A Comparative Machine Learning Modelling Approach for Patients' Mortality Prediction in Hospital Intensive Care Unit

Aldraimli, M., Nazyrova, N., Djumanov, A., Sobirov, I. and Chaussalet, T.J.

A Data Science Approach for Predicting Patient's Susceptibility to Acute Side Effects in Breast Cancer Radiation Therapy

Mahmoud Aldraimli[1], Daniele Soria[2], Diana Grishchuck[3], Samuel Ingram[4], Robert Lyons[5], Anil Mistry[6], Jorge Oliveira[7], Robert Samuel[8], Leila E.A. Shelly[9], Sarah Osman[10], Miriam V. Dwek[11], Christopher J. Talbot[12], Catharine M. West[13], Tim Rattay[12], David Azria[14], Jenny Chang-Claude[15], Sara Gutiérrez-Enríquez[16], Tiziana Rancati[17], Barry S Rosenstein[18], Dirk De Ruysscher[19], Elena Sperk[20], R Paul Symonds[21], Ana Vega[22], Liv Veldeman[23] and Thierry J. Chaussalet[1].

[1] The Health Innovation Ecosystem, University of Westminster, London, UK
[2] School of Computing, University of Kent (Medway), Chatham Maritime, UK
[3] Imperial College Healthcare NHS Trust, London, UK
[4] Division of Cancer Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, UK
[5] Department of Computer Science, Edge Hill University, Ormskirk, Lancashire, UK
[6] Guy's and St Thomas' NHS Foundation Trust, London, UK
[7] Mirada Medical, Oxford, UK
[8] University of Leeds, Leeds Cancer Centre, St. James's University Hospital, Leeds, UK
[9] Department of Engineering, University of Cambridge, Cambridge, UK
[10] Patrick G Johnston Centre for cancer research, Queen's University Belfast, Belfast, UK
[11] School of Life Sciences, University of Westminster, London, UK
[12] Cancer Research Centre, University of Leicester, Leicester, UK
[13] Institute of Cancer Sciences, Christie Hospital, Wilmslow Road, Manchester, UK
[14] University of Montpellier, France
[15] German Cancer Research Center (DKFZ) Division of Cancer Epidemiology, Unit of Genetic Epidemiology, Heidelberg, Germany
[16] Vall d'Hebron Institute of Oncology, Barcelona, Spain
[17] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[18] Mount Sinai School of Medicine, New York, USA
[19] Maastricht Radiation Oncology (MAASTRO Clinic) University Hospital Maastricht, The Netherlands
[20] Department of Radiation Oncology, University Medical Center Mannheim, Medical Faculty Mannheim, Heidelberg University, Germany
[21] Department of Oncology, Leicester Royal Infirmary, UK
[22] Fundación Publica Galega Medicina Xenomica, Santiago de Compostela, Spain
[23] Department of Basic Medical Sciences, University Hospital Ghent, Belgium

Corresponding Author: Mahmoud Aldraimli
The Health Innovation Ecosystem
School of Computer Science and Engineering
University of Westminster
115 New Cavendish Street
London W1W 6UW
United Kingdom
m.aldraimli2@westminster.ac.uk

Abstract

The prediction by classification of side effects incidence in a given medical treatment is a common challenge in medical research. Machine Learning (ML) methods are widely used in the areas of risk prediction and classification. The primary objective of such algorithms is to use several features to predict dichotomous responses (e.g., disease positive/negative). Similar to statistical inference modelling, ML modelling is subject to the problem of class imbalance and is affected by the majority class, increasing the false-negative rate. In this study, seventy-nine ML models were built and evaluated to classify approximately 2000 participants from 26 hospitals in eight different countries into two groups of radiotherapy (RT) side effects incidence based on recorded observations from the international study of RT related toxicity "REQUITE". We also examined the effect of sampling techniques, cost-sensitive learning and meta-learning methods on the models when dealing with class imbalance. The combinations of resampling and meta techniques used had a significant impact on the classification. They resulted in an improvement in incidence status prediction by facilitating an increase in the information contained within each variable. Based on domain expert criteria, the best classification model for RT acute toxicity prediction was identified. The Area Under Receiver Operator Characteristic curve of the models tested with an isolated dataset ranged between 0.50 and 0.77. The scale of improved results is promising and will be used to guide further development of models to predict RT acute toxicities. One new model was optimised and found to be beneficial to identify patients who are at risk of developing acute RT early toxicities during or after breast RT treatment ensuring relevant treatment management interventions can be appropriately targeted. The ML models presented in this paper were developed by a multi-disciplinary collaboration of data scientists, medical physicists, oncologists and surgeons in the UK Radiotherapy Machine Learning Network.

Keywords

Classification; REQUITE; Machine Learning; Imbalanced Learning; Radiotherapy; Early Toxicities; Desquamation; Chemotherapy.

Ethics approval

The questionnaire and methodology for this study were approved by the REQUITE publications committee. The REQUITE study was registered with International Standard Randomised Controlled Trial Number Register (ref: ISRCTN98496463), and written consent was obtained from all participants before their involvement.

Statement of no conflict of interest

The authors whose names are listed immediately below the manuscript title certify that they have NO affiliations with or involvement in any organisation or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

1. Introduction

A common real-world problem facing Machine Learning (ML) is the lack of good data. While data preparation and modelling often consume most of the time of developing ML solutions, data quality is essential for the algorithms to function as intended. Noisy, dirty, and incomplete data are common obstacles to creating ML solutions [1]. Routinely collected health data are data collected without specific a priori research questions developed before collection [2]. Health data of this type are used widely for clinical, pharmacoepidemiologic and health services research. However, the quality of these data remains in question; hence data scientists often need a combination of domain knowledge as well as an in-depth understanding of ML to examine and cleanse such data. Such a process sheds light on the significance of interdisciplinary collaborations in this type of research.

In ML modelling, the imbalance and lack of uniform distribution across patients' groups in health data also form a challenge for both industrial and research domains [3]. There are multiple techniques to tackle class imbalance [4], of which data enrichment is the most straightforward. Other more sophisticated methods include varied sampling techniques [5], cost-sensitive learning [6], [7], feature selection; more complex strategies include meta-learning [8], combining classifiers [9], and algorithmic modifications [10].

Resampling methods often raise questions over their suitability [11]. For example: is the new resampled dataset representative of the population in relation to the response variable? Is it acceptable to artificially generate synthetic data of class subjects when training ML classification models? It has been argued that by using sampling methods, the original class ratio is lost during the training process and that this affects the accuracy metrics [12]. Similarly, training ML models with synthetic data may compromise accuracy measures by deceiving the process of cross-validation [13].

While most learning algorithms train under the assumption that the cost of misclassification is identical across outcome groups [14], penalising classifiers with cost-sensitive classification for incorrect predictions is a practical solution to the problem in many fields, like the medical domain of our study. In the medical domain, defining such a cost is challenging [15]. In treatment management scenarios, the cost of a false positive might be derived by the monetary cost of performing subsequent tests. In contrast, there is not a monetary equivalent cost for administering treatment on a patient and get further health complications.

The primary goal of this study is to identify the best ML models to predict acute desquamation - a common side effect following breast cancer radiotherapy (RT) - before the start of any treatment; such models are of particular interest to cancer clinicians.

The deployment of ML modelling in this study aims at tackling a real-world treatment management challenge. Over 70% of breast cancer patients receive RT during the course of their treatment [16]. RT is recommended to all breast cancer patients who have a local excision following mastectomy in high-risk patients [17]. Radiation treatment reduces the rates of cancer recurrence following local excision and increases long-term survival [18]. As survival from breast cancer continues to improve [19], quality-of-life (QoL) and survivorship have become an increasingly important research priority [20].

Radiation toxicity can be estimated from empirical dosimetric models based on the dose to the target organ and surrounding tissue [21] [22]. However, there is considerable variation between individual patients' normal tissue reaction to RT and resultant toxicities, including skin desquamation [23]. In a significant minority of patients, this can cause substantial patient morbidity and can worsen the cosmetic outcome following breast surgery. At the severe end, acute desquamation (skin loss) can result in the interruption of RT or even a total dose reduction, potentially increasing the risk of local recurrence. Thus, acute radiation toxicities can have an adverse effect on the QoL in a significant minority of breast cancer patients. This effect could be reduced if a patient's individual risk of radiation toxicity was better known. This would allow treatment plans to be personalised and inform discussions about treatment risks and benefits with patients.

Given the paucity of validated predictive models for RT-related toxicity, it is important to build models with optimal predictive performance. ML is well placed to achieve this. Recent studies have demonstrated an ability to develop well-performing predictive models for radiation toxicities [24] [25], including a thermal image-based classifier to predict breast radiation skin toxicity after the first week of RT [26]. In this paper, using the large REQUITE breast cancer cohort, we compare eight different ML algorithms (building a total of 79 models) for predicting acute skin desquamation. The new models were built using Cost-Sensitive (CS) learning [27], Random Under Sampling (RUS), Synthetic Minority Over Sampling Technique (SMOTE) and Random Over Sampling (ROS) techniques [28] [29] applied to highly imbalanced training data. 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 monitored visually by Adaptive Projection Analysis (APA) [30] and numerically via Information Gain (IG) attribute evaluation [31]. The ML models presented in this paper were developed by a multi-disciplinary collaboration of data scientists, medical physicists, oncologists and surgeons in the UK Radiotherapy Machine Learning Network.

The paper is structured as follows: section 2 has a brief description of the study cohort. The methodology, methods and approaches used in this study are presented in section 3. The results and analysis are documented in section 4, with the discussion and next steps in sections 5 and 6, respectively.

2. Study cohort and participants

The study is a cross-sectional assessment of an international, prospective cohort study recruited cancer patients in 26 hospitals in eight countries between April 2014 and March 2017. This study uses collected data from 2069 patients who underwent breast RT. There were 192 patients (9.3%) with acute desquamation (grade 1≥ulceration or grade≥ 3 erythema). The median age of breast patients was 58 years (range 23-80 years), treated with a median dose to the breast of 50 Gy (28.5-56 Gy) in 25 fractions (5-31), and 64 % of patients received boost treatment. Further details on the prevalence of clinical risk factors are widely available [32].

Binary ML classification models were built to predict acute desquamation development in breast RT patients. The REQUITE team provided patients' data, and the questionnaire and methodology for this study were approved by the REQUITE publications committee. The multicentre-REQUITE breast cancer patients cohort was recruited prospectively in seven European countries and the US. The cohort was used for building predictive ML models throughout this study. The REQUITE study was conceived as a multicentre validation cohort for predictive models of radiation toxicity collecting data under a unified protocol [33]. Patient baseline characteristics and methodology have been described in detail elsewhere [34]. All patients were treated with Breast Conservative Surgery (BCS) followed by External Beam RT (EBRT) according to local protocol. Although late toxicity was the primary endpoint in REQUITE, data collected at the end of radiation treatment was used to document acute toxicity. All patients gave written informed consent. The study was approved by local ethics committees in participating countries and registered at the ISRCTN registry [35] (ISRCTN98496463). For the full list and sequence of the methods used in this study see Fig.1 in section 3.

