Journal of Big Data

RESEARCH Open Access

Predicting clinical outcomes of radiotherapy for head and neck squamous cell carcinoma patients using machine learning algorithms

Tarun Gangil[1], Amina Beevi Shahabuddin[1], B. Dinesh Rao[2], Krishnamoorthy Palanisamy[3], Biswaroop Chakrabarti[3] and Krishna Sharan[1]*

*Correspondence: tk.sharan@manipal.edu
[1] Department of Radiotherapy and Oncology, Kasturba Medical College-Manipal, Manipal Academy of Higher Education, Manipal, Karnataka, India
Full list of author information is available at the end of the article

Abstract

Background: Radiotherapy is frequently used to treat head and neck Squamous cell carcinomas (HNSCC). Treatment outcomes being highly uncertain, there is a significant need for robust predictive tools to improvise treatment decision-making and better understand HNSCC by recognizing hidden patterns in data. We conducted this study to identify if Machine Learning (ML) could accurately predict outcomes and identify new prognostic variables in HNSCC.

Method: Retrospective data of 311 HNSCC patients treated with radiotherapy between 2013 and 2018 at our center and having a follow-up of at least three months' duration were collected. Binary-classification prediction models were developed for: Choice of Initial Treatment, Residual disease, Locoregional Recurrence, Distant Recurrence, and Development of New Primary. Clinical data were pre-processed using Imputation, Feature selection, Minority Oversampling, and Feature scaling algorithms. A method to retain original characteristics of dataset in testing samples while performing minority oversampling is illustrated. The classification comparison was performed using Random Forest (RF), Kernel Support Vector Machine (KSVM), and XGBoost classification algorithms for each model.

Results: For the choice of the initial treatment model, the testing accuracy was 84.58% using RF. The distant recurrence, locoregional recurrence, new-primary, and residual models had a testing accuracy (using KSVM) of 95.12%, 77.55%, 98.61%, and 92.25%, respectively. The important clinical determinants were identified using Shapely Values for each classification model, and the mean area under the curve (AUC) for the receiver operating curve was plotted.

Conclusion: ML was able to predict several clinically relevant outcomes, and with additional clinical validation, could facilitate recognition of novel prognostic factors in HNSCC.

Keywords: Squamous cell head and neck cancer, Machine learning, Shapely values, Prognosis, Recurrence pattern, Feature selection, Missing value imputation

Introduction

Head and neck squamous cell cancers (HNSCC) constitute a diverse group of cancers arising from the head and neck region with common risk factors, natural history, and similar treatment principles. Worldwide, they are the 6th most common malignancy [1]. Despite advances in evaluation and treatment, outcomes of HNSCC continue to be poor. Radiotherapy, either as a primary treatment or in addition to surgery, is integral to the management of most patients with HNSCC. Radiotherapy delivered with curative intent for HNSCC is associated with significant toxicity in addition to suboptimal disease outcomes. However, there is substantial variability in outcomes; some patients are cured of their disease while others aren't, and similarly, toxicities of treatment are minimal in some and excessive in others. This observation reflects the underlying heterogeneity among patients and their cancers, primarily a result of incomplete biological understanding of both the disease and the patient [2]. Several ongoing strategies are looking at addressing this inadequacy, including genomics, radiomics, mathematical models, and newer therapeutic strategies [2].

Machine learning (ML) has been increasingly utilized in recent years to advance medicine, including cancer treatment [3]. Artificial Intelligence (AI) has found application in various aspects of medicine, ranging from imaging (where it has been evaluated for screening, diagnosis, and prognostication of radiological, pathological, and other medical images) to treatment planning, execution, and follow-up. With radiation oncology in specific, AI has been explored for image segmentation, radiation dose optimization, quality assurance, and clinical decision support [4]. Regarding clinical decision-making, AI has significant potential in facilitating a better understanding of HNSCC and its treatment. A few publications have investigated the feasibility of the implementation of AI on clinical details of HNSCC patients to determine the various outcomes, such as 5-year recurrence rates [5–8]. Few studies illustrated how ML algorithms aided in predicting nodal metastasis in early oral squamous cell carcinoma using clinical and pathological data [9, 10]. Also, several studies have compared ML models and artificial neural networks to predict locoregional recurrence for HNSCC [11, 12]. These studies represent how ML models could effectively help doctors in clinical decision-making [13, 14].

Pre-processing steps are essential to make the given dataset suitable for ML algorithms. Studies have shown that techniques such as Ordinal Encoding, Onehotencoding, and Feature Hashing have been used to encode mixed data types, having numeric and categorical variables [15]. The dataset may consist of missing entries that must be dealt with beforehand. Studies have shown that various iterative imputation techniques work well on datasets with a significant amount of missing entries [16]. The dataset may contain more columns than rows. Therefore, it is essential to perform appropriate feature selection techniques. There are studies that describe methods of feature selection such as Principal Component Analysis [17], Independent Component Analysis [18], Filter based methods [19], Wrapper based approaches [19], and Embedded approaches [19]. Since we intended to identify prognostic features contributing the most to classification, and PCA converts the original dataset into its principal components, it was not considered [20, 21]. Studies have also found methods such as genetic algorithms suitable for imbalanced datasets for feature selection [22].

