange.com/questions/244507/what-algorithms-need-feature-scaling-beside-from-svm> [Accessed 27 March 2018].

[122] Data Preparation and Feature Engineering for Machine Learning. 2021. Normalization. [online] Available at: <https://developers.google.com/machine-learning/data-prep/transform/normalization> [Accessed 15 July 2021].

[123] Moreno, A., 2020. The complete guide to clean datasets. [online] Medium. Available at: <https://towardsdatascience.com/data-normalization-with-pandas-and-scikit-learn-7c1cc6ed6475> [Accessed 29 November 2020].

[124] Ahsan, M.M., Mahmud, M.A., Saha, P.K., Gupta, K.D. and Siddique, Z., 2021. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3), p.52.

[125] Huilgol, P., 2020. 9 Feature Transformation & Scaling Techniques | Boost Model Performance. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/07/types-of-feature-transformation-and-scaling/ [Accessed 30 December 2020].

[126] Yang, L., Jin, R., Sukthankar, R. and Liu, Y., 2006, July. An efficient algorithm for local distance metric learning. In *AAAI* (Vol. 2, pp. 543-548).

[127] Jain, S., Shukla, S. and Wadhvani, R., 2018. Dynamic selection of normalization techniques using data complexity measures. Expert Systems with Applications, 106, pp.252-262.

[128] Ho, T.K. and Basu, M., 2000, September. Measuring the complexity of classification problems. In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000 (Vol. 2, pp. 43-47). IEEE.

[129] Lorena, A.C. and de Souto, M.C., 2015, November. On measuring the complexity of classification problems. In International Conference on Neural Information Processing (pp. 158-167). Springer, Cham.

[130] Ho, T.K. and Basu, M., 2002. Complexity measures of supervised classification problems. IEEE transactions on pattern analysis and machine intelligence, 24(3), pp.289-300.

[131] Vellido, A., 2019. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Computing and Applications, pp.1-15.

[132] Niu, Z., Shi, S., Sun, J. and He, X., 2011, September. A survey of outlier detection methodologies and their applications. In International Conference on Artificial Intelligence and Computational Intelligence (pp. 380-387). Springer, Berlin, Heidelberg.

[133] Serneels, S., De Nolf, E. and Van Espen, P.J., 2006. Spatial sign preprocessing: a simple way to impart moderate robustness to multivariate estimators. Journal of Chemical Information and Modeling, 46(3), pp.1402-1409.

[134] Gupta, S., Agrawal, A., Gopalakrishnan, K. and Narayanan, P., 2015, June. Deep learning with limited numerical precision. In International conference on machine learning (pp. 1737-1746). PMLR.

[135] Kulisch, U., 1975, November. Mathematical foundation of computer arithmetic. In 1975 IEEE 3rd Symposium on Computer Arithmetic (ARITH) (pp. 1-13). IEEE.

[136] Robben, S., Velikova, M., Lucas, P.J. and Samulski, M., 2010, December. Discretisation does affect the performance of Bayesian networks. In International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 237-250). Springer, London.

[137] Rajbahadur, G.K., Wang, S., Kamei, Y. and Hassan, A.E., 2019. Impact of discretization noise of the dependent variable on machine learning classifiers in software engineering. IEEE Transactions on Software Engineering.

[138] Yang, Y. and Webb, G.I., 2002, August. A comparative study of discretization methods for naive-bayes classifiers. In Proceedings of PKAW (Vol. 2002).

[139] Z. Marzuki and F. Ahmad, "Data Mining Discretization Methods and Performances," Electr. Eng. Informatics, vol. 19, no. 17, pp. 535–537, 2007.

[140] Dougherty, J., Kohavi, R. and Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In Machine learning proceedings 1995 (pp. 194-202). Morgan Kaufmann.

[141] Ian, H.W. and Eibe, F., 2005. Data Mining: Practical machine learning tools and techniques. 3rd edition, pp. 349-356.

[142] Yang, Y. and Webb, G.I., 2001, September. Proportional k-interval discretization for naive-Bayes classifiers. In European Conference on Machine Learning (pp. 564-575). Springer, Berlin, Heidelberg.

[143] Yang, Y. and Webb, G.I., 2003, April. Weighted proportional k-interval discretization for naive-bayes classifiers. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 501-512). Springer, Berlin, Heidelberg.

[144] Nevill-Manning, C.G., Holmes, G. and Witten, I.H., 1995, November. The development of Holte's 1R classifier. In Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems (pp. 239-242). IEEE.

[145] Fayyad, U. and Irani, K., 1993. Multi-interval discretization of continuous-valued attributes for classification learning.

[146] He, H. and Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), pp.1263-1284.

[147] Akosa, J., 2017, April. Predictive accuracy: A misleading performance measure for highly imbalanced data. In Proceedings of the SAS Global Forum (Vol. 12).

[148] He, H. and Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), pp.1263-1284.

[149] Ling, C.X. and Li, C., 1998, August. Data mining for direct marketing: Problems and solutions. In Kdd (Vol. 98, pp. 73-79).

[150] Elhassan, T. and Aljurf, M., 2017. Classification of imbalanced data using Tomek link (T-link) combined with random under-sampling (RUS) as a data reduction method. Global J Technol Optim S, 1.

[151] Aldraimli, M., Soria, D., Grishchuck, D., Ingram, S., Lyon, R., Mistry, A., Oliveira, J., Samuel, R., Shelley, L.E., Osman, S. and Dwek, M.V., 2021. A data science approach for early-stage prediction of patient's

susceptibility to acute side effects of advanced radiotherapy. Computers in Biology and Medicine, p.104624.

[152] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.

[153] Woods, K.S., Solka, J.L., Priebe, C.E., Kegelmeyer Jr, W.P., Doss, C.C. and Bowyer, K.W., 1994. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. In State of The Art in Digital Mammographic Image Analysis (pp. 213-231).

[154] Wang, B.X. and Japkowicz, N., 2004, June. Imbalanced data set learning with synthetic samples. In Proc. IRIS Machine Learning Workshop (Vol. 19, p. 435). sn.

[155] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.

[156] Han, H., Wang, W.Y. and Mao, B.H., 2005, August. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Springer, Berlin, Heidelberg.

[157] Torres, F.R., Carrasco-Ochoa, J.A. and Martínez-Trinidad, J.F., 2016, June. SMOTE-D a deterministic version of SMOTE. In Mexican Conference on Pattern Recognition (pp. 177-188). Springer, Cham.

