A Data Science Approach for Early-Stage Prediction of Patient's Susceptibility to Acute Side Effects of Advanced Radiotherapy

Aldraimli, M., Soria, D., Grishchuck, D., Ingram, S., Lyon, R., Mistry, A., Oliveira, J., Samuel, R., Shelley, L.E.A., Osman, S., Dwek, M., Azria, D., Chang-Claude, J., Gutiérrez-Enríquez, S., De Santis, M.C., Rosenstein, B.S., De Ruysscher, D., Sperk, E., Symonds, R.P., Stobart, H., Vega, A., Veldeman, L., Webb, A, Christopher, J.T., West, C.M., Rattay, T., REQUITE consortium and Chaussalet, T.J.

A Data Science Approach for Early-Stage Prediction of Patient's Susceptibility to Acute Side Effects of Advanced Radiotherapy

Mahmoud Aldraimli[1], Daniele Soria[2], Diana Grishchuck[3], Samuel Ingram[4], Robert Lyon[5], Anil Mistry[6], Jorge Oliveira[7], Robert Samuel[8], Leila E.A. Shelley[9], Sarah Osman[10], Miriam V. Dwek[11], David Azria[12], Jenny Chang-Claude[13], Sara Gutiérrez-Enríquez[14], Maria Carmen De Santis[15], Barry S Rosenstein[16], Dirk De Ruysscher[17], Elena Sperk[18], R Paul Symonds[19], Hilary Stobart[20], Ana Vega[21], Liv Veldeman[22], Adam Webb[23], Christopher J. Talbot[24], Catharine M. West[25], Tim Rattay[24], REQUITE consortium 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] Edinburgh Cancer Centre, Western General Hospital, Crewe Road South, Edinburgh, 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] University of Montpellier, France
[13] German Cancer Research Center (DKFZ) Division of Cancer Epidemiology, Unit of Genetic Epidemiology, Heidelberg, Germany
[14] Vall d'Hebron Institute of Oncology, Barcelona, Spain
[15] Dept of Radiation Oncology 1, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[16] Prostate Cancer Program, Mount Sinai School of Medicine, New York, USA
[17] Maastricht Radiation Oncology (MAASTRO Clinic) University Hospital Maastricht, The Netherlands
[18] Department of Radiation Oncology, University Medical Center Mannheim, Medical Faculty Mannheim, Heidelberg University, Germany
[19] Department of Oncology, Leicester Royal Infirmary, UK
[20] Independent Cancer Patients' Voice, London, UK.
[21] Fundación Publica Galega Medicina Xenomica, Santiago de Compostela, Spain
[22] Department of Basic Medical Sciences, University Hospital Ghent, Belgium
[23] Department of Genetics and Genome Biology, University of Leicester, UK.
[24] Cancer Research Centre, University of Leicester, Leicester, UK
[25] Institute of Cancer Sciences, Christie Hospital, Wilmslow Road, Manchester, UK

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
w1654353@my.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 class imbalance problem 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 and cost-sensitive learning methods on the models when dealing with class imbalance. The combinations of such techniques used had a significant impact on the classification. They resulted in an improvement in incidence status prediction by shifting classifiers' attention to the minority group. The best classification model for RT acute toxicity prediction was identified based on domain experts' success criteria. The Area Under Receiver Operator Characteristic curve of the models tested with an isolated dataset ranged from 0.50

to 0.77. The scale of improved results is promising and will guide further development of models to predict RT acute toxicities. One model was optimised and found to be beneficial to identify patients who are at risk of developing acute RT early-stage toxicities as a result of undergoing breast RT ensuring relevant treatment interventions can be appropriately targeted. The design of the approach presented in this paper resulted in producing a preclinical-valid prediction model. The study was 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.; SMOTE; Meta-Learning; Desquamation.

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.

Data and material availability statement

Research data are stored in an institutional repository and will be shared upon request to the corresponding author and the REQUITE consortium.

Software availability

The workbench for machine learning Waikato Environment for Knowledge Analysis is open source machine learning software licensed under the GNU General Public License and can be accessed through a graphical user interface, standard terminal applications, or a Java API. Version 3.8 used for research and industrial applications. The workbench contains a plethora of built-in tools for standard machine learning tasks, and additionally gives transparent access to toolboxes such as Python SciKit-learn and R for machine learning and statistical analysis.

Consent for publication

Publication consent was obtained from the REQUITE Publications Committee that comprises members of the Steering Committee, a clinical representative from each country, the project manager, observational study manager, database manager, the Patient Advisory Group and any active member of REQUITE Consortium is eligible to serve on the Publications Committee.

Authors' contributions

All authors contributed to the study conception and design. **Data collection and on behalf of the REQUITE consortium:** David Azria1, Jenny Chang-Claude, Sara Gutiérrez-Enríquez, Maria Carmen De Santis, Barry S Rosenstein, Dirk De Ruysscher, Elena Sperk, R Paul Symonds, Hilary Stobart, Ana Vega, Liv Veldeman, Adam Webb, Christopher J. Talbot, Catharine M. West and Tim Rattay. **Conceptualisation**: Mahmoud Aldraimli, Tim Rattay, Sarah Osman, Diana Grishchuck, Samuel Ingram, Robert Lyon, Anil Mistry, Jorge Oliveira, Robert Samuel, Leila E.A. Shelley and Robert Lyon. **Methodology Design:** Mahmoud Aldraimli **Data cleaning and Pre-processing**: Mahmoud Aldraimli, Diana Grishchuck, Anil Mistry, Sarah Osman, Samuel Ingram, Robert Lyon, Jorge Oliveira, Robert Samuel and Tim Rattay. **Formal modelling:** Mahmoud Aldraimli. **Analysis and interpretation** were carried out by Mahmoud Aldraimli, Tim Rattay, Sarah Osman, Diana Grishchuck, Samuel Ingram, Robert Lyon, Anil Mistry, Jorge Oliveira, Robert Samuel, Leila E.A. Shelley and Robert Lyon, Daniele Soria, Miriam V. Dwek and Thierry Chaussalet. **Models Evaluation**: Mahmoud Aldraimli, Sarah Osman, Robert

Lyon, Jorge Oliveira, Diana Grishchuck, Anil Mistry, Samuel Ingram, Robert Samuel and Tim Rattay. **Writing original draft:** Mahmoud Aldraimli. And Finally, **Critical review and editing:** Daniele Soria, Thierry Chaussalet**,** Diana Grishchuck, Samuel Ingram, Robert Lyon, Anil Mistry, Jorge Oliveira, Robert Samuel, Leila E.A. Shelley, Sarah Osman, Miriam V. Dwek, David Azria1, Jenny Chang-Claude, Sara Gutiérrez-Enríquez, Maria Carmen De Santis, Barry S Rosenstein, Dirk De Ruysscher, Elena Sperk, R Paul Symonds, Hilary Stobart, Ana Vega, Liv Veldeman, Adam Webb, Christopher J. Talbot, Catharine M. West and Tim Rattay. **All authors read and approved the final manuscript.**

Funding Statement

This research collaboration was formed by the UK Radiotherapy Machine Learning Network (RTML), funded through the Advanced Radiotherapy 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. The research 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.