2.1 Response variable (endpoint) definition

Radiation toxicity in REQUITE was scored using CTCAE (Common Terminology Criteria for Adverse Events) v4.0 [36]. CTCAE v4.0 has separate scales for radiation dermatitis (erythema or redness) and skin ulceration (skin loss), both of which are relevant to the acute response to RT in the breast. The primary endpoint of this study was acute desquamation (skin loss or moist desquamation) occurring by the end of treatment. Cases were defined as patients who experienced either grade≥3 radiation dermatitis (moist desquamation) or CTCAE grade≥1 skin ulceration, implying that skin integrity has been broken over the breast or in the inframammary fold. There were 192 patients (9.3%) with acute desquamation.

3. Methodology

The methodology corresponds to a merge of several data mining tasks into two key phases which were carried out in collaboration with medical domain experts. The first phase combines data cleaning, preparation and pre-processing tasks; it includes predictors selections, error detection, data labelling and imputation. The second phase combines the modelling, evaluation and simplification tasks.

3.1 Data preparation and pre-processing

Fig.1 shows the sequence of data preparation and pre-processing tasks as they were deployed to this study. The raw REQUITE dataset (n = 2069) contained (m > 300) variables (features). Variables were initially nominated manually using domain expertise in modelling desquamation by clinicians and RT physicists and only a set of m = 136 applicable variables and n = 2058 ($Desq^{+} = 192, Desq^{-} = 1866$) records remained (Case-wise deletion (n=11 with missing class endpoint observations). The nominated set proceeded to preparation and pre-processing; its variables include baseline characteristics, familial history, breast cancer staging information, chemotherapy regimens, lifestyle attributes, medical conditions, household characteristics, sociodemographic factors, medical operations, treatment history, female-specific factors, mental and behavioural disorders, medications, quality of life aspects and breast RT clinical measurements such as normo-fractionation procedure. In data preparation, Boundary Value Analysis (BVA) and Equivalence Class Partitioning (EPC) techniques [37] were used for detecting and correcting or removing corrupt or inaccurate records from the dataset. Also, missingness analysis was performed by cross-checking the data with the

REQUITE study questionnaire design to ascertain the causes of incomplete records and deduce patterns. A combination of non-statistical and statistical imputation techniques was used, non-statistical methods were used to reduce uncertainty via logical rule imputation and variable dropping [38] (see Table 1). The investigation of missing data patterns [38] assisted in 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). The retained dataset for feature engineering transformation and modelling finally had m=123 variables and n=2058 records.

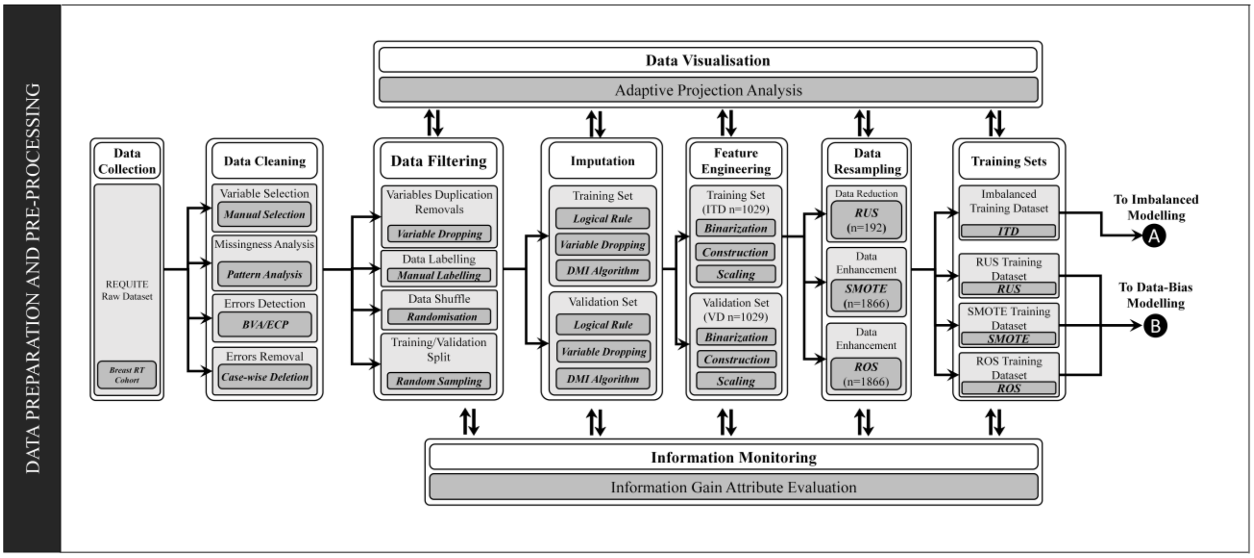


Fig. 1 Data preparation and pre-processing tasks used in this study

Table 1. Percentage of Imputed missing observations 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 dataset-split with class stratification was performed to sample both the raw Imbalanced Training Dataset (raw ITD, n=1029) and the raw Validation Dataset (raw VD, n=1029). The process was followed by applying a state-of-the-art hybrid statistical-ML imputation for each set independently with ML Decision-Tree based Missing-Value Imputation (DMI) Technique [39] 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 [40]. The evaluation of information worth is highly affected by the number of records; hence the 50:50 training-test split to allow for a fair information bias comparison, see supplementary Information Gain Attribute Evaluation Table A.

The retained 123 variables for modelling consisted of:

- 106 raw features.
- Breast size measurement calculated as a single continuous variable by adding bra cup and band sizes, to represent 'sister' sizes equal to the same breast volume [41].
- For instance, a UK size 34B bra holds an approximate breast volume equal to 32C, approximately 390 cc.
- Sixteen additional features described below.

For data pre-processing and feature engineering, sixteen additional 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, we opted to binarise [42] the prescriptions based on their generic chemical names (see table 2). One-Hot Encoding converted chemotherapy drugs categorical values into a form that could be provided to ML algorithms to improve prediction performance [43]. The categorical value represents the administered chemo-drugs combinations in a chemotherapy regime. The combinations values start from zero goes all the way up to N-1 categories. One-Hot encoding binarisation is performed at a category level (single observation level per attribute) which converted every chemo-drug used in a chemotherapy regime into a new feature.

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 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 at the same time. These combinations are known as chemotherapy regimens. Every cancer responds differently to chemotherapy. Common breast cancer chemotherapy regimens include AT, AC, AC+T, CMF, CEF, CAF, TAC and others [44]. NHS UK published a wide range of chemotherapy side effects which may occur to breast cancer patients, some of whom may have plans to undergoing breast RT [45]. Therefore, including chemotherapy attributes in this study was recommended.

Table 2. Illustration of the binarisation of 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

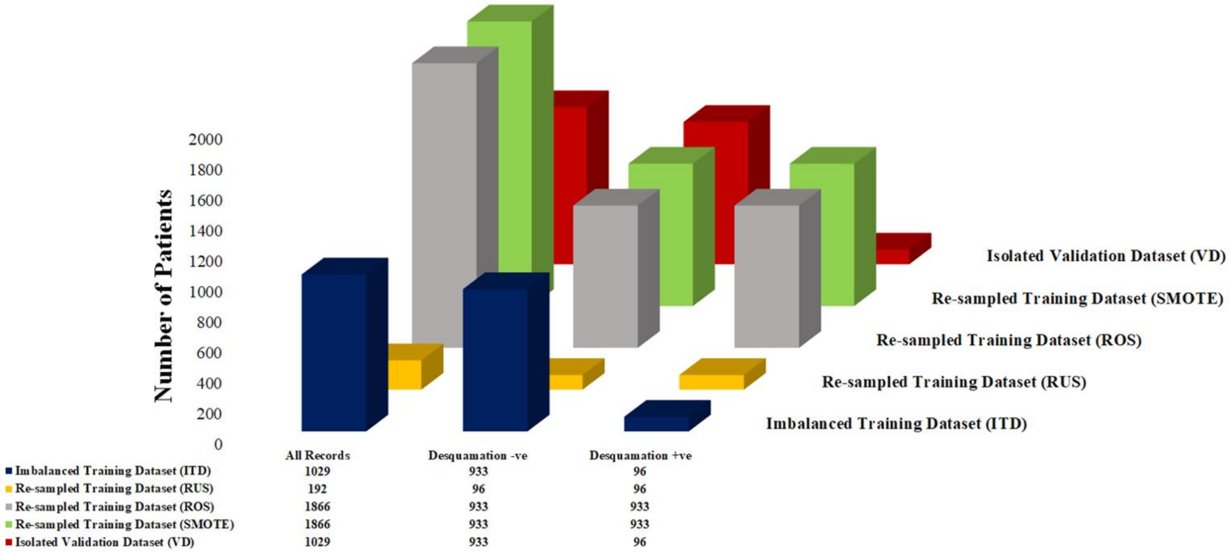
In order to adjust for different RT regimens, the dose was calculated as the biologically effective dose (BED). BED is the product of the number of fractions (n), dose per fraction (d), and a factor determined by the dose and α/β ratio for skin (10 Gy), which is used in radiobiology to describe the slope of the cell survival curve for different irradiated tissues [46]. Three features were constructed by calculating the BED.

$$BED = n\,d\left(1 + \frac{d}{\alpha/\beta}\right)$$

CTCAE endpoint definition was used to label the patients to create a binary response variable. Out of all 123 variables, all numeric features (m=63) were normalised with $\boldsymbol{\mathcal{Z}}$-score standardisation [47]. The $\boldsymbol{\mathcal{Z}}$-score indicates the distance from each value in each variable to its mean in the units of standard deviation. Feature standardisation scales the values of the observations of each feature in the data to have a zero mean. The need for feature scaling (standardisation) emerges in the REQUITE dataset since it contains features which highly vary in magnitudes, units and range. For example, there is a large difference in magnitude of breast volume measurement in cm^3 and the photon radio dose per fraction in Gray (Gy).

$$x' = \frac{x - \bar{x}}{\sigma}$$

In a breast radiation treatment, only a small portion of patients suffer from acute desquamation [48], that is also reflected in the REQUITE dataset known as a problem of class imbalance. This 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. Class imbalance is a common challenge in ML modelling [4]. One approach to tackle class imbalance in the training data is to apply three data resampling techniques to ITD≡DMI(ITD), by which the endpoint response classes of records become equal (see Fig.2); Random Under Sampling (RUS) (n=192, $Desq^{+} = 96, Desq^{-} = 96$), Random Over Sampling (ROS) ($\text{n} = 1866,\ Desq^{+} = 933, Desq^{-} = 933$) and Synthetic Minority Oversampling Technique (SMOTE) ($\text{n} = 1866,\ Desq^{+} = 933, Desq^{-} = 933$) [28] [29]. The effect of such resampling techniques on the training dataset was visualised with a multi-dimensional Adaptive Projection Algorithm (APA) [30] into a 3D point cloud.