[158] Wang, Q., Luo, Z., Huang, J., Feng, Y. and Liu, Z., 2017. A novel ensemble method for imbalanced data learning: bagging of extrapolation SMOTE SVM. Computational intelligence and neuroscience, 2017.

[159] Santos, M.S., Abreu, P.H., García-Laencina, P.J., Simão, A. and Carvalho, A., 2015. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of biomedical informatics, 58, pp.49-59.

[160] Ma, L. and Fan, S., 2017. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC bioinformatics, 18(1), pp.1-18.

[161] He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, June. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.

[162] Zhang, Y.P., Zhang, L.N. and Wang, Y.C., 2010, September. Cluster-based majority under-sampling approaches for class imbalance learning. In 2010 2nd IEEE International Conference on Information and Financial Engineering (pp. 400-404). IEEE.

[163] Santos, M.S., Abreu, P.H., García-Laencina, P.J., Simão, A. and Carvalho, A., 2015. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of biomedical informatics, 58, pp.49-59.

[164] Tomek, I., 1976. Two modifications of CNN.

[165] Kubat, M. and Matwin, S., 1997, July. Addressing the curse of imbalanced training sets: one-sided selection. In Icml (Vol. 97, pp. 179-186).

[166] Maheshwari, S., Agrawal, J. and Sharma, S., 2011. A new approach for classification of highly imbalanced datasets using evolutionary algorithms. Int. J. Sci. Eng. Res, 2(7), pp.1-5.

[167] Hart, P., 1968. The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory, 14(3), pp.515-516.

[168] Wilson, D.L., 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, (3), pp.408-421.

[169] Tomek, I., 1976. An experiment with the edited nearest-neighbor rule.

[170] Mani, I. and Zhang, I., 2003, August. kNN approach to unbalanced data distributions: a case study involving information extraction. In

Proceedings of workshop on learning from imbalanced datasets (Vol. 126). United States: ICML.

[171] Laurikkala, J., 2001, July. Improving identification of difficult small classes by balancing class distribution. In Conference on Artificial Intelligence in Medicine in Europe (pp. 63-66). Springer, Berlin, Heidelberg.

[172] Arruda, J.L., Prudêncio, R.B. and Lorena, A.C., 2020, October. Measuring Instance Hardness Using Data Complexity Measures. In Brazilian Conference on Intelligent Systems (pp. 483-497). Springer, Cham.

[173] Smith, Michael Reed. "An empirical study of instance hardness." (2009).

[174] Anand, A., Pugalenthi, G., Fogel, G.B. and Suganthan, P.N., 2010. An approach for classification of highly imbalanced data using weighting and undersampling. Amino acids, 39(5), pp.1385-1391.

[175] Li, Q., Wang, Y. and Bryant, S.H., 2009. A novel method for mining highly imbalanced high-throughput screening data in PubChem. Bioinformatics, 25(24), pp.3310-3316.

[176] Liu, X.Y. and Zhou, Z.H., 2006, December. The influence of class imbalance on cost-sensitive learning: An empirical study. In Sixth International Conference on Data Mining (ICDM'06) (pp. 970-974). IEEE.

[177] McCarthy, K., Zabar, B. and Weiss, G., 2005, August. Does cost-sensitive learning beat sampling for classifying rare classes?. In Proceedings of the 1st international workshop on Utility-based data mining (pp. 69-77).

[178] Cieslak, D.A., 2009. Finding problems in, proposing solutions to, and performing analysis on imbalanced data. University of Notre Dame.

[179] Quinlan, J.R., 1996, August. Bagging, boosting, and C4. 5. In Aaai/iaai, Vol. 1 (pp. 725-730).

[180] Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F. and Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, pp.1-5.

[181] Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J. and Van Hoecke, S., 2021. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. Artificial Intelligence in Medicine, 111, p.101987.

[182] Azuaje, F., 2011. Witten, I, Frank, E: Data mining: Practical machine learning tools and techniques 3rd edition (pp. 307-308).

[183] Thrun, S.B., Bala, J.W., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., De Jong, K.A., Dzeroski, S., Fisher, D.H., Fahlman, S.E. and Hamann, R., 1991. The monk's problems: A performance comparison of different learning algorithms.

[184] Aha, D.W., 1992. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of Man-Machine Studies, 36(2), pp.267-287.

[185] Langley, P., Iba, W. and Thompson, K., 1992, July. An analysis of Bayesian classifiers. In Aaai (Vol. 90, pp. 223-228).

[186] Almuallim, H. and Dietterich, T.G., 1991, July. Learning With Many Irrelevant Features. In AAAI (Vol. 91, pp. 547-552).

[187] Gennari, J.H., Langley, P. and Fisher, D., 1989. Models of incremental concept formation. Artificial intelligence, 40(1-3), pp.11-61.

[188] Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. Artificial intelligence, 97(1-2), pp.273-324.

[189] Khaire, U.M. and Dhanalakshmi, R., 2019. Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences.

[190] Azuaje, F., 2011. Witten, I, Frank, E: Data mining: Practical machine learning tools and techniques 3rd edition (pp. 308-313)

[191] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), pp.1157-1182.

[192] Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), pp.16-28.

[193] Torkkola, K., 2003. Feature extraction by non-parametric mutual information maximization. *Journal of machine learning research*, *3*(Mar), pp.1415-1438.

[194] Quinlan, J.R., 1986. Induction of decision trees. *Machine learning*, *1*(1), pp.81-106.

[195] Harris, E., 2002, January. Information Gain Versus Gain Ratio: A Study of Split Method Biases. In *ISAIM*.

[196] Salzberg, S.L., 1994. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.

[197] Raileanu, L.E. and Stoffel, K., 2004. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), pp.77-93.

[198] Hall, M.A., 1998. Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

[199] Liu, H. and Setiono, R., 1996, July. A probabilistic approach to feature selection-a filter solution. In *ICML* (Vol. 96, pp. 319-327).

[200] Kitter, J., 1978. Feature set search algorithm. *Pattern Recognition and Signal Processing, Sithoff and Noordhoff, Alphen ann Den Riju, The Netherlands, pp41-60.*

[201] Kohavi, R. and John, G.H., 1998. The wrapper approach. In *Feature extraction, construction and selection* (pp. 33-50). Springer, Boston, MA.

[202] Vafaie, H. and De Jong, K.A., 1992, November. Genetic Algorithms as a Tool for Feature Selection in Machine Learning. In *ICTAI* (pp. 200-203).