Dr Alison M. Dunning was supported by Cancer Research-UK C8197/A16565.

Dr Sara Gutiérrez-Enríquez is supported by the ISCIII Miguel Servet II Program (CP16/00034).

Dr Tim Rattay is currently an NIHR Clinical Lecturer. He was previously funded by a National Institute of Health Research (NIHR) Doctoral Research Fellowship (DRF 2014-07-079). This publication presents independent research funded by the NIHR. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Dr Leila Shelley reports grants from Chief Scientist Office (CSO) Scotland grant (TCS/17/26 - CSO Award).

Dr Elena Sperk was previously supported by the Ministry of Science and Arts of the State of Baden-Württemberg (2017-19) through the Brigitte-Schlieben-Lange-Programme.

Dr Ana Vega is supported by Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds (INT15/00070, INT16/00154, INT17/00133, INT20/00071, PI19/01424, PI16/00046, PI13/02030, PI10/00164), and through the Autonomous Government of Galicia (Consolidation and structuring program: IN607B).

Prof Catharine West is supported by Cancer Research UK (C1094/A18504, C147/A25254) and by the NIHR Manchester Biomedical Research Centre.

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]. For example, 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 and 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]. When models are built with such strategies, careful consideration of performance metrics assessment must be taken into account. The evaluation of the models' performances is likely to require input from the domain experts.

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 technique 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 realm, defining such a cost is challenging [15]. For example, in treatment management scenarios, the cost of a false positive might be derived from the monetary cost of performing subsequent tests. In contrast, there is no monetary equivalent cost for administering treatment on a patient and getting further health complications.

This paper presents a new comparative ML classification approach and new toxicity prediction models, including a simple clinically valid classification model assessed by a UK national collaboration of data scientists, medical physicists, oncologists, and surgeons in the UK Radiotherapy Machine Learning Network. The models predict breast cancer patients' susceptibility to early-stage radiotherapy skin toxicity (acute desquamation). Findings in this study are considered confirmatory; hence carefully describes the data preprocessing techniques, algorithmic modifications, and evaluation metrics, including those to account for data quality and imbalance. The new models were built with both sampled and unsampled datasets using Random Under Sampled (RUS) [5], Synthetic Minority Over-sampling Technique (SMOTE) [5], Random Over-Sampling (ROS) [5] and Cost-Sensitive Classification techniques [6] [7] as well as the original highly imbalanced training data (ITD). The study then takes a systematic comparative approach to compare eighty-nine parametric and non-parametric models built with eight classification algorithms. Finally, this study suggests the most suitable model meeting the domain experts' success criteria proven to be of particular interest to cancer radiotherapy clinicians. The data imbalance characteristic causing the transition in classifier training performance was captured visually by Adaptive Projection Analysis (APA) [16] and numerically via Information Gain (IG) attribute evaluation [17].

The deployment of ML modelling in this study aims to give researchers a new multi-stage approach to effectively 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 buildup under the skin, impacting their Quality of Life (QoL) [18]. RT-treated patients' QoL has become an increasingly important research priority [18]. RT reduces

the rates of cancer recurrence and increases long-term survival. Hence over 70% of breast cancer patients receive RT during the course of their treatment [19]. Typically, the incidence rate of acute desquamation range between 11% to 71% in breast cancer RT patients [18].

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. 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 [20]. Such models are based on a simplified characterization 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 [21]. Such an approach makes large scale analysis of RT toxicities limited 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 [22]. A different approach utilized hundreds of clinical variables as model inputs raising an issue in interpretability [23].

The REQUITE study provides a comprehensive means of assessing the relationship between the patients' 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 [24]. 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 study is to identify a simple and clinically valid ML prediction model to predict the occurrence of acute desquamation in the REQUITE breast cancer cohort. 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 multicenter breast cancer patients' cohort was recruited prospectively in seven European countries and the US. All patients gave written informed consent [25]. The study was approved by local ethics committees in participating countries and registered at the ISRCTN registry [26] (ISRCTN98496463). The study is a cross-sectional assessment of 2069 patients from the REQUITE international multicenter cohort, aged 23-80 and treated with breast RT between April 2014 and March 2017.

The paper is structured as follows: The methodology, techniques, algorithms and metrics of this study are presented in section 2. The results and analysis are documented in section 3, with the discussion and clinical next steps in sections 4 and 5, respectively.

2. Methodology

For RT complication prediction, 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. Out of the REQUITE cohort (n=2069, m>300), a final 2058 patients' records and 123 variables were deemed viable for modelling by RTML experts, hence retained. A randomly class-stratified sample without replacement (n=1029) of patients was used to train eight ML algorithms using 10-fold cross-validation with two different strategies. Finally, all the trained models were tested with the same isolated remaining patients' data (n=1029).

The original REQUITE dataset underwent a rigorous data preparation and pre-processing phase by the RTML network specialists (See Fig.1), followed by the modelling, evaluation and simplification phase (See Fig.3). 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 [27], Support Vector Machine [28], Logistic Regression [29], Artificial Neural Network [30], C4.5 Decision Tree [31], Logistic Model Tree [32], Random Forest [33] and K-Nearest Neighbour [35]). 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. The descriptive statistics of the REQUITE dataset variables are reported in a previous study [35].

Clinical ML 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 [36].

2.1 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. The RTML clinicians manually labelled all records for acute desquamation outcome based on the CTCAE v4.0 endpoint definition: grade 1 ≥ 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 ($Desq^{+} = 192, Desq^{-} = 1866$) 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 input variables used in this study are easily obtainable at 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.

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 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). The retained dataset for feature engineering transformation and modelling finally had m=123 variables and n=2058 records.