	All Records	Desquamation -ve	Desquamation +ve
🟦 Imbalanced Training Dataset (ITD)	1029	933	96
🟨 Re-sampled Training Dataset (RUS)	192	96	96
⬜ Re-sampled Training Dataset (ROS)	1866	933	933
🟩 Re-sampled Training Dataset (SMOTE)	1866	933	933
🟥 Isolated Validation Dataset (VD)	1029	933	96

Fig. 2 The visualisation of samples size for ITD, RUS, ROS, SMOTE training datasets and validation dataset VD.

3.2 Modelling, Evaluation and Simplification

In this second phase, we apply a complex mix of model building, evaluation and simplification tasks, which flow is shown in Fig.3. The training set (ITD) n=1029 breast cancer patients who underwent breast RT is used to train eight ML algorithms (each of a different learning scheme) with 10-Fold Cross-Validation [13] to avoid the problem of overfitting. In relation to their cohort, the trained models are tested on the isolated validation dataset (VD) n=1029. The description of the REQUITE dataset variables is reported in a previous study [34]. Both ITD and VD are equally imbalanced ($Desq^{+} = 96, Desq^{-} = 933$).

The resampled datasets RUS, ROS and SMOTE, are used to train each of the same ML algorithms. These algorithms are Discretised Naïve Bayes (NB) [49], Logistic Regression with Ridge Estimator (LR) [50], Artificial Neural Networks (ANN) with a multi-layer perceptron architecture [51], Support Vector Machine (SVM) with polynomial

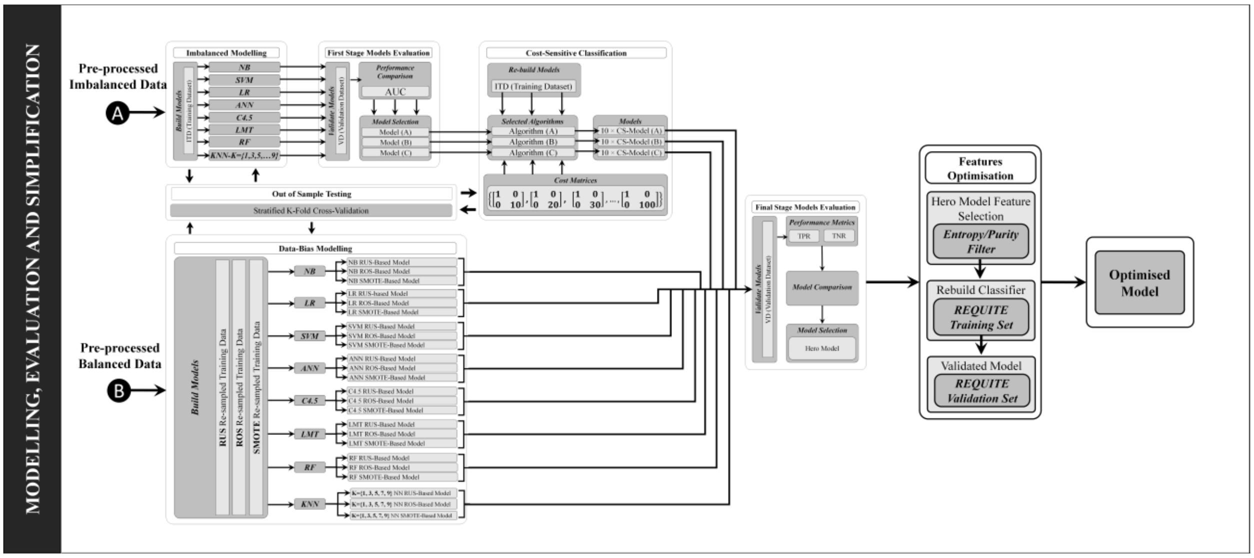


Fig. 3 Models Building, evaluation and simplification methodology used in this study

kernel and Logistic calibrator [52], K-Nearest Neighbour (KNN) [53] with K={1,3,5,7,9}, Decision Trees (C4.5) [54], Logistic Model Tree (LMT) [55] and Random Forest (RF) [56]. Alternative meta-learning approach to overcome class imbalance known as Cost-Sensitive Classification (CS) [27] was used to impose penalties (costs) for the misclassification of the positive group (false negative prediction) only during the model training process with the imbalanced training dataset (ITD). In this study the cost for a false negative prediction is not linked to a monetary value; instead, a ten-step Incremental Inverse Class Distribution cost was used [57], ITD has a (96:933 ≅ 1:10) ratio of examples in the positive class to examples in the negative class. This ratio is inverted to penalise false negative (FN) with a ten-step incrementation at an initial cost x: 1 of 10:1 increasing to 100:1. The cost is applied in the form of Charles Elkan's explicit cost matrix notation below [27].

$$Cost\ Matrix\ Combinations\ \begin{bmatrix} \mathrm{FP}(1) & \mathrm{TN}(0) \\ \mathrm{TP}(0) & \mathrm{FN}(x) \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 20 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 30 \end{bmatrix}, \cdots, \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix} \right\}$$

Three ML algorithms out of the competing eight were systematically selected for Cost-Sensitive Learning modelling. All algorithms used for this study were implemented in Waikato Environment for Knowledge Analysis (WEKA) 3.8.3 (with the default models' parameters settings), with the C4.5 using the J48 implementation, KNN using the IBK implementation and SVM using SMO implementation.

A two-stage performance evaluation was applied, three performance metrics were used to compare and assess the performance of all models after being validated on test datasets (VD). At the first stage, the Area Under Receiver Operator Characteristic Curve (AUC-ROC) [57] was used to select classifiers trained with ITD which achieved the highest AUC for CS modelling improvement, while the Sensitivity (True Positive Rate TPR) and the Specificity (True Negative Rate TNR) [57] were used at the second stage to compare the final models' performances and contribute to its interpretability. Having a model with a large number of features makes its interpretability complex or even opaque. Opaque models are hard to trust by clinicians and physicians. Having a smaller number of features improves interpretability and performance. The clinical specialists made it clear that the requirement is to model with all carefully selected features to understand their impact and importance. A purity filter was used to select fewer features to optimise the final model [58].

4. Results analysis

The APA visualisation [30] in Fig. 4 can be used to indicate the classes which can be separated, the attribute combinations which are most associated with each class, 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 and how resampling techniques are addressing the balance.

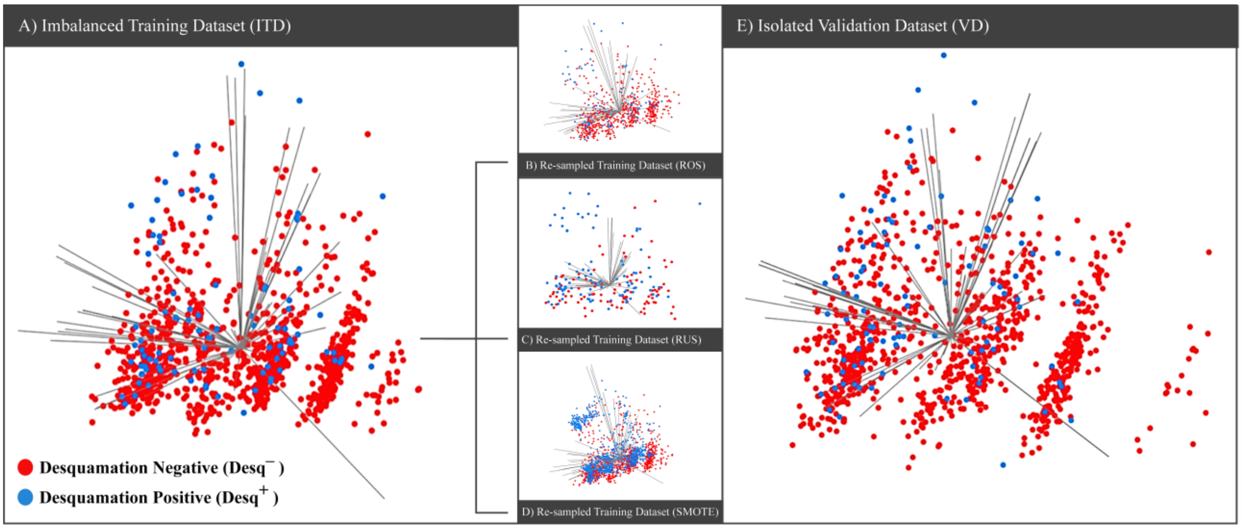


Fig. 4 The APA visualisation of imputed ITD, RUS, ROS, SMOTE training datasets and validation dataset.

ROS training dataset shows somewhat widely scattered positive class records since ROS re-sampling technique randomly duplicated records from the positive class. While SMOTE resampling technique has intensified the existing positive class records by generating synthetic prototype records analogous to the positive class records, these records seem to cluster near the original positive records. The RUS visualisation depicts how a balanced dataset may be able to 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 similar, thus suggesting that the randomised data split did not introduce any major bias into either dataset and that the training dataset is representative of the whole data.

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 data was split using just one feature. We used IG to monitor any bias that occurs in either training or validation datasets. Entropy and purity could vary as a result of 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 Fig. 5, it is notable that most of their features preserved close purity and entropy levels before and after imputation. Features that showed dominance in IG evaluation before DMI imputation have also maintained power after DMI imputation. Note that the imputation of ITD and VD separately removes the opportunity of both datasets sharing the same statistical parameter setting used by the imputation algorithm. This execution makes both the training and validation datasets utterly independent from each other and entirely isolated.

As for the models built with ITD dataset, a single model was built and validated for each of the eight ML algorithms, with the exception to KNN, for which five models were built with ITD and validated with VD, to account for the different values of K parameter, where K= {1,3,5,7,9} [59]. Table 3 shows the models' AUC, TPR and TNR [57] for all twelve models in training and validation. The training and validation performance results illustrate the impact 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) [57]. At this stage of modelling, for an imbalanced model to compete for selection for further improvement with cost-sensitive classification modelling, the selected imbalanced model needs to achieve the highest AUC in validation which indicates the highest degree of discrimination of at least one of the classes or both. The improvement is achieved with incremental inverse-class distribution cost matrix to penalise the classifier for the misclassification of FN records. The incremental penalty will skew correct classification towards the positive group as there are no further improvements required for the negative class. The highest three champions in AUC performance in validation with (VD) LMT ranked first with AUC of 0.746; RF was not far behind with AUC of 0.742 and NB in third place with AUC of 0.737 all show a good AUC > 0.70; however, the TPR is poor. The highest sensitivity was achieved by the NB model of 0.500, followed by LMT with 0.042 and RF with 0.010. The confusion matrices in Table 4 describe the numeric count of correctly classified patients, FP (type I) and FN (type II) errors misclassifications).