[203] Peng, H., Long, F. and Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence, 27(8), pp.1226-1238.*

[204] Mundra, P.A. and Rajapakse, J.C., 2009. SVM-RFE with MRMR filter for gene selection. *IEEE transactions on nanobioscience, 9(1), pp.31-37.*

[205] Booker, L.B., Goldberg, D.E. and Holland, J.H., 1989. Classifier systems and genetic algorithms. *Artificial intelligence*, *40*(1-3), pp.235-282.

[206] Salin, E.D. and Winston, P.H., 1992. Machine learning and artificial intelligence: an introduction. *Analytical chemistry (Washington, DC)*, *64*(1), pp.49A-60A.

[207] John, G.H., Kohavi, R. and Pfleger, K., 1994. Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994* (pp. 121-129). Morgan Kaufmann.

[208] Almuallim, H. and Dietterich, T.G., 1991, July. Learning With Many Irrelevant Features. In *AAAI* (Vol. 91, pp. 547-552).

[209] Kibler, D. and Aha, D.W., 1987, January. Learning representative exemplars of concepts: An initial case study. In *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 24-30). Morgan Kaufmann.

[210] Cardie, C., 1993. Using decision trees to improve case-based learning. In *Proceedings of the tenth international conference on machine learning* (pp. 25-32).

[211] Holmes, G. and Nevill-Manning, C.G., 1995. Feature selection via the discovery of simple classification rules.

[212] Kira, K. and Rendell, L.A., 1992. A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249-256). Morgan Kaufmann.

[213] Gilad-Bachrach, R., Navot, A. and Tishby, N., 2004, July. Margin based feature selection-theory and algorithms. In *Proceedings of the twenty-first international conference on Machine learning* (p. 43).

[214] Hall, M.A., 2000. Correlation-based feature selection of discrete and numeric class machine learning.

[215] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), pp.389-422.

[216] Gutlein, M., Frank, E., Hall, M. and Karwath, A., 2009, March. Large-scale attribute selection using wrappers. In *2009 IEEE symposium on computational intelligence and data mining* (pp. 332-339). IEEE.

[217] Liu, M., Xu, C., Luo, Y., Xu, C., Wen, Y. and Tao, D., 2017. Cost-sensitive feature selection by optimizing F-measures. *IEEE Transactions on Image Processing*, *27*(3), pp.1323-1335.

[218] Beer, A. and Mohacsi, S., 2008, April. Efficient test data generation for variables with complex dependencies. In *2008 1st International Conference on Software Testing, Verification, and Validation* (pp. 3-11). IEEE.

[219] Bhat, A. and Quadri, S.M.K., 2015, March. Equivalence class partitioning and boundary value analysis-a review. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1557-1562). IEEE.

[220] Arnicane, V., 2009. Complexity of equivalence class and boundary value testing methods. *International Journal of Computer Science and Information Technology, 751*, pp.80-101.

[221] Shtatland, E.S., Kleinman, K. and Cain, E.M., 2005. Model building in PROC PHREG with automatic variable selection and information

criteria. *Philadelphia, PA, SAS Users Group International Paper*, pp.206-30.

[222] Vittinghoff, E. and McCulloch, C.E., 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. *American journal of epidemiology*, *165*(6), pp.710-718.

[223] Van Smeden, M., de Groot, J.A., Moons, K.G., Collins, G.S., Altman, D.G., Eijkemans, M.J. and Reitsma, J.B., 2016. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, *16*(1), pp.1-12.

[224] van Smeden, M., Moons, K.G., de Groot, J.A., Collins, G.S., Altman, D.G., Eijkemans, M.J. and Reitsma, J.B., 2019. Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical methods in medical research*, *28*(8), pp.2455-2474.

[225] Foley, D., 1972. Considerations of sample and feature size. *IEEE Transactions on Information Theory*, *18*(5), pp.618-626.

[226] Sevey, R., 2017. *How Much Data is Needed to Train a (Good) Model?* [online] Data Robot AI Cloud. Available at: <https://www.datarobot.com/blog/how-much-data-is-needed-to-train-a-good-model/> [Accessed 9 March 2019].

[227] Somorjai, R.L., Dolenko, B. and Baumgartner, R., 2003. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, *19*(12), pp.1484-1491.

[228] Sechidis, K., Calvo, B. and Brown, G., 2014, September. Statistical hypothesis testing in positive unlabelled data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 66-81). Springer, Berlin, Heidelberg.

[229] Ripley, B.D., 2007. *Pattern recognition and neural networks*. Cambridge university press, p.354.

[230] Brownlee, J., 2017. *What is the Difference Between Test and Validation Datasets?*. [online] Machine Learning Mastery. Available

at: <https://machinelearningmastery. com/difference-test-validation-datasets/ > [Accessed 10 September 2018].

[231] Guyon, I., 1997. A scaling law for the validation-set training-set size ratio. *AT&T Bell Laboratories*, *1*(11).

[232] Box, G.E. and Meyer, R.D., 1986. An analysis for unreplicated fractional factorials. *Technometrics*, *28*(1), pp.11-18.

[233] Newman, M.E., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, *46*(5), pp.323-351.

[234] Mezaal, M.R., Pradhan, B. and Rizeei, H.M., 2018. Improving landslide detection from airborne laser scanning data using optimized Dempster–Shafer. *Remote Sensing*, *10*(7), p.1029.

[235] Pawluszek-Filipiak, K. and Borkowski, A., 2020. On the importance of train–test split ratio of datasets in automatic landslide detection by supervised classification. *Remote Sensing*, *12*(18), p.3054.

[236] Aguilera, P.A., Fernández, A., Reche, F. and Rumí, R., 2010. Hybrid Bayesian network classifiers: Application to species distribution models. *Environmental Modelling & Software*, *25*(12), pp.1630-1639.

[237] Albert, A. and Zhang, L., 2010. A novel definition of the multivariate coefficient of variation. *Biometrical Journal*, *52*(5), pp.667-675.

[238] Zhao, Y., 2013. Machine learning algorithms for predicting roadside fine particulate matter concentration level in Hong Kong Central. *Computational Ecology and Software*, *3*(3), p.61.

[239] Marcot, B.G. and Hanea, A.M., 2021. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?. *Computational Statistics*, *36*(3), pp.2009-2031.

[240] Schneider, J. and Moore, A., 1997. *A Locally Weighted Learning Tutorial*. [online] Cs.cmu.edu. Available at: <https://www.cs.cmu .edu/~schneide/tut5/tut5.html> [Accessed 16 June 2018].

[241] Krstajic, D., Buturovic, L.J., Leahy, D.E. and Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, *6*(1), pp.1-15.