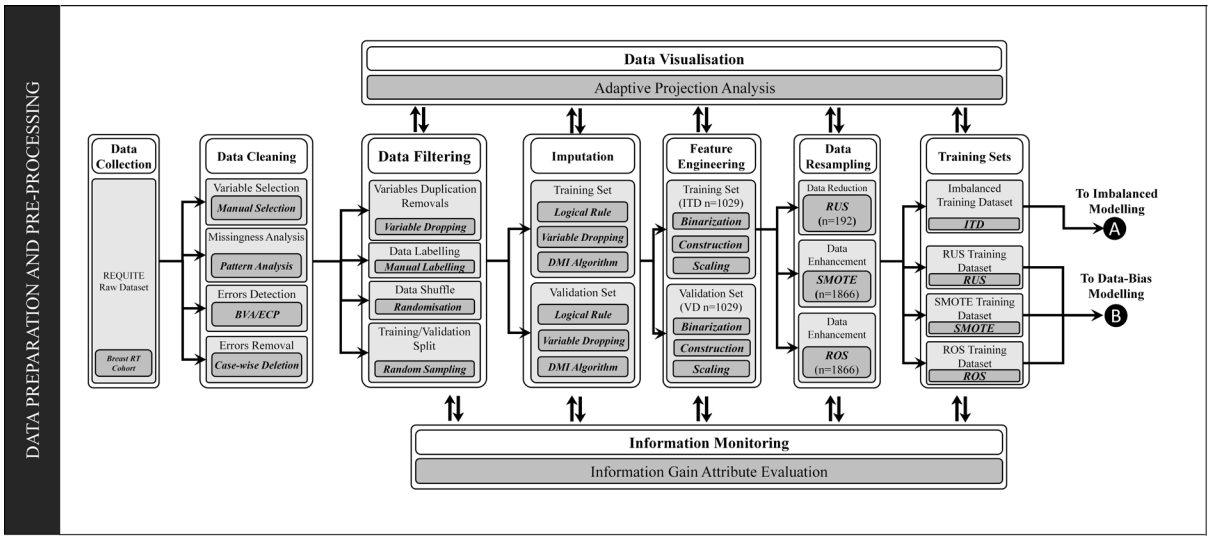

Fig. 1 Data preparation and pre-processing tasks used in this study

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. 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 Decision-Tree based Missing-Value Imputation (DMI) Algorithm [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 allows for a fair information bias comparison (see supplementary Information Gain Attribute Evaluation Table A).

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 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 [41]. 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 binarized [42] based on their generic chemical names (see Table 2). The chemotherapy drugs categorical values were converted to One-Hot Encoding, which is a format that could be provided to ML algorithms to improve prediction performance [43]. The categorical values represent the administered chemo-drug combinations in a chemotherapy regime. The combinations' values start from zero goes all the way up to N-1 categories. One-Hot encoding binarization is performed at a category level (single observation level per attribute), converting every chemo-drug used in a chemotherapy regime into a new feature.

Table 2. Illustration examples of binarized 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 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 [44]. NHS UK published a wide range of chemotherapy side effects that 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.

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 the slope of the cell survival curve for different irradiated tissues [46]. Three features were constructed by calculating the BED.

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

Out of all 123 variables, all numeric features (m=63) were normalised with $\mathcal{Z}$-score standardisation [47] to eliminate the impact of larger magnitudes variables when modelling with distance-based algorithms.

The REQUITE dataset shows that in a breast radiation treatment, only a small portion of patients suffered from acute desquamation [48], 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. Class imbalance is a common challenge in ML modelling [4]. One strategy 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) (n = 1866, $Desq^{+} = 933, Desq^{-} = 933$) and Synthetic Minority Oversampling Technique (SMOTE) (n = 1866, $Desq^{+} = 933, Desq^{-} = 933$). The effect of such resampling techniques on the training dataset was visualised with a multi-dimensional Adaptive Projection Algorithm (APA) [16] into a 3D point cloud.

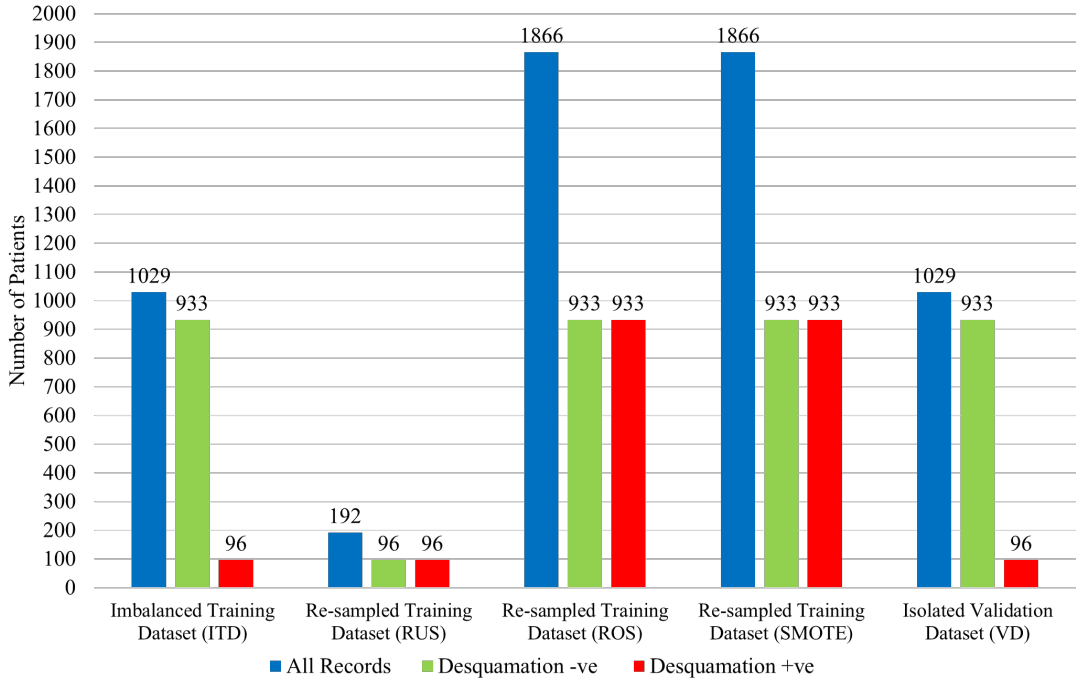

Fig. 2 Visualisation of samples size for ITD, RUS, ROS, SMOTE training datasets and test dataset VD

2.2 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. In order to verify the impact of the class imbalance on modelling, the training set (ITD) n=1029 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 test dataset (VD) n=1029. Both ITD and VD are equally imbalanced ($Desq^{+} = 96, Desq^{-} = 933$).

The resampled datasets RUS, ROS, and SMOTE, are used to train each ML algorithm. These algorithms are Discretised Naïve Bayes (NB) [27], Logistic Regression with Ridge Estimator (LR) [29], Artificial Neural Networks (ANN) with a multi-layer perceptron architecture [30], Support Vector Machine (SVM) with polynomial kernel and Logistic calibrator [28], K-Nearest Neighbour (KNN) [34] with K={1,3,5,7,9}, Decision Trees (C4.5) [31], Logistic Model Tree (LMT) [32] and Random Forest (RF) [56]. An alternative strategy 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). Three ML algorithms out of the competing eight were systematically selected for Cost-Sensitive Learning modelling.

A false negative prediction cost is not linked to a monetary value; instead, a ten-step Incremental Inverse Class Distribution cost was used [49]. ITD has a (96:933 ≅ 1:10) ratio of examples in the positive class to examples in the negative group. 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 [50].