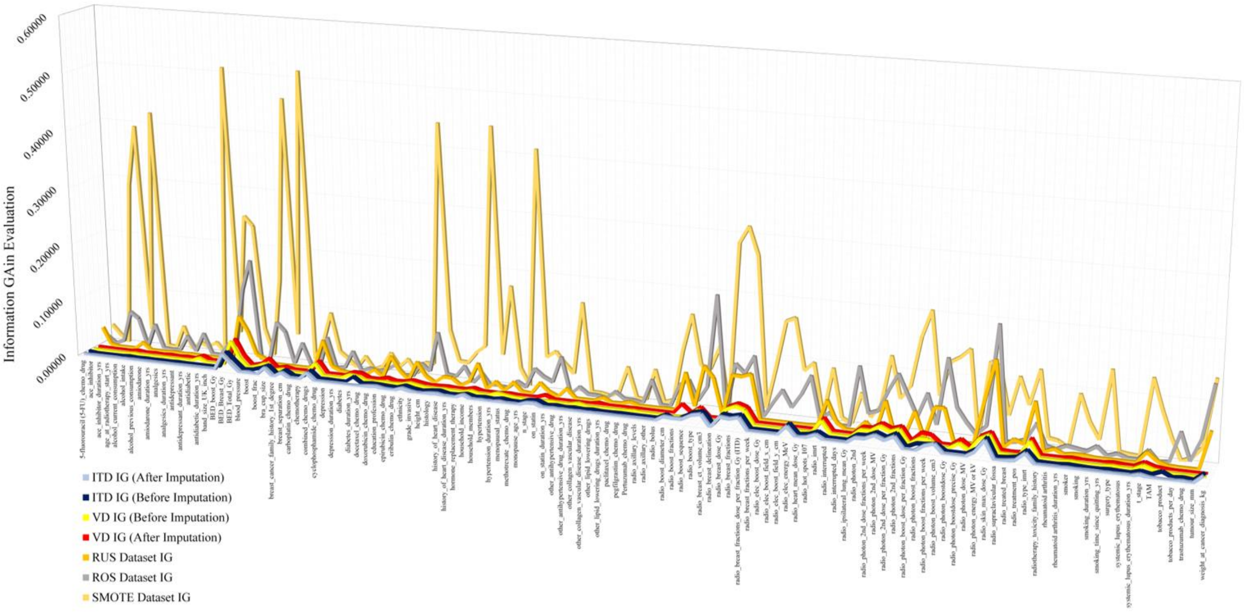


Fig. 5 The IG levels comparison of ITD, RUS, ROS, SMOTE training datasets and validation dataset.

Table 3. Imbalanced ML models' performances with ITD training set

Classifier	**Training with ITD (n=1029)**			**Validation with VD (n=1029)**			**Rank**
	Specificity (TNR)	**Sensitivity (TPR)**	**AUC**	**Specificity (TNR)**	**Sensitivity (TPR)**	**AUC**	
LMT	0.996	0.010	0.578	0.995	0.042	0.746	1
RF	0.998	0.021	0.725	1.000	0.010	0.742	2
NB	0.810	0.438	0.697	0.833	0.500	0.737	3
ANN	0.945	0.198	0.694	0.953	0.177	0.676	4
KNN (K=9)	0.999	0.031	0.660	0.999	0.042	0.665	5
KNN (K=5)	0.985	0.042	0.624	0.989	0.063	0.651	6
KNN (K=7)	0.996	0.031	0.648	0.998	0.052	0.644	7
KNN (K=3)	0.975	0.094	0.601	0.979	0.125	0.627	8
KNN (K=1)	0.908	0.167	0.548	0.923	0.292	0.607	9
LR	0.910	0.188	0.567	0.959	0.135	0.596	10
SVM	0.966	0.156	0.561	0.976	0.146	0.561	11
C4.5	0.985	0.083	0.575	0.979	0.125	0.496	12

Table 4. The validation confusion matrices of LMT, RF and NB imbalanced ML models (Trained with ITD)

LMT				**RF**				**NB**			
Predicted				***Predicted***				***Predicted***			
Desq -ve	*Desq +ve*			*Desq -ve*	*Desq +ve*			*Desq -ve*	*Desq +ve*		
928	5	*Desq -ve*	***Actual***	933	0	*Desq -ve*	***Actual***	777	156	*Desq -ve*	***Actual***
92	4	*Desq +ve*		95	1	*Desq +ve*		48	48	*Desq +ve*	

Fig. 6 shows the validation results of the cost-sensitive RF, NB and LMT models. The effect of applying incremental cost is indirectly proportional to a decrease of specificity per model, and the false positive (FP) increases as the penalty increases. A significant improvement is made in TPR; models with higher penalty showed higher sensitivity. The TPR improvement is rapid for all models as the cost of FN increases. NB sensitivity ranges from 0.50 in the unpenalised model to 0.771 for a penalty of 100. The greatest improvement in sensitivity was achieved by RF ranging from 0.010

for the unpenalised model to 0.792 at a penalty of 100. LMT sensitivity improved from 0.042 without a penalty to 0.646 at a penalty of 100.

The TNR and TPR validation performances of resampling techniques RUS, ROS and SMOTE for RF, LMT, NB, C4.5, ANN, KNN, SVM and LR classifiers are also in Fig. 6. It shows that resampling techniques improved sensitivity across all classifiers with RUS achieving the least variance between specificity and improved sensitivity.

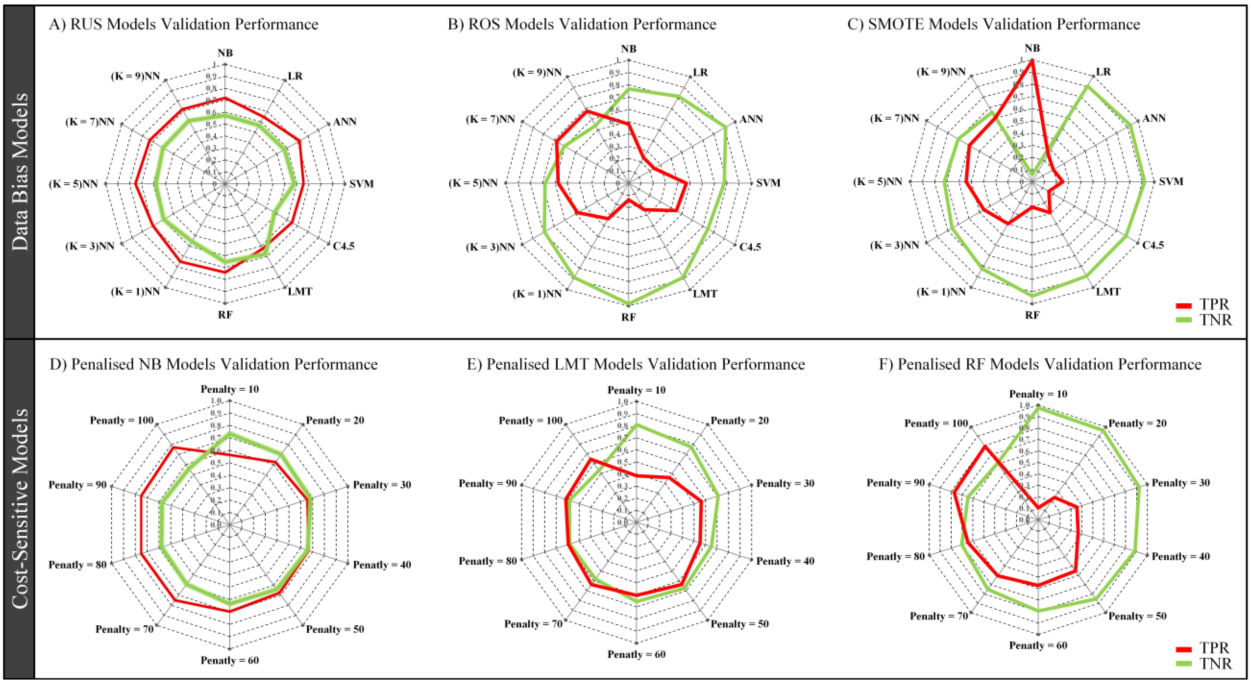


Fig. 6 Radar charts plotting The Ture Positive Rate (TPR) and Ture Negative Rate (TNR) for RUS, ROS, SMOTE and Cost-sensitive validated models. Penalty values refer to FN prediction costs in the explicit cost-sensitive models. While FP predictions costs are kept at a value of 1, both TP and TN predictions costs always remained at the value of zero

4.1 Model's selection and simplification

Based on all models' validation TPR and TNR evaluations and the clinicians' trade-off between TPR and TNR in Fig. 7, it is found that there are two trade-off conditions that models compete towards, based on a lower and upper threshold values of 0.63 and 0.70 respectively. These conditions are (TPR ≥ 0.63 & TNR ≥ 0.70) and (TNR ≥ 0.63 & TPR ≥ 0.70). Five models met both conditions. They are CS-RF(FN:FP=90:1, TNR=0.65, TPR=0.77, AUC=0.76), RUS-RF(TNR=0.65, TPR=0.74, AUC=0.74), CS-NB(FN:FP=60:1, TNR=0.64, TPR=0.70, AUC= 0.72), CS-RF(FN:FP=80:1, TNR=0.70, TPR=0.65, AUC=0.75) and CS-NB(FN:FP=20:1, TNR=0.70. TPR=0.63, AUC=0.73). The confusion matrices for the compliant five validated models are found in Table 5.

Table 5. The confusion matrices of the compliant five validated models with VD

Cost-sensitive RF Cost ratio (FN : FP = 90:1)			RUS-based RF Sampling ratio (r = 1)			Cost-sensitive NB Cost ratio (FN : FP = 60:1)			Cost-sensitive RF Cost ratio (FN : FP = 80:1)			Cost-sensitive NB Cost ratio (FN : FP = 20:1)		
	←*Predicted*	1st		←*Predicted*	2nd		←*Predicted*	3rd		←*Predicted*	4th		←*Predicted*	5th
$Desq^-$	$Desq^+$	*Actual* ↓	$Desq^-$	$Desq^+$	*Actual* ↓	$Desq^-$	$Desq^+$	*Actual* ↓	$Desq^-$	$Desq^+$	*Actual* ↓	$Desq^-$	$Desq^+$	*Actual* ↓
602	331	$Desq^-$	608	325	$Desq^-$	592	341	$Desq^-$	655	278	$Desq^-$	654	279	$Desq^-$
22	74	$Desq^+$	25	71	$Desq^+$	29	67	$Desq^+$	34	62	$Desq^+$	36	60	$Desq^+$
Accuracy = 66%			Accuracy = 66%			Accuracy = 64%			Accuracy = 70%			Accuracy = 69%		

Maximising TPs is essential; therefore, specialists' consensus concluded that the best performing model was CS-RF(FN:FP = 90:1) for exceeding all other models' sensitivity and AUC performances while maintaining a competitive specificity. The ranking of the compliant performing models based on domain experts' success criteria are in table 6.