[242] Cawley, G.C. and Talbot, N.L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, pp.2079-2107.

[243] Brownlee, J., 2020. *Nested Cross-Validation for Machine Learning with Python*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/> [Accessed 10 November 2020].

[244] Shi, L., Campbell, G., Jones, W.D., Campagne, F., Wen, Z., Walker, S.J., Su, Z., Chu, T.M., Goodsaid, F.M., Pusztai, L. and Shaughnessy Jr, J.D., 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, *28*(8), pp.827-838.

[245] Kotlar, A.M., Iversen, B.V. and de Jong van Lier, Q., 2019. Evaluation of parametric and nonparametric machine-learning techniques for prediction of saturated and near-saturated hydraulic conductivity. *Vadose Zone Journal*, *18*(1), pp.1-13.

[246] Khire, S., Ganorkar, P., Apastamb, A. and Panicker, S., 2021. Investigating the Impact of Data Analysis and Classification on Parametric and Nonparametric Machine Learning Techniques: A Proof of Concept. In *Computer Networks and Inventive Communication Technologies* (pp. 211-227). Springer, Singapore.

[247] Brownlee, J., 2019. *A Tour of Machine Learning Algorithms*. [online] Machine Learning Mastery. Available at: <https:// machinelearningmastery.com/a-tour-of-machine-learning-algorithms/> [Accessed 5 December 2019].

[248] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. Wiley, 2001.

[249] Wolpert, D.H. and Macready, W.G., 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), pp.67-82.

[250] Bhuvaneswari, R. and Kalaiselvi, K., 2012. Naive Bayesian classification approach in healthcare applications. *International Journal of Computer Science and Telecommunications*, *3*(1), pp.106-112.

[251] Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), pp.43-48.

[252] Al-Aidaroos, K.M., Bakar, A.A. and Othman, Z., 2012. Medical data classification with Naive Bayes approach. *Information Technology Journal*, *11*(9), p.1166.

[253] Schober, P. and Vetter, T.R., 2021. Logistic regression in medical research. *Anesthesia and analgesia*, *132*(2), p.365.

[254] Le Cessie, S. and Van Houwelingen, J.C., 1992. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *41*(1), pp.191-201.

[255] Amato, F., López, A., Peña-Méndez, E.M., Vaňhara, P., Hampl, A. and Havel, J., 2013. Artificial neural networks in medical diagnosis. *Journal of applied biomedicine*, *11*(2), pp.47-58.

[256] Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), p.386.

[257] Nwankpa, C., Ijomah, W., Gachagan, A. and Marshall, S., 2018. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378.*

[258] Abu Alfeilat, H.A., Hassanat, A.B., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B., Eyal Salman, H.S. and Prasath, V.S., 2019. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data*, *7*(4), pp.221-248.

[259] Hassanat, A.B., Abbadi, M.A., Altarawneh, G.A. and Alhasanat, A.A., 2014. Solving the problem of the K parameter in the KNN classifier

using an ensemble learning approach. *arXiv preprint arXiv:1409.0919.*

[260] Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003, November. KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.

[261] Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, "Iknn: Informative k-nearest neighbor pattern classification," in Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, 2007, pp. 248-264.

[262] Hamamoto, Y., Uchimura, S. and Tomita, S., 1997. A bootstrap technique for nearest neighbor classifier design. *IEEE transactions on pattern analysis and Machine intelligence*, *19*(1), pp.73-79.

[263] Enas, G.G. and Choi, S.C., 1986. Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. In *Statistical Methods of Discrimination and Classification* (pp. 235-244). Pergamon.

[264] M. Jirina and M. J. Jirina, "Classifiers Based on Inverted Distances," in New Fundamental Technologies in Data Mining, K. Funatsu, Ed. InTech, 2011, vol. 1, ch. 19, pp. 369-387.

[265] Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.

[266] Daumé, H., 2017. *A course in machine learning* (pp. 149-155). Hal Daumé III.

[267] Platt, J., 1998. Fast training of support vector machines using sequential minimal optimization, In, B. Scholkopf, C. Burges, A. Smola,(eds.): Advances in Kernel Methods-Support Vector Learning.

[268] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K., 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural computation*, *13*(3), pp.637-649.

[269] Joshi, S., Pandey, B. and Joshi, N., 2015. Comparative analysis of Naive Bayes and J48 Classification Algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(12), pp.813-817.

[270] Deepali Kharche, K.R., 2014. Comparison Of Different Datasets Sing Various Classification Techniques With Modified WEKA. *International Journal of Computer Science and Mobile Computing, IJCSMC*, pp.389-393.

[271] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), pp.10-18.

[272] Sherman, D., 2017. [online] Schankacademy.com. Available at: <https://www.schankacademy.com/demos/data-analytics/xt/lib/docs/0/j48_parameters.pdf> [Accessed 3 December 2018].

[273] Saravana, N. and Gayathri, D.V., 2018. Performance and classification evaluation of J48 algorithm and Kendall's based J48 algorithm (KNJ48). *Int. J. Comput. Trends Technol. (IJCTT)--Volume*, *59*.

[274] Stiglic, G., Kocbek, S., Pernek, I. and Kokol, P., 2012. Comprehensive decision tree models in bioinformatics. *PloS one*, *7*(3), p.e33812.

[275] Kwok, S.W. and Carter, C., 1990. Multiple decision trees. In *Machine Intelligence and Pattern Recognition* (Vol. 9, pp. 327-335). North-Holland.

[276] Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

[277] Biau, G. and Scornet, E., 2016. A random forest guided tour. *Test*, *25*(2), pp.197-227.

[278] Schwarz, D.F., König, I.R. and Ziegler, A., 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, *26*(14), pp.1752-1758.

[279] Díaz-Uriarte, R. and De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, *7*(1), pp.1-13.

[280] Genuer, R., Poggi, J.M. and Tuleau-Malot, C., 2010. Variable selection using random forests. *Pattern recognition letters*, *31*(14), pp.2225-2236.

[281] Breiman, L. and Cutler, A., 2003. Setting up, using, and understanding random forests V4. 0. *University of California, Department of Statistics*.

[282] Archer, K.J. and Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis*, *52*(4), pp.2249-2260.

[283] Auret, L. and Aldrich, C., 2011. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, *105*(2), pp.157-170.

[284] Toloşi, L. and Lengauer, T., 2011. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), pp.1986-1994.