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

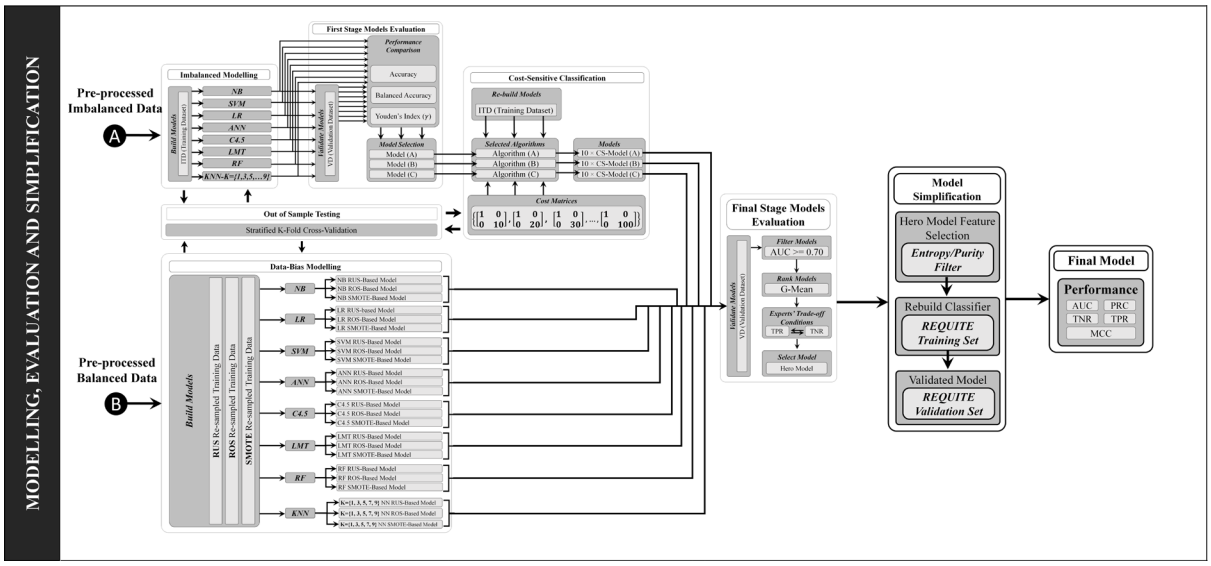

Fig. 3 Models Building, evaluation and simplification methodology used in this study

All algorithms used for this study were implemented in the Waikato Environment for Knowledge Analysis (WEKA) 3.8.3 (with the default algorithm's configuration settings), with the C4.5 decision tree using the J48 implementation, KNN using the IBK implementation and SVM using SMO implementation.

For models' assessments, a two-stage performance evaluation was applied. At the first stage of evaluation, all eight algorithms were trained with ITD and tested with VD. We used three performance metrics: Accuracy, Balanced Accuracy and Youden's index (γ) [51]. The measurements of these performance metrics show the most and the least class imbalance impact on all the eight classifiers. The top two severely impacted classifiers by class imbalance, and the bottom two affected the least were selected to undergo further improvement with cost-sensitive (CS) modelling.

At the final evaluation stage, all models built by incorporating resampling and CS strategies to tackle the class imbalance problem were compared. Multiple metrics were used: The Area Under Receiver Operator Characteristic Curve (AUC-ROC) [52], All models with AUC scores below 0.70 were discarded, the remaining models were then ranked with their Geometric Mean (G-Mean) [51]. The G-Mean measures the classifier avoidance of overfitting the negative class and underfitting the positive labels [51]. The ranked models' sensitivity (True Positive Rate TPR) and specificity (True Negative Rate TNR) were also compared, and the radiotherapy clinicians and physicists also determined a trade-off performance threshold. Models that met the TNR-TPR trade-off criteria were ranked based on their TPR performance, and a single model was nominated as a "Hero Model".

The clinical specialists made it clear that the requirement is to model with all carefully selected features to understand their impact and importance. Therefore, domain experts manually selected all the input variables to the models based on their empirical observations on patients who underwent radiotherapy and their correlation with acute desquamation occurrence in previously published studies (discussed in section 4).

The Hero Model was simplified further by utilising a purity filter to reduce the number of features to produce a "Final Model" [53]. The final model performance was reported in terms of AUC-ROC score, Precision-Recall Curve (AUC-PR) score [54] and Matthew Correlation Coefficient (MCC) [52] while highlighting any improvement in TPR and TNR.

3. Results analysis

3.1 Datasets visualisation interpretation

The APA visualisation [16] in Fig. 4 can indicate 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 and how resampling techniques are addressing the balance.

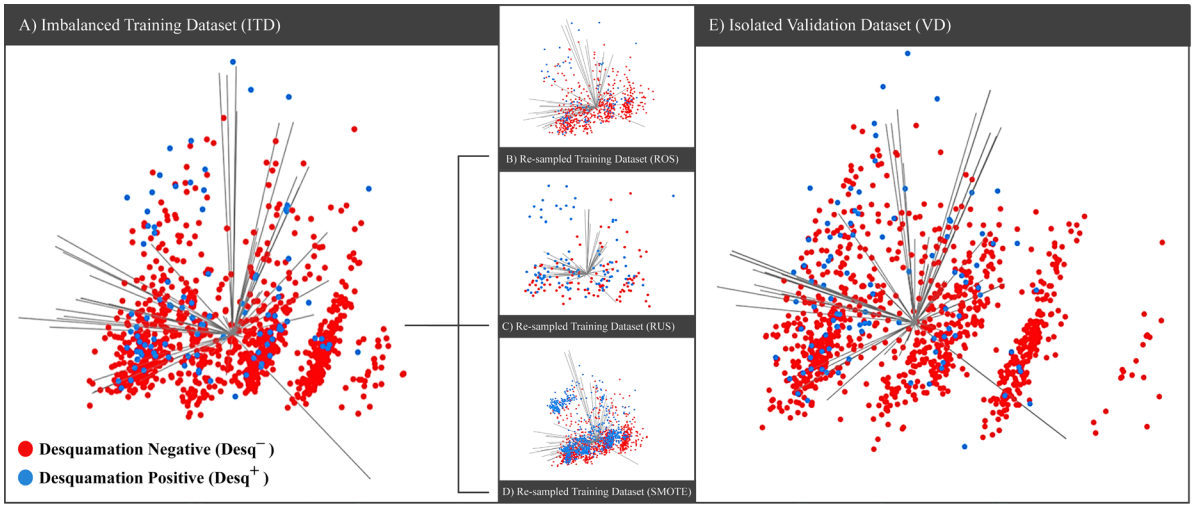

Fig. 4 APA visualisation of imputed ITD, RUS, ROS, SMOTE training datasets and test dataset

ROS training dataset shows somewhat widely scattered positive class records since the 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 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.

3.2 The Information Gain (IG) evaluations

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. 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 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 test datasets utterly independent from each other and entirely isolated.