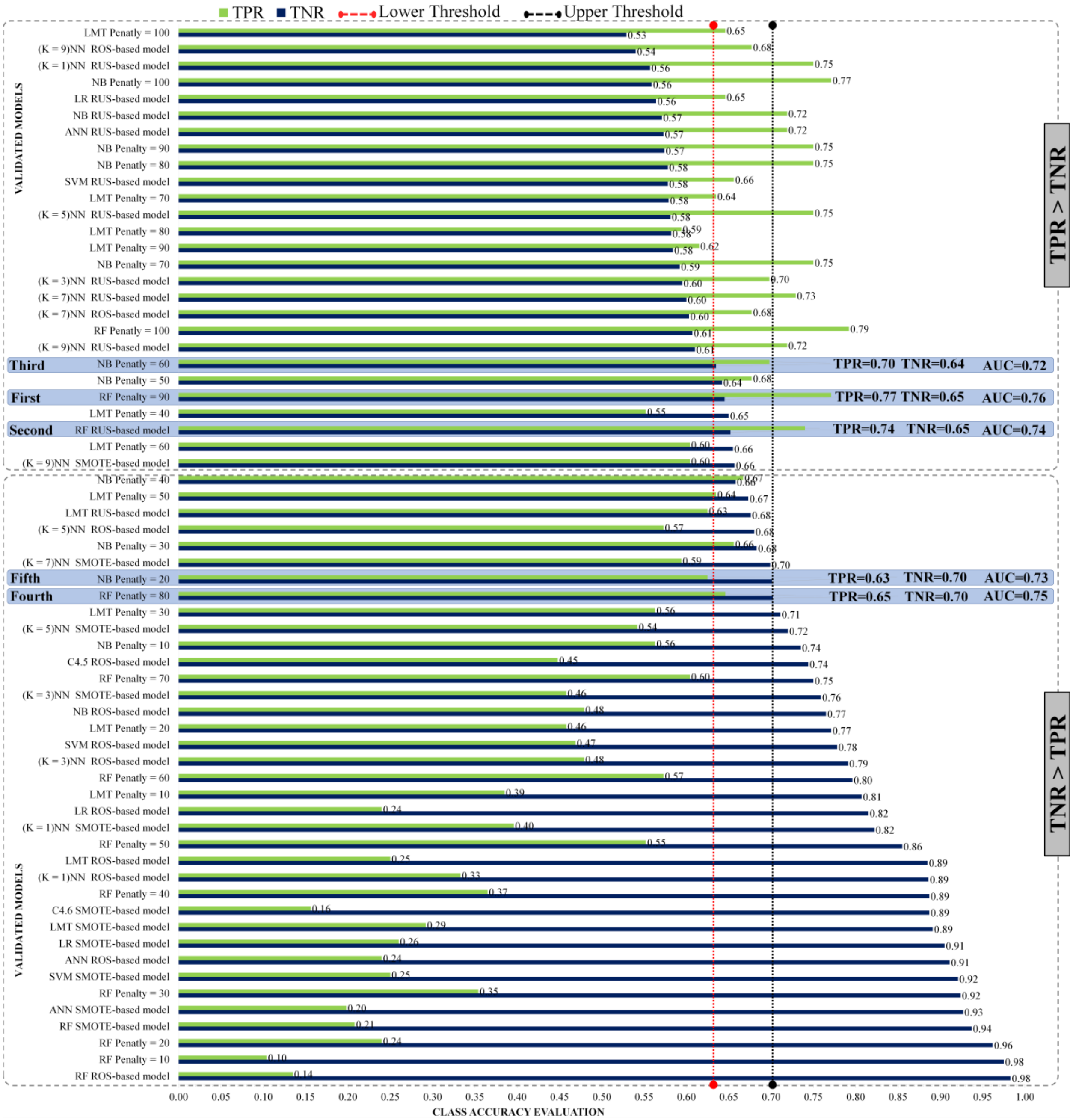


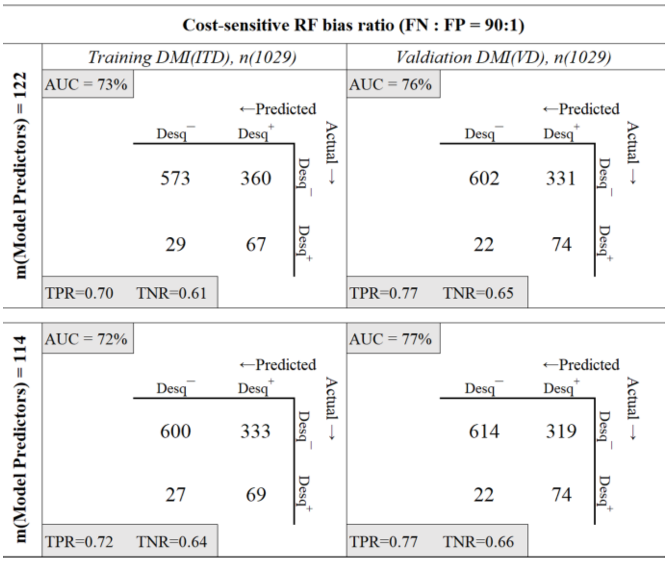
Fig. 7 The True Positive Rate (TPR) and True Negative Rate (TNR) trade-offs threshold lines for all validated models with VD. Penalty values refer to FN prediction costs in the explicit cost-sensitive models. While FP predictions costs are kept at a value of 1, both TP and TN predictions costs always remained at the value of zero

Table 6. The performance ranking of the compliant five validated models with VD

Rank	Learner	Bias type	Bias ratio	TNR	TPR	AUC
1	RF	Cost-sensitive	Misclassification cost (FN:FP = 90:1)	0.65	0.77	0.76
2	RF	Data re-sampling	RUS (r = 1)	0.65	0.74	0.74
3	NB	Cost-sensitive	Misclassification cost (FN:FP = 60:1)	0.64	0.70	0.72
4	RF	Cost-sensitive	Misclassification cost (FN:FP = 80:1)	0.70	0.65	0.75
5	NB	Cost-sensitive	Misclassification cost (FN:FP = 20:1)	0.70	0.63	0.73

The top-performing model has many predictors M=122, which makes its interpretability quite complicated. Feature importance in RF was calculated with Mean Decrease Impurity [58]. Eight features were estimated to have zero importance for the model CS-RF(FN:FP = 90:1). In order to simplify the model, these features were removed, and the model was rebuilt and validated. As a result, the simplified model performance slightly improved its specificity to 0.66 and AUC to 0.77, while its sensitivity remained unchanged. Feature importance is described in the Supplementary Material Tables B and C. The final simplified Hero model's performance is described in Table 7.

Table 7. The simplified final model training and validation confusion matrices performances



Cost-sensitive RF bias ratio (FN : FP = 90:1)

	Training DMI(ITD), n(1029)			*Valdiation DMI(VD), n(1029)*		
m(Model Predictors) = 122	AUC = 73%	←Predicted		AUC = 76%	←Predicted	
	Actual ↓	$Desq^-$	$Desq^+$	Actual ↓	$Desq^-$	$Desq^+$
	$Desq^-$	573	360	$Desq^-$	602	331
	$Desq^+$	29	67	$Desq^+$	22	74
	TPR=0.70 TNR=0.61			TPR=0.77 TNR=0.65		
m(Model Predictors) = 114	AUC = 72%	←Predicted		AUC = 77%	←Predicted	
	Actual ↓	$Desq^-$	$Desq^+$	Actual ↓	$Desq^-$	$Desq^+$
	$Desq^-$	600	333	$Desq^-$	614	319
	$Desq^+$	27	69	$Desq^+$	22	74
	TPR=0.72 TNR=0.64			TPR=0.77 TNR=0.66		

5. Discussion

The overall goal of this study was to predict radiation therapy acute toxicity desquamation in breast cancer patient's participants from the REQUITE cohort and to apply ML methods to classify these subjects into susceptibility to toxicity occurrence or non-occurrence categories. The ability to predict and classify this variable, using simple clinical routinely collected data will have a significant impact on 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 except imaging and genomic factors which previously demonstrated to correlate with acute desquamation significantly. [32]

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. A purity based ranking technique, IG was calculated to evaluate the worth of each input feature independently. When observing IG evaluation after the randomised and stratified training/validation data split, it was noted that few variables in the validation dataset (VD) contained a different worth of information as compared to the training set (ITD). 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, yet, this purity measure differs from correlation association, and it is not utilised as a feature selection in this study. Observed IG evaluation also showed that some variables in the VD contained a higher worth of information as compared to 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 sored 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 approach into the models' features 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 [60 – 66].

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 to 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 distinguish between both class labels at the endpoint. Some ML models may still calculate otherwise and utilise them in building the predictive models depending on the learning mechanism, hence including all 122 predictors in the modelling process.

For ML modelling, tackling the imbalanced class problem has a significant impact on the performance of standard ML algorithms. Also, the classification modelling performance in the training phase is severely impacted by class separability. Training standard ML algorithms with highly imbalanced classes without any adjustment to the training set results in an accuracy bias towards the majority class. In this study, we tackled that bias by applying two approaches. In the first approach, resampling techniques (RUS, ROS and SMOTE) were 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. The second approach (a cost-sensitive approach) awarded higher weights for 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 as to whether other remedies for imbalanced data classifications, such as Ensembles Learning (which are implemented at the algorithmic level), could result in better performances [8] [9] [10]. The advantages of resampling techniques evaluated here, however, include simplicity and transportability. Nevertheless, they are limited by the amount of IG manipulation because of their application resulting in biased predictions towards 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. The complexity becomes recursive since the attention to the minority records of different ML classifiers of various learning schemes is shifted differently for the same misclassification penalty when building predictive models.

The REQUITE breast RT dataset utilised in this study showed that applying the correct level of resampling without disrupting the original data distribution in the case of RUS-based method, together with the desired choice of performance metrics and slight manipulation of IG levels, produced a prediction solution which competed with further developed models with algorithmic modifications in the case of cost-sensitive classification. Among all 79 models reported in this study, five models satisfied the trade-off threshold conditions (see table 6). However, one "hero" model was selected for this specific domain problem that is a cost-sensitive RF model with 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. In the resampled models' results analysis, the learning scheme's impact was seen to decrease with the class imbalance severity in datasets compared to balanced datasets. In cost-sensitive classification, classifiers behaved very differently for the same cost matrix when trained on the same dataset.

The "hero" model was further simplified by discarding eight features which were deemed unimportant according to RF model-based feature selection method Mean Decrease Impurity (MDI) zero value, and the "hero" classifier is rebuilt with the remaining 114 features. The performance of the "hero" model continued to show a slight improvement in TNR. When using the MDI, which is an impurity-based ranking filter, feature selection based on impurity reduction is biased towards preferring variables with more categories [67]. This bias is not a problem in our study, since MDI was only used to optimise (simplify) a model with known performance. However, if the dataset contains two (or more) correlated features, then from the model's point of view, any of these correlated features can be used as a top predictor, with no preference of 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, when interpreting the model, it can provide a misleading perception that one of the variables is a strong predictor while the others in the same group are unimportant, while in fact, they are very closely associated with the response endpoint (see Table 8 and Table 9). The effect of 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 [68].

Like most biomedical case studies, when biochemical tests are performance assessed, in our study, the data obtained is heavily skewed (imbalanced). 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 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, and the area under the ROC curve gives an idea about the benefit of using the test in question. However, the imbalanced datasets tend to provide a much better ROC curve; therefore, visual interpretation and comparisons of ROC curves for ML models trained with imbalanced datasets can be misleading [69] [70] as observed in all ITD-based models in Table 3. Therefore, additional performance metrics are required to provide a more accurate representation of the models' validation. 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.