[285] Bengio, Y., 2009. Learning deep architectures for AI. Found Trends Mach Learn 2:1–127

[286] Welbl, J., 2014, September. Casting random forests as artificial neural networks (and profiting from it). In *German Conference on Pattern Recognition* (pp. 765-771). Springer, Cham.

[287] Landwehr, N., Hall, M. and Frank, E., 2005. Logistic model trees. *Machine learning*, *59*(1-2), pp.161-205.

[288] Szymanski, A., 2020. *Combining logistic regression and decision tree.* [online] Medium. Available at: <https://towardsdatascience .com/combining-logistic-regression-and-decision-tree-1adec36a4b3 f> [Accessed 3 December 2020].

[289] Loh, W.Y., 2006. Logistic regression tree analysis. *Handbook of engineering statistics*, pp.537-549.

[290] Plato.stanford.edu. 2021. *Simpson's Paradox (Stanford Encyclopedia of Philosophy).* [online] Available at: <https://plato.stanford.edu/entries/paradox-simpson/> [Accessed 5 July 2021].

[291] Archive.ics.uci.edu. 2005. *UCI Machine Learning Repository*. [online] Available at: <https://archive.ics.uci.edu/ml/index.php> [Accessed 6 April 2018].

[292] Frank, E. and Hall, M., 2001, September. A simple approach to ordinal classification. In *European conference on machine learning* (pp. 145-156). Springer, Berlin, Heidelberg.

[293] Medium. 2019. *Simple Trick to Train an Ordinal Regression with any Classifier*. [online] Available at: <https://towardsdata science.com/simple-trick-to-train-an-ordinal-regression-with-any-classifier-6911183d2a3c> [Accessed 5 August 2020].

[294] Zadrozny, B. and Elkan, C., 2001, August. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 204-213).

[295] Provost, F., 2000, July. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets* (Vol. 68, No. 2000, pp. 1-3). AAAI Press.

[296] Japkowicz, N. and Stephen, S., 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), pp.429-449.

[297] Ling, C.X. and Sheng, V.S., 2008. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, *2011*, pp.231-235.

[298] McCarthy, K., Zabar, B. and Weiss, G., 2005, August. Does cost-sensitive learning beat sampling for classifying rare classes?. In *Proceedings of the 1st international workshop on Utility-based data mining* (pp. 69-77).

[299] DataRobot AI Cloud. 2012. Model Fitting. [online] Available at: <https://www.datarobot.com/wiki/fitting/> [Accessed 18 July 2020].

[300] Akalin, A., 2022. *5.1 How are machine learning models fit? | Computational Genomics with R*. [online] Compgenomr.github.io. Available at: <https://compgenomr.github.io/book/how-are-machine-learning-models-fit.html#machine-learning-vs-statistics> [Accessed 12 December 2021].

[301] Akalin, A., 2020. *Computational Genomics with r*. Chapman and Hall/CRC.

[302] Koehrsen, W., 2018. *Overfitting vs. Underfitting: A Complete Example*. [online] Medium. Available at: <https://towardsdata science.com/overfitting-vs-underfitting-a-complete-example-d05d d7e19765> [Accessed 23 July 2018].

[303] Bekkar, M., Djemaa, H.K. and Alitouche, T.A., 2013. Evaluation measures for models' assessment over imbalanced data sets. *J Inf Eng Appl*, *3*(10).

[304] Aldraimli, M., Soria, D., Parkinson, J., Thomas, E.L., Bell, J.D., Dwek, M.V. and Chaussalet, T.J., 2020. Machine learning prediction of susceptibility to visceral fat associated diseases. *Health and Technology*, *10*(4), pp.925-944.

[305] Wu, J. and Hicks, C., 2021. Breast Cancer Type Classification Using Machine Learning. *Journal of personalized medicine*, *11*(2), p.61.

[306] Kubat, M. and Matwin, S., 1997, July. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml* (Vol. 97, pp. 179-186).

[307] Ertekin, S., Huang, J., Bottou, L. and Giles, L., 2007, November. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 127-136).

[308] Zhang, Y. and Wang, D., 2013, January. A cost-sensitive ensemble method for class-imbalanced datasets. In *Abstract and applied analysis* (Vol. 2013). Hindawi.

[309] Biggerstaff, B.J., 2000. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in medicine*, *19*(5), pp.649-663.

[310] Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006, December. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015-1021). Springer, Berlin, Heidelberg.

[311] Chountas, P 2021, *Seminar 10: Data mining and Machine Learning,* lecture notes, Model Performance Metric and Ensembles, University of Westminster, delivered 7 December 2021.

[312] Batuwita, R. and Palade, V., 2009, December. A new performance measure for class imbalance learning. application to bioinformatics problems. In *2009 International Conference on Machine Learning and Applications* (pp. 545-550). IEEE.

[313] Chawla, N.V., Cieslak, D.A., Hall, L.O. and Joshi, A., 2008. Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery, 17(2), pp.225-252.

[314] Youden, W.J., 1950. Index for rating diagnostic tests. *Cancer*, *3*(1), pp.32-35.

[315] Ding, Z., 2011. Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics.

[316] Provost, F. and Fawcett, T., 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions In: Proc of the 3rd International Conference on Knowledge Discovery and Data Mining.

[317] Drummond, C. and Holte, R.C., 2000, August. Explicitly representing expected cost: An alternative to ROC representation. In *Proceedings*

of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 198-207).

[318] Muschelli, J., 2020. ROC and AUC with a binary predictor: a potentially misleading metric. *Journal of Classification*, *37*(3), pp.696-708.

[319] Halligan, S., Altman, D.G. and Mallett, S., 2015. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, *25*(4), pp.932-939.

[320] McClish, D.K., 1989. Analyzing a portion of the ROC curve. *Medical decision making*, *9*(3), pp.190-195.

[321] Weng, C.G. and Poon, J., 2008, November. A new evaluation measure for imbalanced datasets. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87* (pp. 27-32).

[322] Saito, T. and Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), p.e0118432.

[323] Magalhães, P.S.T., Revett, K. and Santos, H.D.D., 2006. Keystroke dynamics: stepping forward in authentication.

[324] Familydoctor.org. 2020. *Surgery for Obesity - familydoctor.org*. [online] Available at: <https://familydoctor.org/surgical-treatment-obesity/> [Accessed 16 October 2020].

[325] Medicalxpress.com. 2021. *Data science approach helps oncologists predict which patients will suffer side effects from radiotherapy*. [online] Available at: <https://medicalxpress.com/ news/2021-07-science-approach-oncologists-patients-side.html> [Accessed 19 November 2021].