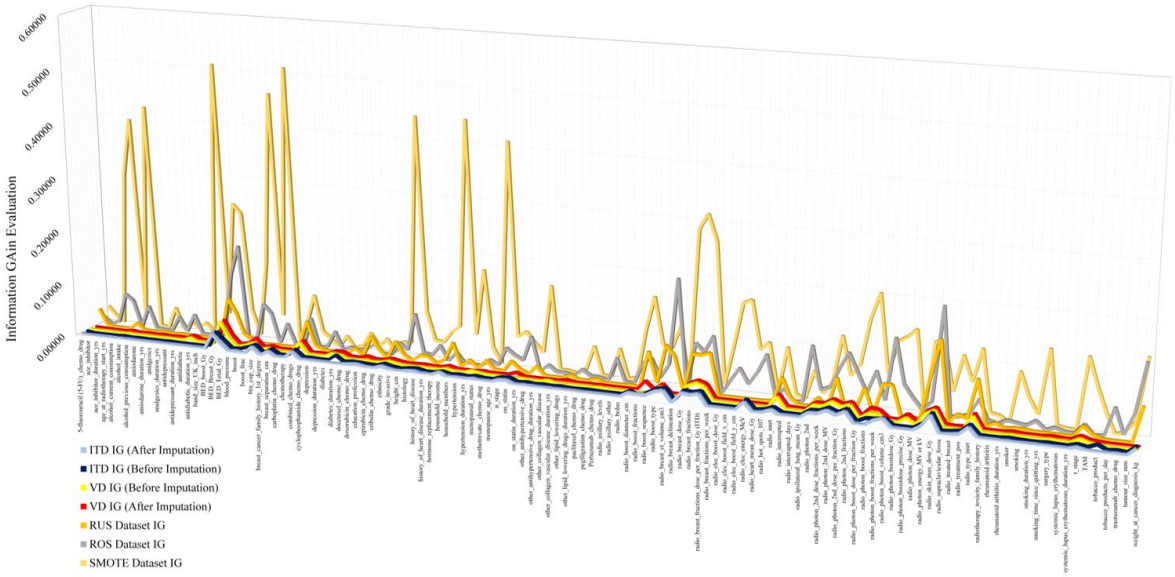

Fig. 5 IG evaluations of predictors in ITD, RUS, ROS, SMOTE training datasets and test dataset (VD)

3.3 The imbalanced models' 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} [55]. Table 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 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) [52]. 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 performances ranking was observed after 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 3, NB scored the highest in both balanced accuracy and Youden's index (γ) 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 (γ) 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 in misclassification avoidance in both classes proved with the lowest TPR of 0.01. RF, a famous ensemble algorithm in the data science community for its accomplishments, missed the lowest performance on both balanced accuracy metric and Youden's index (γ) and ranked just before LMT with 0.10 and 0.019, respectively, showing its severe bias towards the majority class.

Table 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

By analysing the training performances of the classifiers with ITD in table 3, the question of class importance in this particular domain problem arises when selecting algorithms for seeking further improvement with a CS strategy. The higher the balanced accuracy and Youden's index (γ), the higher degree of discrimination between both classes in the imbalanced setting. In contrast, the lowest measurements on both 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. 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 (γ) in training, i.e. LMT or RF, for CS modelling, may indicate caring about both classes equally. Since 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.

The RT potential benefits have to be weighed against the possibilities of causing damage to the healthy tissue, with the final aim of maximizing curative response while minimizing the probability of complications [56]. 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 both 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.

Domain experts 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 4 describe the numeric count of correctly classified patients, FP (type I) and FN (type II) errors misclassifications.

Table 4. Training and test confusion matrices of LMT, RF, ANN and NB imbalanced ML models

		LMT		RF		ANN		NB	
Training with ITD		*Predicted*		*Predicted*		*Predicted*		*Predicted*	
		Desq -ve	*Desq +ve*	*Desq -ve*	*Desq +ve*	*Desq -ve*	*Desq +ve*	*Desq -ve*	*Desq +ve*
	Actual *Desq -ve*	929	4	931	2	882	51	756	177
	Actual *Desq +ve*	95	1	94	2	77	19	54	42
Test with VD		**LMT**		**RF**		**ANN**		**NB**	
		Predicted		*Predicted*		*Predicted*		*Predicted*	
		Desq -ve	*Desq +ve*	*Desq -ve*	*Desq +ve*	*Desq -ve*	*Desq +ve*	*Desq -ve*	*Desq +ve*
	Actual *Desq -ve*	928	5	933	0	889	44	777	156
	Actual *Desq +ve*	92	4	95	1	79	17	48	48

3.4 The Cost-Sensitive (CS) classification 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 towards the positive group as there are no further improvements required for the negative class. Forty models were built with ITD accompanied by a defined cost matrix then tested with VD. In order to evaluate such an improvement, four metrics measurements in test were reported: the AUC-ROC, G-mean, TPR and TNR. (See table 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).

From table 5, the CS classification showed a consistent deterioration of TNR for all models across all four algorithms. ANN 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 from 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 Fig. 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.

Table 5. Selected ITD algorithms Test performance with CS classification strategy

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

Fig. 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, then a constant massive change rate for both TNR and TPR occurs. Fig. 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.

Fig. 6 Absolute change in TNR and TPR test performance per FN penalty in LMT, RF, NB and ANN models

In Fig. 6 – B, for FN penalties 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 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.

For FN penalties from 70 to 100 (Fig. 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.

Fig. 7 shows the AUC, G-Mean, TPR and TNR performance combinations for LMT, RF, NB and ANN for all incremental FN penalties. Fig. 7 – (A, B, C and 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 [54], 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.

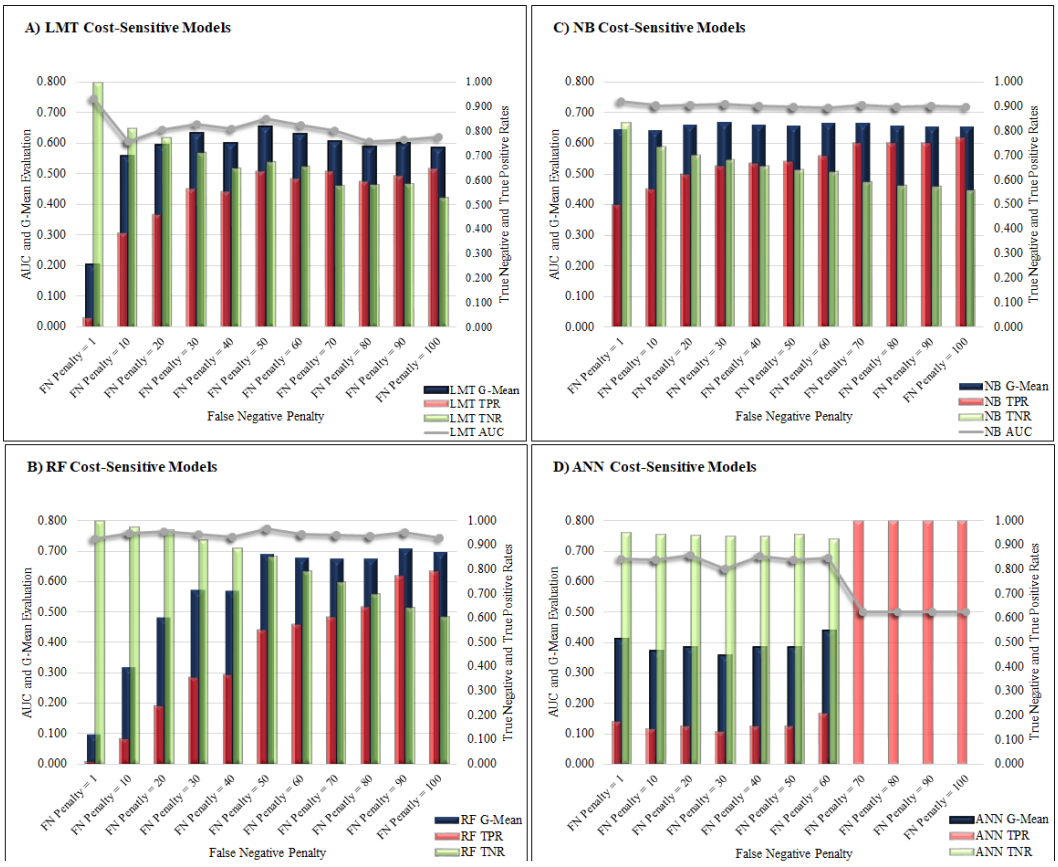

Fig. 7 TNR, TPR, G-Mean and AUC computations per FN penalty in LMT, RF, NB and ANN models

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. In Fig. 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 in Fig. 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.