A limitation of this and many other ML papers used in radiation oncology is the number of variables used compared to routine practice. Real-world applicability is 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. 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 this approach to predict RT toxicities can move beyond the research setting and into clinical practice.

6. Clinical implication and next steps

Our study shows that the application of traditional ML algorithms to datasets of phenotype and clinical variables offers a fast and inexpensive solution to predict acute toxicities (moist desquamation) for breast cancer RT patients by aligning the classification task to predict specific adverse skin effects based on a Common Terminology Criteria for Adverse Events. The selection of a binary-class prediction task in this study is strategic to include patients classed within severe, life-threatening and death criteria. It identifies patients who are at higher risk of developing acute desquamation condition 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 testing the models with an isolated group of patients of similar ratio to the training data makes this study suitable for follow-up research in medical screening to identify subjects that may require treatment intervention.

This domain problem is the first to use the clinical features only 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 any other research from the same domain. Nevertheless, this work will be followed by further analyses where additional methods to improve the outcomes will be investigated.

Acknowledgement

This research collaboration was formed by the UK Radiotherapy Machine Learning (RTML) Network supported by members of the REQUITE steering group funded through the Advanced RT Challenge+ by the Science and Technology Facilities Council (STFC). The REQUITE study received funding from the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 601826. This publication presents independent research funded by the NIHR. The study was supported by the Quintin Hogg Trust research awards award no.165435391. The workshops were hosted by the University of Manchester and the Health and Innovation Ecosystem at the University of Westminster. Members of the REQUITE steering group are Ananya Choudhury, Alison Dunning, Rebecca M Elliott, Anusha Müller, Petra Seibold, Hilary Stobart and Adam Webb. Members of RTML steering group are Andrew Green and Nigel Mason OBE. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. We sincerely thank all patients who participated in the REQUITE study and all REQUITE staff involved at the participating hospitals.

References

1. L’heureux, A., Grolinger, K., Elyamany, H.F. and Capretz, M.A., 2017. Machine learning with big data: Challenges and approaches. IEEE Access, 5, pp.7776-7797.

2. Nicholls, S.G., Langan, S.M. and Benchimol, E.I., 2017. Routinely collected data: the importance of high-quality diagnostic coding to research. *CMAJ*, *189*(33), pp.E1054-E1055.

3. Yang, Q., Wu, X.: 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making 5(4), 597-604 (2006).

4. Gu, J., Zhou, Y., Zuo, X.: Making Class Bias Useful: A Strategy of Learning from Imbalanced Data. In: Yin H., Tino P., Corchado E., Byrne W., Yao X. (eds.) IDEAL 2007, LNCS, vol. 4881, pp 287-295. Springer, Heidelberg (2007).

5. More, A.: Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv:1608.06048 [stat.AP] (2016).

6. Weiss G.M., McCarthy, K., Zabar, B.: Cost-Sensitive Learning vs Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? In: Proceedings of the 2007 International Conference on Data Mining, pp. 35-41, Las Vegas, USA (2007).

7. Bekkar, M., Taklit, A.A.: Imbalanced Data Learning Approaches Review. International Journal of Data Mining & Knowledge Management Process (IJDKP) 3(4), 15-33 (2013).

8. Ensemble Learning to Improve Machine Learning Results, https://blog.statsbot.co/ensemble-learning-d1dcd548e936, last accessed: 2019/02/19.

9. 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).

10. Choi, J.M.: A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines. Iowa State University (Graduate Theses and Dissertation) (2010).

11. Unbalanced Data Is a Problem? No, Balanced Data Is Worse, https://matloff.wordpress.com/2015/09/29/un-balanced-data-is-a-problem-no-balanced-data-is-worse/, last accessed: 2019/02/24.

12. When should I balance classes in a training data set? https://stats.stackexchange.com/questions/227088/when-should-i-balance-classes-in-a-training-data-set, last accessed: 2018/11/22.

13. Bharat Rao, R., Fung, G., Rosales R.: On the Dangers of Cross-Validation. An Experimental Evaluation. In: Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 588-596 (2008).

14. Ling, C.X. and Sheng, V.S., 2008. Cost-sensitive learning and the class imbalance problem. Encyclopedia of machine learning, 2011, pp.231-235.

15. McCarthy, K., Zabar, B., Weiss, G, M., (2005), "Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes?", Proc. Int’l Workshop Utility-Based Data Mining, pp 69-77

16. UK, C. R. (2014) 'Cancer Research UK statitistics'.

17. National Collaborating Centre for Cancer (UK, 2009. Early and locally advanced breast cancer: diagnosis and treatment.

18. Early Breast Cancer Trialists' Collaborative Group, 2011. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials. The Lancet, 378(9804), pp.1707-1716.

19. Murphy, R., 2020. Cancer Survival In England - Office For National Statistics. [online] Ons.gov.uk. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancersurvivalinenglandadultsdiagnosed/2010and2014andfollowedupto2015> [Accessed 25 November 2020].

20. National Institutes of Health (NIH). 2010. National Cancer Institute (NCI). [online] Available at: <https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-cancer-institute-nci> [Accessed 25 November 2020].

21. Emami, B., Lyman, J., Brown, A., Cola, L., Goitein, M., Munzenrider, J.E., Shank, B., Solin, L.J. and Wesson, M., 1991. Tolerance of normal tissue to therapeutic irradiation. International Journal of Radiation Oncology* Biology* Physics, 21(1), pp.109-122.

22. Fowler, J.F., 1989. The linear-quadratic formula and progress in fractionated radiotherapy. The British journal of radiology, 62(740), pp.679-694.

23. 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.

24. Dean, J., Wong, K., Gay, H., Welsh, L., Jones, A.B., Schick, U., Oh, J.H., Apte, A., Newbold, K., Bhide, S. and Harrington, K., 2018. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. Clinical and translational radiation oncology, 8, pp.27-39.

25. Lee, S., Kerns, S., Ostrer, H., Rosenstein, B., Deasy, J.O. and Oh, J.H., 2018. Machine learning on a genome-wide association study to predict late genitourinary toxicity after prostate radiation therapy. International Journal of Radiation Oncology* Biology* Physics, 101(1), pp.128-135.

26. 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.

27. Elkan, C., 2001, August. The foundations of cost-sensitive learning. International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 973-978).

28. Chawla, N.V., 2009. Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886). Springer, Boston, MA.

29. Luis Torgo, PB and Ribeiro, R., 2016. A survey of predictive modeling under imbalanced distributions. ACM Comput. Surv, 49(2), pp.1-31.

30. Faith, J., Mintram, R., Angelova, M.: Gene expression Targeted projection pursuit for visualising gene expression data classifications. Bioinformatics 22(21), 2667–2673 (2006).

31. Harris, E., 2002, January. Information Gain Versus Gain Ratio: A Study of Split Method Biases. In ISAIM.

32. 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.

33. 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.

34. 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.

35. Isrctn.com. 2020. ISRCTN - Search Results. [online] Available at: <http://www.isrctn.com/search?q=ISRCTN98496463> [Accessed 25 November 2020].

36. Chen, A.P., Setser, A., Anadkat, M.J., Cotliar, J., Olsen, E.A., Garden, BC and Lacouture, M.E., 2012. Grading dermatologic adverse events of cancer treatments: the Common Terminology Criteria for Adverse Events Version 4.0. Journal of the American Academy of Dermatology, 67(5), pp.1025-1039.

37. Arnicane, V., 2009. Complexity of equivalence class and boundary value testing methods. International Journal of Computer Science and Information Technology, 751, pp.80-101.

38. 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.

39. 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).

40. Quinlan, J.R.: Induction of Decision Trees. Machine Learning 1(1), 81-106 (1986).

41. Sizechart.com. 2020. Bra Sister Size. [online] Available at: <http://www.sizechart.com/brasize/sistersize/index.html> [Accessed 25 November 2020].

42. DeepAI. 2020. Binarization. [online] Available at: <https://deepai.org/machine-learning-glossary-and-terms/binarization> [Accessed 9 April 2020].

43. 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.

44. Hassan, M.S.U., Ansari, J., Spooner, D. and Hussain, S.A., 2010. Chemotherapy for breast cancer. Oncology reports, 24(5), pp.1121-1131.

45. nhs.uk. 2020. Breast Cancer In Women - Treatment. [online] Available at: <https://www.nhs.uk/conditions/breast-cancer/treatment/> [Accessed 9 April 2020].

46. 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.

47. Sebastianraschka. 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].

48. 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.

49. Efron, B., 2013. Bayes' theorem in the 21st century. Science, 340(6137), pp.1177-1178.

50. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. Logistic regression. New York: Springer-Verlag.

51. Graupe, D., 2013. Principles of artificial neural networks (Vol. 7). World Scientific.

52. Platt, J., 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimisation. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning.

53. Aha, D.W., Kibler, D. and Albert, M.K., 1991. Instance-based learning algorithms. Machine learning, 6(1), pp.37-66.

54. Quinlan, J.R., 1996. Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 4, pp.77-90.

55. Landwehr, N., Hall, M. and Frank, E., 2005. Logistic model trees. Machine learning, 59(1-2), pp.161-205.

56. Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

57. 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).

58. 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).

59. 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.

60. Twardella, D., Popanda, O., Helmbold, I., Ebbeler, R., Benner, A., von Fournier, 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.

61. Back, M., Guerrieri, M., Wratten, C. and Steigler, A., 2004. Impact of radiation therapy on acute toxicity in breast conservation therapy for early breast cancer. Clinical Oncology, 16(1), pp.12-16.

62. 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.

63. 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.

64. 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.

65. Sharp, L., Johansson, H., Hatschek, T. and Bergenmar, M., 2013. Smoking as an independent risk factor for severe skin reactions due to adjuvant radiotherapy for breast cancer. The breast, 22(5), pp.634-638.

66. 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.

67. Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1), p.25.

68. Zhu, S., Wang, D., Yu, K., Li, T. and Gong, Y., 2008. Feature selection for gene expression using model-based entropy. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(1), pp.25-36.

69. 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).

70. Ekelund, S., 2017. Precision-Recall Curves – What Are They And How Are They Used?. [online] Acutecaretesting.org. Available at: <https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used> [Accessed 22 April 2020].

Supplementary material

Table A. Information Gain Attribute Evaluation.

Information and entropy levels within independent variables were monitored using an Information Gain Attribute Evaluator (IG) Algorithm. This algorithm evaluates the worth of each attribute by measuring information (purity) with respect to the class in combination with a ranker algorithm that ranks the attributes by their influence on the class. IG assisted in spotting and removing variables duplications but mainly helped to monitor and report any information bias introduced as a result of data splitting, imputation and resampling. This supplementary table shows the information gain evaluation for each predictor per data set.