[326] Chaiken, S. and Ledgerwood, A., 2011. A theory of heuristic and systematic information processing. *Handbook of theories of social psychology: Volume one*, pp.246-166.

[327] Requite.eu. 2014. *Information for the scientific community | REQUITE*. [online] Available at: <https://www.requite.eu /node/33> [Accessed 25 September 2020].

[328] Ukbiobank.ac.uk. 2022. *About us*. [online] Available at: <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/ about-us> [Accessed 31 January 2022].

[329] Rapaio, 2015. *strings as features in decision tree/random forest*. [online] Data Science Stack Exchange. Available at: <https://datascience.stackexchange.com/questions/5226/strings-as-features-in-decision-tree-random-forest> [Accessed 16 October 2021].

[330] Crusoveanu, L., Thai, N., Schimandle, T., Goebelbecker, E., Roza, G., Stalla, A., Oberle, C. and Gulden, M., 2020. *Data Normalization Before or After Splitting a Data Set?*. [online] Baeldung. Available at: <https://www.baeldung.com/cs/data-normalization-before-after-splitting-set> [Accessed 4 August 2021].

[331] Lo, I., 2018. *Should scaling be done on both training data and test data for machine learning? Can one do scaling on only the training data?*. [online] Quora. Available at: <https://www.quora.com/Should-scaling-be-done-on-both-training-data-and- test -data-for-machine-learning-Can-one-do-scaling-on-only-the-training-data/answer/Ian-Lo-7> [Accessed 29 March 2021].

[332] Ramos-Pérez, I., Arnaiz-González, Á., Rodríguez, J.J. and García-Osorio, C., 2022. When is resampling beneficial for feature selection with imbalanced wide data?. *Expert Systems With Applications*, *188*, p.116015.

[333] Revill, J. (2006) *Are you a tofi? (that's thin on the outside, Fat Inside)*, *The Guardian*. Available at: https://www.theguardian.com/science /2006/dec/10/medicineandhealth.health (Accessed: 22 January 2021).

[334] Wang, Y.C., McPherson, K., Marsh, T., Gortmaker, S.L. and Brown, M., 2011. Health and economic burden of the projected obesity trends in the USA and the UK. *The Lancet*, *378*(9793), pp.815-825.

[335] Sam, S. and Mazzone, T., 2014. Adipose tissue changes in obesity and the impact on metabolic function. *Translational Research*, *164*(4), pp.284-292.

[336] Dattilo, A.M. and Kris-Etherton, P.M., 1992. Effects of weight reduction on blood lipids and lipoproteins: a meta-analysis. *The American journal of clinical nutrition*, *56*(2), pp.320-328.

[337] Fox, C.S., Massaro, J.M., Hoffmann, U., Pou, K.M., Maurovich-Horvat, P., Liu, C.Y., Vasan, R.S., Murabito, J.M., Meigs, J.B., Cupples, L.A. and D'Agostino Sr, R.B., 2007. Abdominal visceral and subcutaneous adipose tissue compartments: association with metabolic risk factors in the Framingham Heart Study. *Circulation*, *116*(1), pp.39-48.

[338] Després, J.P., Lemieux, I., Bergeron, J., Pibarot, P., Mathieu, P., Larose, E., Rodés-Cabau, J., Bertrand, O.F. and Poirier, P., 2008. Abdominal obesity and the metabolic syndrome: contribution to global cardiometabolic risk. *Arteriosclerosis, thrombosis, and vascular biology*, *28*(6), pp.1039-1049.

[339] Chin, S.H., Kahathuduwa, C.N. and Binks, M., 2016. Physical activity and obesity: what we know and what we need to know. *Obesity Reviews*, *17*(12), pp.1226-1244.

[340] Golabi, P., Bush, H. and Younossi, Z.M., 2017. Treatment strategies for nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Clinics in Liver Disease*, *21*(4), pp.739-753.

[341] Uusitupa, M., Lindi, V., Louheranta, A., Salopuro, T., Lindström, J., Tuomilehto, J. and Finnish Diabetes Prevention Study Group, 2003. Long-term improvement in insulin sensitivity by changing lifestyles of people with impaired glucose tolerance: 4-year results from the Finnish Diabetes Prevention Study. *Diabetes*, *52*(10), pp.2532-2538.

[342] Brouwers, B., Hesselink, M.K., Schrauwen, P. and Schrauwen-Hinderling, V.B., 2016. Effects of exercise training on intrahepatic lipid content in humans. *Diabetologia*, *59*(10), pp.2068-2079.

[343] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. and Liu, B., 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, *12*(3), p.e1001779.

[344] Parkinson, J.R., 2019. Visceral adipose tissue, thigh adiposity and liver fat fraction: a cross-sectional analysis of the UK biobank. *UK Biobank.*

[345] Parkinson, J.R., Gerbault, P., Alenaini, W., Elliot, B., Wilman, H., Bell JD., Thomas, E.L., 2019. Physical activity, visceral adipose tissue, thigh adiposity and liver fat fraction: a cross sectional analysis of the UK Biobank. *UK Biobank.*

[346] Shuster, A., Patlas, M., Pinthus, J.H. and Mourtzakis, M., 2012. The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. *The British journal of radiology*, *85*(1009), pp.1-10.

[347] Thomas, E.L., Parkinson, J.R., Frost, G.S., Goldstone, A.P., Doré, C.J., McCarthy, J.P., Collins, A.L., Fitzpatrick, J.A., Durighel, G., Taylor-Robinson, S.D. and Bell, J.D., 2012. The missing risk: MRI and MRS phenotyping of abdominal adiposity and ectopic fat. *Obesity*, *20*(1), pp.76-87.

[348] Palmer, B.F. and Clegg, D.J., 2015. The sexual dimorphism of obesity. *Molecular and cellular endocrinology*, *402*, pp.113-119.

[349] Machann, J., Thamer, C., Schnoedt, B., Haap, M., Haring, H.U., Claussen, C.D., Stumvoll, M., Fritsche, A. and Schick, F., 2005. Standardized assessment of whole body adipose tissue topography by MRI. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *21*(4), pp.455-462.

[350] Training.parthenos-project.eu. 2020. *The Data Heterogeneity problem – Parthenos training*. [online] Available at: https://training.

parthenos-project.eu/sample-page/formal-ontologies-a-complete-novices-guide/what-is-data-heterogeneity/ [Accessed 25 March 2022].