By examining the G-Mean for all CS-models in Fig.7 and table 5, it is observed 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}.

3.5 Resampling models results

Table 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. By analysing the effect of resampling techniques on both TNR and TPR in Fig. 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. 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

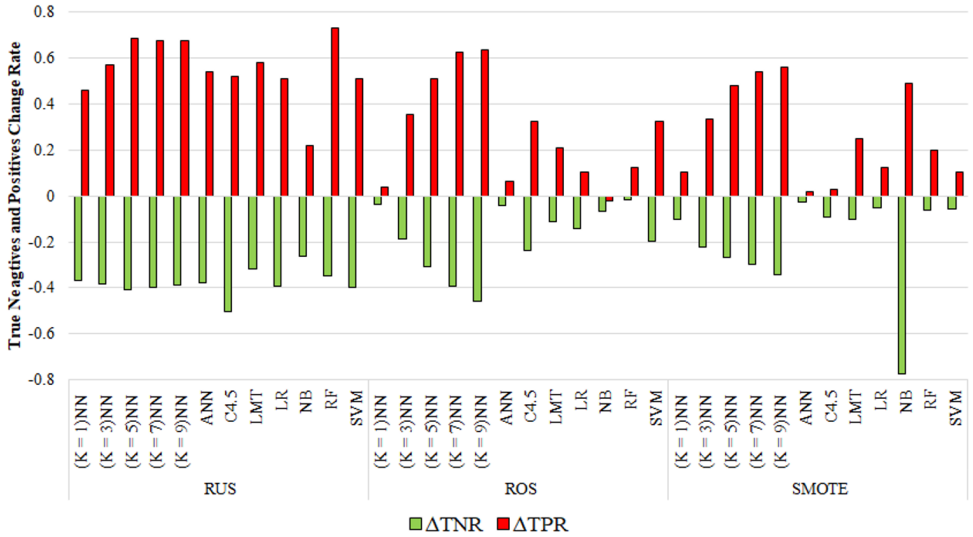

Fig. 8 TN and TP change rates in test per classifier in RUS, ROS and SMOTE models

Fig .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.

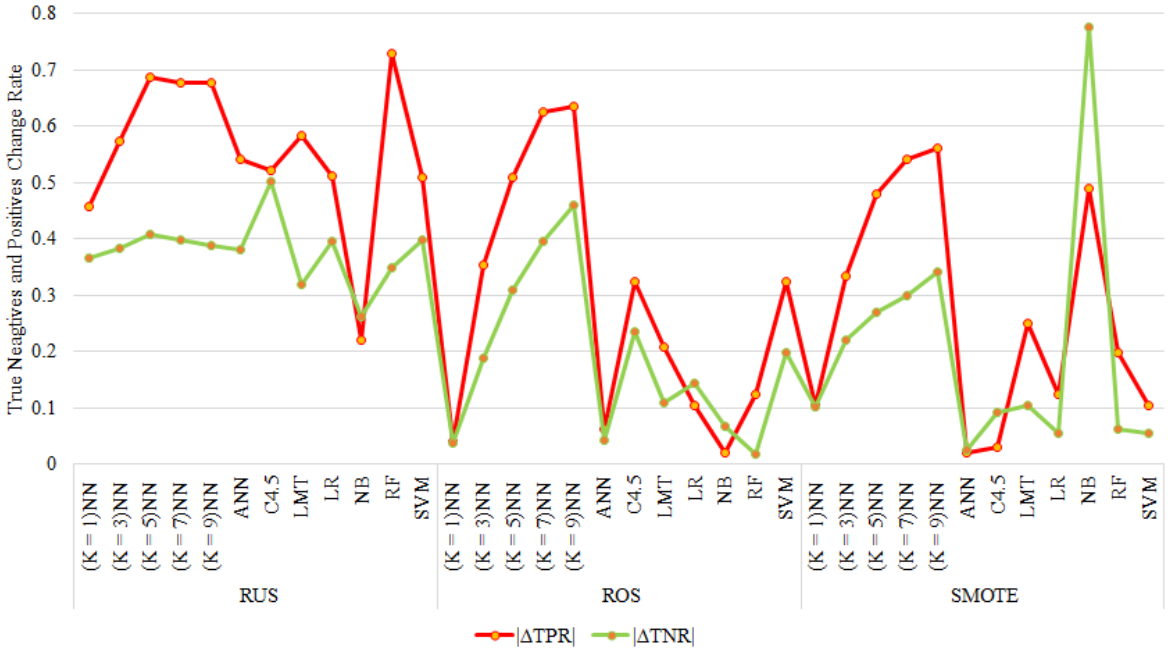

Fig. 9 Absolute TN and TP change rates in test per classifier in RUS, ROS and SMOTE models

Fig. 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 (Fig. 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 (Fig. 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.

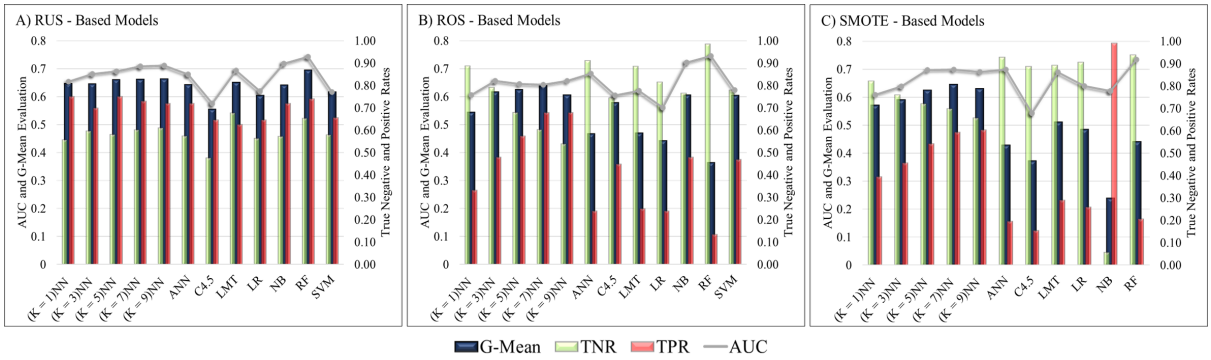

Fig. 10 TNR, TPR, G-Mean and AUC test performances per classifier in RUS, ROS and SMOTE models

In SMOTE-based 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.