Variable Name	Data Type	Imbalanced Training Data (ITD) N=1029			RUS Training Data N=192	ROS Training Data N=1866	SMOTE Training Data N=1866	Validation Data (VD) N=1029		
		IG(Raw)	IG(Imputed)	ΔIG	IG(RUS)	IG(ROS)	IG(SMOTE)	IG(Raw)	IG(Imputed)	ΔIG
5-fluorouracil (5-FU) _chemo_drug	CAT	0.00186	0.00186	0.00000	0.03211	0.01134	0.02820	0.00074	0.00074	0.00000
ace_inhibitor	CAT	0.00002	0.00002	0.00000	0.00330	0.00001	0.01242	0.00039	0.00039	0.00000
ace_inhibitor_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00646	0.00000	0.00000	0.00000	0.00000
age_at_radiotherapy_start_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.06043	0.28719	0.00000	0.00000	0.00000
alcohol_current_consumption	NUM	0.00000	0.00000	0.00000	0.00000	0.04980	0.39272	0.00000	0.00000	0.00000
alcohol_intake	CAT	0.00092	0.00111	0.00019	0.01246	0.00144	0.02850	0.00155	0.00185	0.00031
alcohol_previous_consumption	NUM	0.00000	0.00000	0.00000	0.00000	0.04232	0.41889	0.00000	0.00000	0.00000
amiodarone	CAT	0.00041	0.00041	0.00000	0.00000	0.00107	0.00161	0.00059	0.00059	0.00000
amiodarone_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
analgesics	CAT	0.00025	0.00025	0.00000	0.00084	0.00076	0.03930	0.00079	0.00079	0.00000
analgesics_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.02784	0.00000	0.00000	0.00000	0.00000
antidepressant	CAT	0.00050	0.00050	0.00000	0.00084	0.00242	0.01402	0.00071	0.00071	0.00000
antidepressant_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.03538	0.00707	0.00000	0.00000	0.00000
antidiabetic	CAT	0.00005	0.00005	0.00000	0.00000	0.00031	0.01662	0.00661	0.00661	0.00000
antidiabetic_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
band_size_UK_inch	NUM	0.00000	0.00000	0.00000	0.00000	0.02131	0.50857	0.00000	0.00000	0.00000
BED_boost_Gy	NUM	0.00000	0.00000	0.00000	0.00000	0.02415	0.04063	0.00000	0.00000	0.00000
BED_Breast_Gy	NUM	0.02970	0.02970	0.00000	0.07932	0.12273	0.25004	0.04354	0.04354	0.00000
BED_total_Gy	NUM	0.01495	0.01495	0.00000	0.05387	0.17604	0.23385	0.01529	0.01529	0.00000
blood_pressure	CAT	0.00132	0.00132	0.00000	0.01372	0.00086	0.05213	0.00002	0.00002	0.00000
boost	CAT	0.00262	0.00262	0.00000	0.00778	0.00226	0.00611	0.00357	0.00357	0.00000
boost_frac	NUM	0.00737	0.00000	-0.00737	0.00000	0.07024	0.14193	0.01035	0.01527	0.00492
bra_cup_size	NUM	0.01383	0.01406	0.00024	0.00000	0.05494	0.46227	0.00000	0.00000	0.00000
breast_cancer_family_history_1st_degree	CAT	0.00001	0.00001	0.00000	0.00012	0.00013	0.04822	0.00347	0.00345	-0.00001
breast_separation_cm	NUM	0.00903	0.00903	0.00000	0.00000	0.03786	0.51206	0.00000	0.00000	0.00000
carboplatin_chemo_drug	CAT	0.00031	0.00031	0.00000	0.00000	0.00098	0.00721	0.00008	0.00008	0.00000
chemotherapy	CAT	0.00005	0.00005	0.00000	0.00621	0.00003	0.03693	0.00020	0.00020	0.00000
combined_chemo_drugs	CAT	0.01366	0.01366	0.00000	0.05236	0.05304	0.09239	0.02102	0.02102	0.00000
cyclophosphamide_chemo_drug	CAT	0.00031	0.00031	0.00000	0.00838	0.00047	0.02510	0.00000	0.00000	0.00000
depression	CAT	0.00046	0.00046	0.00000	0.00181	0.00283	0.01370	0.00024	0.00024	0.00000
depression_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.03309	0.00000	0.00000	0.00000	0.00000
diabetes	CAT	0.00001	0.00001	0.00000	0.00103	0.00003	0.02156	0.00763	0.00762	-0.00001
diabetes_duration_yrs	NUM	0.01067	0.00000	-0.01067	0.00000	0.00646	0.00000	0.00610	0.00763	0.00152
docetaxel_chemo_drug	CAT	0.00064	0.00064	0.00000	0.01099	0.00328	0.00682	0.00037	0.00037	0.00000
doxorubicin_chemo_drug	CAT	0.00252	0.00252	0.00000	0.00000	0.00753	0.03255	0.00043	0.00043	0.00000
education_profession	CAT	0.00215	0.00391	0.00176	0.03741	0.01803	0.01005	0.00175	0.00463	0.00288
epirubicin_chemo_drug	CAT	0.00106	0.00106	0.00000	0.01359	0.00177	0.02509	0.00069	0.00069	0.00000
eribulin_chemo_drug	CAT	0.00055	0.00055	0.00000	0.00000	0.00161	0.00215	0.00152	0.00152	0.00000
ethnicity	CAT	0.00571	0.00570	0.00000	0.03271	0.02589	0.02189	0.00509	0.00508	-0.00001

grade_invasive	CAT	0.00187	0.00228	0.00041	0.00971	0.01402	0.02199	0.00246	0.00226	-0.00020
height_cm	NUM	0.00000	0.00000	0.00000	0.00000	0.08316	0.44196	0.00000	0.00000	0.00000
histology	CAT	0.00234	0.00237	0.00003	0.01183	0.01176	0.08485	0.00057	0.00060	0.00003
history_of_heart_disease	CAT	0.00354	0.00353	-0.00001	0.01157	0.00952	0.03197	0.00127	0.00127	0.00000
history_of_heart_disease_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.02948	0.03102	0.00000	0.00000	0.00000
hormone_replacement_therapy	CAT	0.00029	0.00066	0.00037	0.00910	0.00257	0.05089	0.00037	0.00029	-0.00008
household_income	CAT	0.00356	0.00703	0.00347	0.04210	0.01992	0.06340	0.00351	0.00408	0.00057
household_members	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.44358	0.00000	0.00000	0.00000
hypertension	CAT	0.00132	0.00132	0.00000	0.01372	0.00086	0.05213	0.00002	0.00002	0.00000
hypertension_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.17291	0.00509	0.00000	-0.00509
menopausal_status	CAT	0.00237	0.00231	-0.00006	0.01637	0.01302	0.03152	0.00246	0.00138	-0.00108
methotrexate _chemo_drug	CAT	0.00025	0.00025	0.00000	0.00130	0.00479	0.00308	0.00074	0.00008	-0.00066
monopause_age_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.03857	0.41090	0.00010	0.00000	-0.00010
n_stage	CAT	0.00525	0.00545	0.00020	0.02052	0.02645	0.05619	0.00000	0.00059	0.00059
on_statin	CAT	0.00644	0.00644	0.00000	0.00691	0.01914	0.06728	0.00057	0.00602	0.00545
on_statin_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.06413	0.04844	0.00127	0.00000	-0.00127
other_antihypertensive_drug	CAT	0.00145	0.00145	0.00000	0.01611	0.00160	0.03251	0.00000	0.00000	0.00000
other_antihypertensive_drug_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.01555	0.15670	0.00037	0.00000	-0.00037
other_collagen_vascular_disease	CAT	0.00096	0.00096	0.00000	0.00000	0.00430	0.00376	0.00351	0.00013	-0.00338
other_collagen_vascular_disease_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00430	0.00376	0.00000	0.00000	0.00000
other_lipid_lowering_drugs	CAT	0.00104	0.00104	0.00000	0.00742	0.00124	0.00045	0.00002	0.00277	0.00276
other_lipid_lowering_drugs_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.01449	0.00811	0.00000	0.00000	0.00000
paclitaxel_chemo_drug	CAT	0.00006	0.00006	0.00000	0.00056	0.00336	0.05403	0.00015	0.00015	0.00000
pegfilgrastim_chemo_drug	CAT	0.00055	0.00055	0.00000	0.00523	0.00322	0.00215	0.00008	0.00027	0.00020
Pertuzumab_chemo_drug	CAT	0.00027	0.00027	0.00000	0.00000	0.00000	0.00107	0.00144	0.00037	-0.00107
radio_axillary_levels	NUM	0.00000	0.00000	0.00000	0.00000	0.04207	0.05464	0.00000	0.00000	0.00000
radio_axillary_other	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
radio_bolus	CAT	0.00001	0.00001	0.00000	0.00078	0.00001	0.00575	0.00000	0.00000	0.00000
radio_boost_diameter_cm	NUM	0.00774	0.00918	0.00143	0.00000	0.03798	0.09225	0.00000	0.00000	0.00000
radio_boost_fractions	NUM	0.00000	0.00824	0.00824	0.06593	0.04896	0.15592	0.00000	0.01748	0.01748
radio_boost_sequence	CAT	0.00857	0.00857	0.00000	0.01071	0.01516	0.07039	0.00436	0.00436	0.00000
radio_boost_type	CAT	0.01700	0.01700	0.00000	0.08043	0.04059	0.06648	0.01575	0.01575	0.00000
radio_breast_ct_volume_cm3	NUM	0.02000	0.02047	0.00048	0.06228	0.19793	0.10627	0.00000	0.00000	0.00000
radio_breast_delineation	CAT	0.00027	0.00027	0.00000	0.00000	0.00107	0.00107	0.00059	0.00059	0.00000
radio_breast_dose_Gy	NUM	0.01966	0.01966	0.00000	0.07054	0.08518	0.28445	0.02210	0.02210	0.00000
radio_breast_fractions	NUM	0.01813	0.01813	0.00000	0.06984	0.07038	0.31260	0.02926	0.02926	0.00000
radio_breast_fractions_dose_per_fract_Gy	NUM	0.02204	0.02204	0.00000	0.07547	0.10130	0.26556	0.02415	0.02415	0.00000
radio_breast_fractions_per_week	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.01054	0.00000	0.00000	0.00000
radio_elec_boost_dose_Gy	NUM	0.00000	0.00000	0.00000	0.00000	0.02557	0.08132	0.00000	0.00000	0.00000
radio_elec_boost_field_x_cm	NUM	0.00000	0.00000	0.00000	0.00000	0.04908	0.16020	0.00000	0.00000	0.00000
radio_elec_boost_field_y_cm	NUM	0.00000	0.00000	0.00000	0.00000	0.02164	0.16766	0.00000	0.00000	0.00000
radio_elec_energy_MeV	NUM	0.01686	0.01686	0.00000	0.00000	0.04548	0.08072	0.00000	0.00000	0.00000
radio_heart_mean_dose_Gy	NUM	0.00000	0.00000	0.00000	0.00000	0.07873	0.09566	0.00000	0.00000	0.00000
radio_hot_spots	CAT	0.00211	0.00214	0.00003	0.00152	0.00515	0.00655	0.00009	0.00010	0.00001
radio_imrt	CAT	0.00848	0.00843	-0.00005	0.04575	0.02009	0.08996	0.02141	0.02127	-0.00014
radio_interrupted	CAT	0.00002	0.00002	0.00000	0.01050	0.00017	0.00762	0.00057	0.00057	0.00000
radio_interrupted_days	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
radio_ipsilateral_lung_mean_Gy	NUM	0.00000	0.00000	0.00000	0.00000	0.10337	0.05360	0.00000	0.00000	0.00000
radio_photon_2nd	CAT	0.01060	0.01060	0.00000	0.01690	0.02592	0.01454	0.01341	0.01341	0.00000
radio_photon_2nd_dose_fract_per_wk	NUM	0.01127	0.01127	0.00000	0.00000	0.03197	0.03582	0.01363	0.01363	0.00000