[351] Bisschop, C.N.S., Peeters, P.H., Monninkhof, E.M., van der Schouw, Y.T. and May, A.M., 2013. Associations of visceral fat, physical activity and muscle strength with the metabolic syndrome. *Maturitas*, *76*(2), pp.139-145.

[352] Pasdar, Y., Darbandi, M., Mirtaher, E., Rezaeian, S., Najafi, F. and Hamzeh, B., 2019. Associations between muscle strength with different measures of obesity and lipid profiles in men and women: results from RaNCD cohort study. *Clinical nutrition research*, *8*(2), pp.148-158.

[353] Van Poppel, M.N., Chinapaw, M.J., Mokkink, L.B., Van Mechelen, W. and Terwee, C.B., 2010. Physical activity questionnaires for adults: a systematic review of measurement properties.. *Sports medicine*, *40*(7), pp.565-600.

[354] Helmerhorst, H.H.J., Brage, S., Warren, J., Besson, H. and Ekelund, U., 2012. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), pp.1-55.

[355] Hagströmer, M., Bergman, P., De Bourdeaudhuij, I., Ortega, F.B., Ruiz, J.R., Manios, Y., Rey-López, J.P., Phillipp, K., Von Berlepsch, J. and Sjöström, M., 2008. Concurrent validity of a modified version of the International Physical Activity Questionnaire (IPAQ-A) in European adolescents: The HELENA Study. *International journal of obesity*, *32*(5), pp.S42-S48.

[356] Ferrari, P., Friedenreich, C. and Matthews, C.E., 2007. The role of measurement error in estimating levels of physical activity. *American journal of epidemiology*, *166*(7), pp.832-840.

[357] Miyatake, N., Nishikawa, H., Morishita, A., Kunitomi, M., Wada, J., Suzuki, H., Takahashi, K., Makino, H., Kira, S. and Fujii, M., 2002. Daily walking reduces visceral adipose tissue areas and improves

insulin resistance in Japanese obese subjects. *Diabetes research and clinical practice*, *58*(2), pp.101-107.

[358] Mytton, O.T., Ogilvie, D., Griffin, S., Brage, S., Wareham, N. and Panter, J., 2018. Associations of active commuting with body fat and visceral adipose tissue: a cross-sectional population based study in the UK. *Preventive medicine*, *106*, pp.86-93.

[359] Weiss, G.M., McCarthy, K. and Zabar, B., 2007. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. *Dmin*, *7*(35-41), p.24.

[360] Grainger, A.T., Tustison, N.J., Qing, K., Roy, R., Berr, S.S. and Shi, W., 2018. Deep learning-based quantification of abdominal fat on magnetic resonance images. *PloS one*, *13*(9), p.e0204071.

[361] Deist, T.M., Dankers, F.J., Valdes, G., Wijsman, R., Hsu, I.C., Oberije, C., Lustberg, T., van Soest, J., Hoebers, F., Jochems, A. and El Naqa, I., 2019. Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers (vol 45, pg 3449, 2018). Medical Physics, 46(2), pp.1080-1087.

[362] Saednia, K., Tabbarah, S., Lagree, A., Wu, T., Klein, J., Garcia, E., Hall, M., Chow, E., Rakovitch, E., Childs, C. and Sadeghi-Naini, A., 2020. Quantitative thermal imaging biomarkers to detect acute skin toxicity from breast radiation therapy using supervised machine learning. International Journal of Radiation Oncology* Biology* Physics, 106(5), pp.1071-1083.

[363] Bentzen, S.M. and Overgaard, J., 1994, April. Patient-to-patient variability in the expression of radiation-induced normal tissue injury. In Seminars in radiation oncology (Vol. 4, No. 2, pp. 68-80). WB Saunders.

[364] Reddy, J., Lindsay, W.D., Berlind, C.G., Ahern, C.A. and Smith, B.D., 2018. Applying a machine learning approach to predict acute toxicities during radiation for breast cancer patients. *International Journal of Radiation Oncology• Biology• Physics*, *102*(3), p.S59.

[365] Delishaj, D., D'amico, R., Corvi, D., De Nobili, G., Alghisi, A., Colangelo, F., Cocchi, A., Declich, F. and Soatti, C.P., 2020. Management of grade 3 acute dermatitis with moist desquamation after adjuvant chest wall radiotherapy: a case report. *Radiation Oncology Journal*, *38*(4), p.287.

[366] Delishaj, D., D'amico, R., Corvi, D., De Nobili, G., Alghisi, A., Colangelo, F., Cocchi, A., Declich, F. and Soatti, C.P., 2020. Management of grade 3 acute dermatitis with moist desquamation after adjuvant chest wall radiotherapy: a case report. *Radiation Oncology Journal*, *38*(4), p.287.

[367] UK, C. R. (2014) 'Cancer Research UK statitistics'.

[368] Krishnankutty, B., Bellary, S., Kumar, N.B. and Moodahadu, L.S., 2012. Data management in clinical research: An overview. *Indian journal of pharmacology*, *44*(2), p.168.

[369] West, C., Azria, D., Chang-Claude, J., Davidson, S., Lambin, P., Rosenstein, B., De Ruysscher, D., Talbot, C., Thierens, H., Valdagni, R. and Vega, A., 2014. The REQUITE project: validating predictive models and biomarkers of radiotherapy toxicity to reduce side-effects and improve quality of life in cancer survivors. Clinical oncology, 26(12), pp.739-742.

[370] Isrctn.com. 2020. ISRCTN - Search Results. [online] Available at: <http://www.isrctn.com/search?q=ISRCTN98496463> [Accessed 25 November 2020].

[371] Garciarena, U. and Santana, R., 2017. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Systems with Applications, 89, pp.52-65.

[372] Sizechart.com. 2020. Bra Sister Size. [online] Available at: <http://www.sizechart.com/brasize/sistersize/index.html> [Accessed 25 November 2020].

[373] Lustgarten, J.L., Gopalakrishnan, V., Grover, H. and Visweswaran, S., 2008. Improving classification performance with discretisation on biomedical datasets. In AMIA annual symposium proceedings (Vol. 2008, p. 445). American Medical Informatics Association.

[374] Hassan, M.S.U., Ansari, J., Spooner, D. and Hussain, S.A., 2010. Chemotherapy for breast cancer. Oncology reports, 24(5), pp.1121-1131.

[375] nhs.uk. 2020. Breast Cancer in Women - Treatment. [online] Available at: https://www.nhs.uk/conditions/breast-cancer/treatment/ [Accessed 9 April 2020].