3.6 Model's selection and simplification

In clinical trials, it is known that the AUC-ROC metric preserves the discriminant validity in treatment comparisons in balanced data [57] 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. In the previous results sub-sections, 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, where different models are deemed suitable, choosing a single model as a Hero model becomes challenging. Hence, domain experts should set an additional success criterion to define an acceptable level of TPR-TNR trade-off.

To select 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 (see Fig. 11). Seven models with a high AUC-ROC associated with a low G-Mean indicated overfitting of the majority negative class and underfitting the minority positive class were also dropped out.

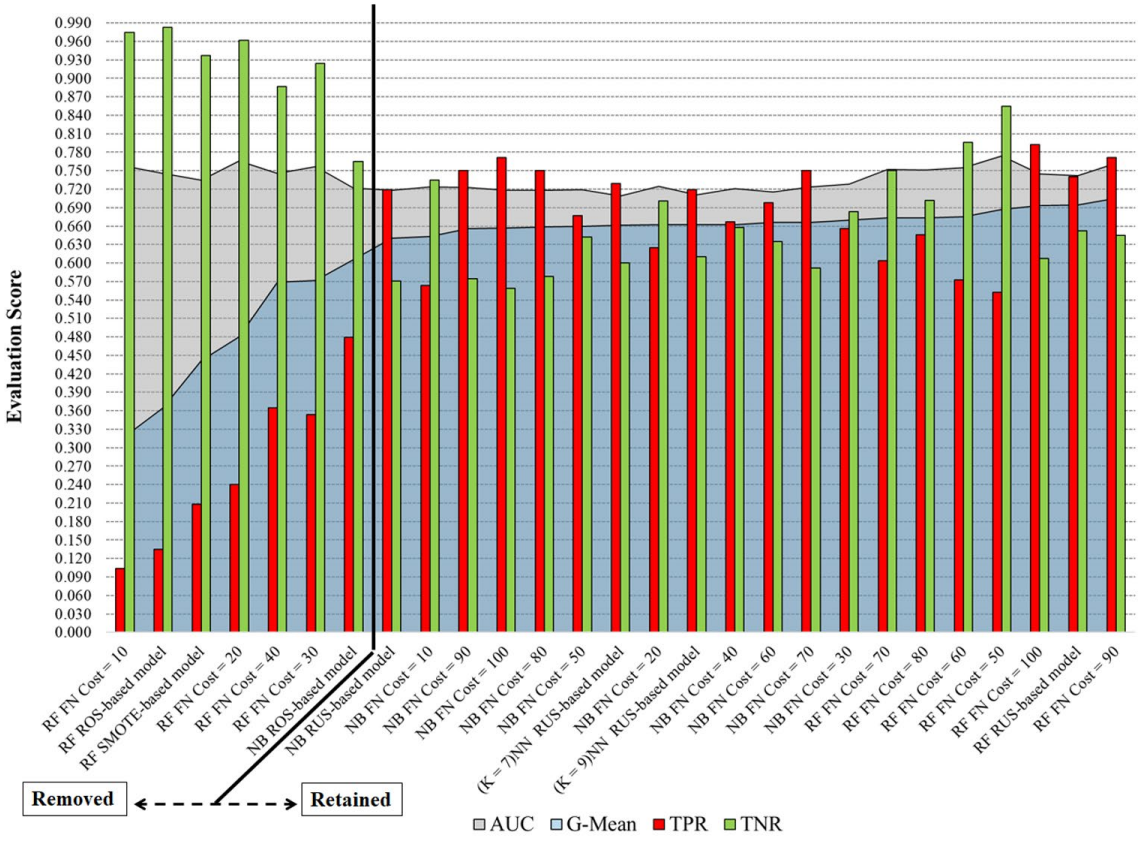

Fig. 11 Filtered models and their associated TNR, TPR, G-Mean and AUC-ROC test performances

Based on all models' validation TPR and TNR evaluations and the clinicians' trade-off between TPR and TNR in Fig. 12, 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). The confusion matrices for the compliant three tested models are found in Table 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 paper.

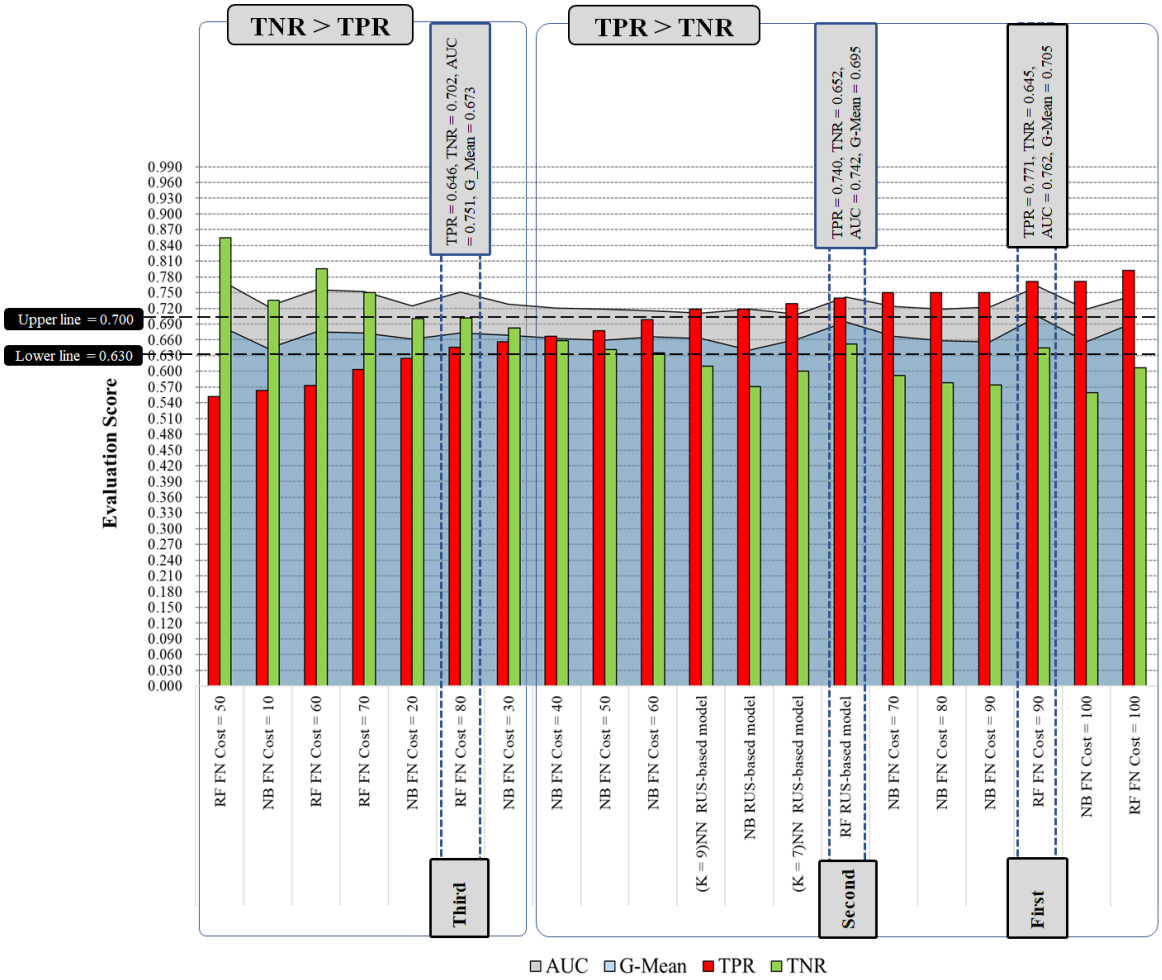

Fig. 12 True Positive Rate (TPR) and True Negative Rate (TNR) trade-offs threshold lines for all tested models with VD. FN prediction costs refer to Penalty values in the explicit cost-sensitive models. While FP predictions costs are kept at a value of 1, both TP and TN predictions costs always remained at the value of zero