radio_photon_2nd_dose_MV	NUM	0.01843	0.01843	0.00000	0.05581	0.06771	0.12095	0.02328	0.02328	0.00000
radio_photon_2nd_dose_per_fract_Gy	NUM	0.01228	0.01228	0.00000	0.00000	0.09747	0.05150	0.01629	0.01629	0.00000
radio_photon_2nd_fractions	NUM	0.02037	0.02037	0.00000	0.00000	0.06359	0.07346	0.02186	0.02186	0.00000
radio_photon_boost_dose_per_fract_Gy	NUM	0.00000	0.00000	0.00000	0.04376	0.02956	0.15682	0.00000	0.00000	0.00000
radio_photon_boost_fractions	NUM	0.00737	0.00000	-0.00737	0.00000	0.07024	0.20287	0.01035	0.01527	0.00492
radio_photon_boost_fractions_per_week	NUM	0.00800	0.01066	0.00267	0.05360	0.02002	0.06328	0.01042	0.01330	0.00287
radio_photon_boost_volume_cm3	NUM	0.01033	0.01574	0.00541	0.05411	0.13251	0.12075	0.00000	0.00000	0.00000
radio_photon_boostdose_Gy	NUM	0.00000	0.00000	0.00000	0.00000	0.05234	0.13049	0.00000	0.00000	0.00000
radio_photon_boostdose_precise_Gy	NUM	0.00000	0.00000	0.00000	0.00000	0.02963	0.14553	0.00991	0.01182	0.00191
radio_photon_dose_MV	NUM	0.00000	0.00000	0.00000	0.00000	0.01107	0.00000	0.00000	0.00000	0.00000
radio_photon_energy_MV or kV	NUM	0.00970	0.00965	-0.00006	0.07597	0.02628	0.12818	0.02097	0.02097	0.00000
radio_skin_max_dose_Gy	NUM	0.03073	0.03088	0.00015	0.14315	0.19629	0.12209	0.02948	0.02912	-0.00035
radio_supraclavicular_fossa	CAT	0.00027	0.00027	0.00000	0.00020	0.00368	0.04354	0.00115	0.00115	0.00000
radio_treated_breast	CAT	0.00159	0.00159	0.00000	0.01542	0.00618	0.10882	0.00023	0.00023	0.00000
radio_treatment_pos	CAT	0.00396	0.00396	0.00000	0.01001	0.01182	0.06438	0.00094	0.00093	-0.00001
radio_type_imrt	CAT	0.01754	0.01749	-0.00005	0.08163	0.04062	0.12413	0.02651	0.02637	-0.00014
radiotherapy_toxicity_family_history	CAT	0.00047	0.00045	-0.00002	0.00078	0.00505	0.01303	0.00001	0.00002	0.00002
rheumatoid arthritis	CAT	0.00007	0.00007	0.00000	0.00742	0.00127	0.01021	0.00002	0.00002	0.00000
rheumatoid arthritis_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00918	0.00000	0.00000	0.00000	0.00000
smoker	CAT	0.00145	0.00132	-0.00013	0.00239	0.00650	0.09127	0.00140	0.00146	0.00006
smoking_status	CAT	0.00059	0.00059	0.00000	0.00204	0.00364	0.04721	0.00015	0.00015	0.00000
smoking_duration_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.01408	0.01982	0.00000	0.00000	0.00000
smoking_time_since_quitting_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.13645	0.00000	0.00000	0.00000
surgery_type	CAT	0.00105	0.00105	0.00000	0.00000	0.00574	0.00344	0.00155	0.00155	0.00000
systemic_lupus_erythematosus	CAT	0.00027	0.00027	0.00000	0.00000	0.00000	0.00107	0.00027	0.00027	0.00000
systemic_lupus_erythematosus_yrs	NUM	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
t_stage	CAT	0.00454	0.00446	-0.00008	0.01970	0.01723	0.12975	0.00806	0.00815	0.00008
TAM	CAT	0.00118	0.00108	-0.00010	0.00661	0.00000	0.07888	0.00291	0.00269	-0.00022
tobacco_product	CAT	0.00764	0.00030	-0.00734	0.00074	0.00145	0.03533	0.00061	0.00072	0.00011
tobacco_products_per_day	NUM	0.00000	0.00000	0.00000	0.00000	0.05251	0.00000	0.00000	0.00000	0.00000
trastuzumab_chemo_drug	CAT	0.00166	0.00166	0.00000	0.00000	0.00700	0.00646	0.00010	0.00010	0.00000
tumour_size_mm	NUM	0.00000	0.00000	0.00000	0.00000	0.04472	0.02545	0.00000	0.00000	0.00000
weight_at_cancer_diagnosis_kg	NUM	0.01264	0.01382	0.00117	0.06476	0.13548	0.13946	0.00000	0.00000	0.00000

Table B. Feature Importance of Cost-Sensitive RF Model's with MDI (Pre-simplification)

Model's Features	MDI	Model's Features	MDI
5-fluorouracil (5-FU)_chemo_drug	0.37	radio_photon_2nd_dose_MV	0.19
radio_imrt	0.35	analgesics	0.19
ace_inhibitor	0.34	radio_photon_2nd_dose_fractions_per_week	0.19
Smoking	0.32	radio_interrupted_days	0.19
chemotherapy_performed	0.32	surgery_type	0.19
docetaxel_chemo_drug	0.32	radio_breast_fractions_dose_per_fraction_Gy	0.18
other_antihypertensive_drug	0.31	alcohol_intake	0.18
tumour_size_mm	0.30	radio_photon_boostdose_precise_Gy	0.18
radio_treated_breast	0.30	radio_elec_boost_dose_Gy	0.18
grade_invasive	0.29	tobacco_product	0.18
histology	0.28	radio_treatment_pos	0.18
tobacco_products_per_day	0.28	radio_photon_2nd	0.18
Band_size_UK	0.27	combined_chemo_drugs	0.17
monopause_age_yrs	0.27	household_income	0.17
boost	0.27	radio_elec_boost_field_y_cm	0.17
epirubicin_chemo_drug	0.27	radio_photon_boost_fractions	0.17
radio_axillary_other	0.27	radio_boost_diameter_cm	0.17
radio_breast_ct_volume_cm3	0.26	radio_supraclavicular_fossa	0.17
radio_heart_mean_dose_Gy	0.26	antidepressant	0.17
BED_breast	0.26	radio_breast_fractions	0.16
TAM	0.26	radio_elec_boost_field_x_cm	0.16
radio_hot_spots_107	0.26	doxorubicin_chemo_drug	0.16
breast_separation	0.25	radio_boost_type	0.15
t_stage	0.25	radio_elec_energy_MeV	0.15
smoking_time_since_quitting_yrs	0.25	radio_photon_energy_MV or kV	0.15
blood_pressure	0.25	diabetes	0.15
cyclophosphamide_chemo_drug	0.25	carboplatin_chemo_drug	0.15
rheumatoid_arthritis_duration_yrs	0.25	depression_duration_yrs	0.14
methotrexate_chemo_drug	0.25	depression	0.13
boost_fractions	0.24	ace_inhibitor_duration_yrs	0.13
alcohol_previous_consumption	0.24	radiotherapy_toxicity_family_history	0.13
radio_skin_max_dose_Gy	0.23	other_lipid_lowering_drugs	0.13
radio_ipsilateral_lung_mean_Gy	0.23	antidiabetic	0.13
height_cm	0.23	radio_axillary_levels	0.12
alcohol_current_consumption	0.23	Ethnicity	0.12
radio_photon_boost_volume_cm3	0.23	radio_photon_2nd_fractions	0.12
n_stage	0.23	analgesics_duration_yrs	0.11
BED_boost	0.23	on_statin	0.11
radio_photon_boostdose_Gy	0.23	radio_photon_boost_fractions_per_week	0.11
hypertension_duration_yrs	0.23	diabetes_duration_yrs	0.11
smoker	0.22	trastuzumab	0.11
menopausal_status	0.22	radio_photon_2nd_dose_per_fraction_Gy	0.10
BED_total	0.21	antidepressant_duration_yrs	0.10

smoking_duration_yrs	0.21	radio_breast_fractions_per_week	0.10
radio_type_imrt	0.21	radio_boost_sequence	0.08
radio_boost_fractions	0.21	on_statin_duration_yrs	0.08
hypertension	0.21	history_of_heart_disease_duration_yrs	0.07
paclitaxel	0.21	radio_bolus	0.07
hormone_replacement_therapy	0.21	radio_interrupted	0.07
weight_at_cancer_diagnosis_kg	0.20	history_of_heart_disease	0.06
age_at_radiotherapy_start_yrs	0.20	antidiabetic_duration_yrs	0.04
bra_cup_size	0.20	pegfilgrastim	0.03
education_profession	0.20	other_collagen_vascular_disease	0.02
breast_cancer_family_history_1st_degree	0.20	systemic_lupus_erythematosus_duration_yrs	0.00
radio_photon_dose_MV	0.20	systemic_lupus_erythematosus	0.00
other_lipid_lowering_drugs_duration_yrs	0.20	radio_breast_delineation	0.00
rheumatoid_arthritis	0.20	pertuzumab_chemo_drug	0.00
radio_breast_dose_Gy	0.19	other_collagen_vascular_disease_duration_yrs	0.00
household_members	0.19	eribulin_chemo_drug	0.00
other_antihypertensive_drug_duration_yrs	0.19	amiodarone_duration_yrs	0.00
radio_photon_boost_dose_per_fraction_Gy	0.19	amiodarone	0.00

Table C. Feature Importance of the simplified cost-sensitive RF model with MDI

Model's Feature	MDI	Model's Feature	MDI
other_lipid_lowering_drugs_duration_yrs	0.52	alcohol_current_consumption	0.20
surgery_type	0.41	smoking_time_since_quitting_yrs	0.20
radio_bolus	0.40	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
band_size_UK	0.30	radio_photon_2nd	0.19
radio_treated_breast	0.30	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
TAM	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_dose_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	0.20	other_collagen_vascular_disease	0.03
rheumatoid arthritis_duration_yrs	0.20	rheumatoid arthritis	0.02