[376] Williams, M.V., Denekamp, J. and Fowler, J.F., 1985. A review of αβ ratios for experimental tumors: implications for clinical studies of altered fractionation. International Journal of Radiation Oncology* Biology* Physics, 11(1), pp.87-96.

[377] Sebastian raschka. 2014. About Feature Scaling And Normalization And The Effect Of Standardization For Machine Learning Algorithms. [online] Available at: <https://sebastianraschka.com/Articles/2014_about_feature_scaling.html> [Accessed 9 April 2020].

[378] Kirkham, E. and Conroy, R., 2015. *Should outliers be removed before or after data transformation?*. [online] ResearchGate. Available at: <https://www.researchgate.net/post/Should_outliers_be_removed_before_or_after_data_transformation/55ce70575dbbbd4c758b45ef/citation/download.> [Accessed 7 March 2021].

[379] Wright, J.L., Takita, C., Reis, I., Zhao, W. and Hu, J.J., 2012. Rate of Moist Desquamation in Patients Receiving Radiation for Breast Cancer After Mastectomy Versus Breast-Conserving Surgery. International Journal of Radiation Oncology• Biology• Physics, 84(3), p.S222.

[380] Ozenne, B., Subtil, F. and Maucort-Boulch, D., 2015. The precision–recall curve overcame the optimism of the receiver operating

characteristic curve in rare diseases. *Journal of clinical epidemiology*, *68*(8), pp.855-859

[381] Pham, B., Cranney, A., Boers, M., Verhoeven, A.C., Wells, G. and Tugwell, P., 1999. Validity of area- under-the-curve analysis to summarize effect in rheumatoid arthritis clinical trials. *The Journal of rheumatology*, *26*(3), pp.712-716.

[382] Louppe, G., Wehenkel, L., Sutera, A. and Geurts, P., 2013. Understanding variable importances in forests of randomised trees. In Advances in neural information processing systems (pp. 431-439).

[383] Fritz, G., Henninger, C. and Huelsenbeck, J., 2011. Potential use of HMG-CoA reductase inhibitors (statins) as radioprotective agents. *British medical bulletin*, *97*(1), pp.17-26.

[384] De Langhe, S., Mulliez, T., Veldeman, L., Remouchamps, V., van Greveling, A., Gilsoul, M., De Schepper, E., De Ruyck, K., De Neve, W. and Thierens, H., 2014. Factors modifying the risk for developing acute skin toxicity after whole-breast intensity-modulated radiotherapy. BMC cancer, 14(1), p.711.

[385] Twardella,D.,Popanda,O.,Helmbold,I.,Ebbeler,R.,Benner,A.,vonFournier,D.,Haase,W.,Sautter- Bihl, M.L., Wenz, F., Schmezer, P. and Chang-Claude, J., 2003. Personal characteristics, therapy modalities and individual DNA repair capacity as predictive factors of acute skin toxicity in an unselected cohort of breast cancer patients receiving radiotherapy. Radiotherapy and Oncology, 69(2), pp.145-153.

[386] Back,M.,Guerrieri,M.,Wratten,C.andSteigler,A.,2004.Impactofradiationtherapyonacutetoxicity in breast conservation therapy for early breast cancer. Clinical Oncology, 16(1), pp.12-16.

[387] Deantonio, L., Gambaro, G., Beldì, D., Masini, L., Tunesi, S., Magnani, C. and Krengli, M., 2010. Hypofractionated radiotherapy after conservative surgery for breast cancer: analysis of acute and late toxicity. Radiation Oncology, 5(1), p.112.

[388] Barnett, G.C., Wilkinson, J.S., Moody, A.M., Wilson, C.B., Twyman, N., Wishart, G.C., Burnet, N.G. and Coles, C.E., 2011. The Cambridge Breast Intensity-modulated Radiotherapy Trial: patient-and treatment-related factors that influence late toxicity. Clinical oncology, 23(10), pp.662-673.

[389] Terrazzino, S., La Mattina, P., Masini, L., Caltavuturo, T., Gambaro, G., Canonico, P.L., Genazzani, A.A. and Krengli, M., 2012. Common variants of eNOS and XRCC1 genes may predict acute skin toxicity in breast cancer patients receiving radiotherapy after breast-conserving surgery. Radiotherapy and Oncology, 103(2), pp.199-205.

[390] Sharp,L.,Johansson,H.,Hatschek,T.andBergenmar,M.,2013.Smoking asanindependentriskfactor for severe skin reactions due to adjuvant radiotherapy for breast cancer. The breast, 22(5), pp.634-638.

[391] Tortorelli, G., Di Murro, L., Barbarino, R., Cicchetti, S., di Cristino, D., Falco, M.D., Fedele, D., Ingrosso, G., Janniello, D., Morelli, P. and Murgia, A., 2013. Standard or hypofractionated radiotherapy in the post-operative treatment of breast cancer: a retrospective analysis of acute skin toxicity and dose inhomogeneities. BMC cancer, 13(1), p.230.

[392] Dzeroski, S., Zenko, B.: Is Combining Classifiers Better than Selecting the Best One? In: Proceedings of the Nineteenth International Conference on Machine Learning, San Francisco, Morgan Kaufmann (2002).

[393] Strobl,C.,Boulesteix,A.L.,Zeileis,A.andHothorn,T.,2007.Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), p.25.

[394] Semenenko, V.A. and Li, X.A., 2008. Lyman–Kutcher–Burman NTCP model parameters for radiation pneumonitis and xerostomia based on combined analysis of published clinical data. *Physics in Medicine & Biology*, *53*(3), p.737.

[395] Gulliford, S.L., Partridge, M., Sydes, M.R., Webb, S., Evans, P.M. and Dearnaley, D.P., 2012. Parameters for the Lyman Kutcher Burman (LKB) model of Normal Tissue Complication Probability (NTCP) for specific rectal complications observed in clinical practise. *Radiotherapy and Oncology*, *102*(3), pp.347-351.

[396] Rattay, T., Seibold, P., Aguado Barrera, M.E., Altabas, M., Azria, D., Barnett, G.C., Bultijnck, R., Chang-Claude, J., Choudhury, A., Coles, C.E. and Dunning, A., 2020. External validation of prediction models for acute skin toxicity in the REQUITE breast cohort. *Frontiers in Oncology*, *10*, p.2153.

[397] Sugak, E., 2018, June. Occupational risks management as a basis of industrial injuries and occupational disease prevention. In *IOP Conference Series: Materials Science and Engineering* (Vol. 365, No. 6, p. 062038). IOP Publishing.