Table 7. Performance ranking of the compliant three tested models with 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)		
	Predicted	1st		*Predicted*	2nd		*Predicted*	3rd
	$Desq^-$	$Desq^+$		$Desq^-$	$Desq^+$		$Desq^-$	$Desq^+$
$Desq^-$	602	331	$Desq^-$	608	325	$Desq^-$	655	278
$Desq^+$	22	74	*Actual*	25	71	*Actual*	34	62
G-Mean = 0.705			**G-Mean = 0.695**			**G-Mean = 0.673**		

3.7 Model's simplification

The hero model has many predictors M=122, which makes its interpretability quite complicated. Feature importance in RF was calculated with Mean Decrease Impurity [53]. Eight features were estimated to have zero importance for the model CS-RF(FN:FP = 90:1). In order to simplify the hero model, these features were removed, and a final model was rebuilt and tested. As a result, the simplified model performance slightly improved its specificity to 0.658, AUC to 0.771 and G-Mean to 0.712, while its sensitivity remained unchanged. Feature

importance is described in supplementary material tables B and C. The final simplified model's performance is described in Table 8.

Table 8. Simplified hero and final simplified models' test confusion matrices and performances

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: $Desq^-$ 602, $Desq^+$ 331; Actual: $Desq^-$ 602, $Desq^+$ 22	Predicted: $Desq^-$ 614, $Desq^+$ 319; Actual: $Desq^-$ 614, $Desq^+$ 22
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 final model, Matthew Correlation Coefficient (MCC) and the Area Under the Precision-Recall Curve (AUC-PR) 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 final model predictions are not random and leaning towards strong predictions. The AUC-PR of the final model was 0.902 indicates a good detection of positive outcomes and a strong prediction performance on the final model.

4. 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 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 previously demonstrated to correlate with acute desquamation significantly. Imaging and genomic risk factors were excluded. [58]

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/test data split, it was noted that few variables in the test 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. 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 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 [59 – 65].

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 all 122 predictors were included in the modelling process.

For ML modelling, tackling the imbalanced class problem has a significant impact on the performance of standard parametric and non-parametric ML algorithms. Also, the classification modelling performance in the training phase is severely impacted by class separability. The training of the standard ML algorithms with highly imbalanced classes without adjusting the training set results in an accuracy bias towards the majority class. In this study, we tackled that bias by applying two approaches. In one approach, 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 approach) awarded incremental 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. 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 become curtail. As previously described, 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 AUC-PR.

This study showed that applying the correct level of resampling without disrupting the original data distribution in the RUS-based method, together with the desired choice of performance metrics and slight manipulation of IG levels, produced a good prediction solution [66]. The RF-RUS model competed with further developed models with algorithmic modifications in the case of cost-sensitive classification. Among all 89 models reported in this study, three models satisfied the trade-off threshold conditions (see table 7). However, one "hero" model was selected for this specific domain problem: 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. 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 towards higher misclassification costs to improve their original TPR in the imbalanced data 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 datasets compared to balanced datasets. Classifiers behaved very 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 [67]. 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 (See Fig. 5). 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 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 [69] 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' 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 each class's importance in the domain experts view and allows for a pragmatic final model selection.

Currently, mechanistic models are embedded within the treatment planning systems to predict RT complications, these are Lyman–Kutcher–Burman models [70][71]. These models allow for effective biological optimization of the delivered radiation dose among competing treatment strategies; however, the handmade exceptions in their algorithms means 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 [56][21][23] that produced clinically valid ML models for detecting acute side effects of breast RT. In one study [21], models were built based on the detection of 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 [23] 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. The prediction models are complex, as they used 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 [72] attempted to create simpler toxicity prediction models that excluded dosimetry and radiomic data. The model did not clinically validate in the REQUITE cohort.

Unlike the previous studies, our study accounts for less and easy to obtain variables in the course of the RT treatment planning phase, incorporating the largest cohort among other studies and our model is considered clinically valid. Data-driven studies often lack reporting the data preparation and pre-processing techniques

involved to build their ML prediction models. By reporting the full methodology designed and delivered by an interdisciplinary team of experts, we build researchers confidence in our findings. Our 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.

Limitations of this and many other ML papers used in radiation oncology are the number of variables used compared to routine practice and the different toxicity scales and grades for an acute skin reaction and ulceration defining the class end point.

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, to the best of our knowledge, this is the first effort to summarize the field's current state and produce the simplest clinically valid prediction model.

5. Clinical implication and next steps

Our study shows that applying 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. This was done 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 testing the models with an isolated group of patients of a 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.

Our successful final ML model has the potential to aid clinical facilities and practitioners in minimizing side effects and increasing the chance of RT positive outcomes. Before being embedded into applications, the model can undergo further clinical assessment in line with the optimised radiation dose output obtained from the current mechanistic models. The final model can utilise the treatment dosimetry measurements obtained from the current treatment planning system to predict acute desquamation accurately. Decisions obtained from the legacy system and the new model are recorded and compared.

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 other research from the same domain. Nevertheless, further analyses will be followed where additional methods to improve the outcomes will be investigated.

Acknowledgements

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 and Petra Seibold. 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 Classify-ing Rare Classes?", Proc. Int'l Workshop Utility-Based Data Mining, pp 69-77
16. Faith, J., Mintram, R., Angelova, M.: Gene expression Targeted projection pursuit for visualising gene expression data classifications. Bioinformatics 22(21), 2667–2673 (2006).
17. Harris, E., 2002, January. Information Gain Versus Gain Ratio: A Study of Split Method Biases. In ISAIM.

18. 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.

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

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

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

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

28. 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.

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

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

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

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

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

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

35. 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.

36. 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.

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. 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).

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

51. 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.

52. 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).

53. 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).

54. 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.

55. 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.

56. Isaksson, L.J., Pepa, M., Zaffaroni, M., Marvaso, G., Alterio, D., Volpe, S., Corrao, G., Augugliaro, M., Starzyńska, A., Leonardi, M.C. and Orecchia, R., 2020. Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy. Frontiers in oncology, 10, p.790.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.

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. 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.

71. 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.

72. 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.

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_performed	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_performed	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