This study was conducted to identify if ML would be able to predict clinically pertinent outcomes in HNSCC treated with radiotherapy. This paper represents an approach to help design classification models using clinical data. The methods used to pre-process and clean the data are also elaborated upon. Additionally, a method to retain the original characteristics of the dataset in the testing dataset while performing minority class oversampling is illustrated. Finally, the classification results using the best-performing ML model are presented. Predictors having significant classification importance were identified to see if they could offer additional clinical understanding. Implementation of ML algorithms using the collected data was performed using python programming language, executed on Google Colaboratory platform.

Methodology

Collection and structuring of clinical data

A total of 311 patients with HNSCC treated at our center between 2013 and 2018 with radiotherapy were included in this study. After treatment, a patient should have had a minimum follow-up for three months. Raw data was collected from the hospital medical records. The collected dataset constituted of 401 non-mutually exclusive clinical variables. Broadly, the heads under which variables were collected are described in Table 1.

Data collected was structured using Excel sheets. Each row represents an individual patient. A single patient's variables should not be present in multiple rows. Columns were structured such that all possible non-mutually exclusive variables (according to the possibilities found in collected 311 samples) were separated, and any given patient's data variable (samples used for validation of the designed algorithm) would be contained in this defined structure (Fig. 1).

According to the data type collected, Table 2 shows the output labels on which the input variables were meant to be mapped to the output vector using classification ML models (supervised learning). There are five models designed in this study (Table 2). Before beginning to build individual models, the column variables were segregated in such a way that every column held clinical meaning in use for the

Table 1 Brief description of collected dataset

S. No.	Clinico-pathological details	No. of variables (n = 401)
1	General details	46
2	Presenting symptoms	64
3	Addictions/substance abuse	17
4	Comorbidities	27
5	Site of cancer	41
6	The extent of primary and Lymph nodes	90
7	Clinical and pathological staging	8
8	Primary treatment details (surgical, radiotherapy, and chemotherapy details)	33
9	Histopathological details	19
10	Acute and late toxicities	33
11	Disease outcome	23

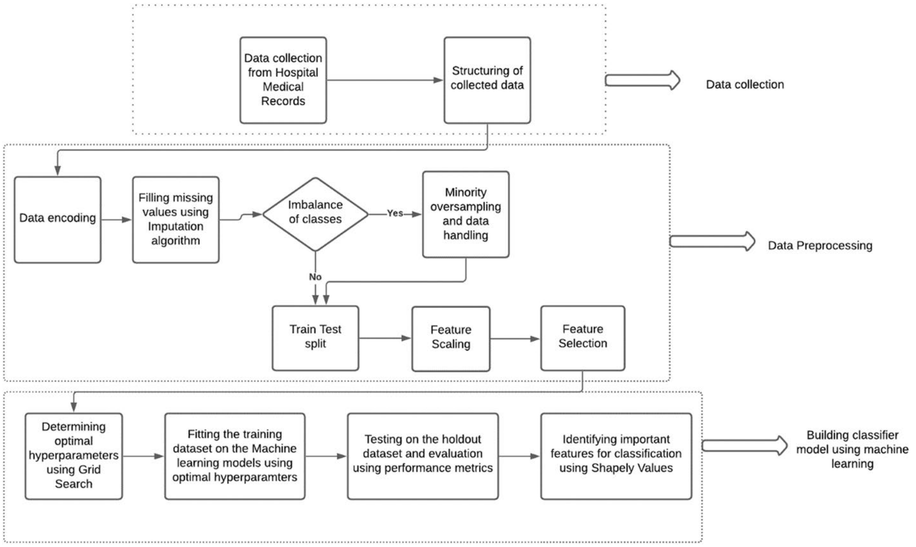
Data collection from Hospital Medical Records
Structuring of collected data
Data collection
Data encoding
Filling missing values using Imputation algorithm
Imbalance of classes
Yes
No
Minority oversampling and data handling
Data Preprocessing
Train Test split
Feature Scaling
Feature Selection
Determining optimal hyperparameters using Grid Search
Fitting the training dataset on the Machine learning models using optimal hyperparamters
Testing on the holdout dataset and evaluation using performance metrics
Identifying important features for classification using Shapely Values
Building classifier model using machine learning

Fig. 1 Workflow of the study. Each model was developed in steps, as illustrated

Table 2 Classification models designed along with their respective encoded output vectors

Model number	Model name	Output label
1	Choice of Initial Treatment (the first treatment that was given to the patient, typically surgery or radiotherapy)	Label 1: Surgery Label 0: Radiation Therapy
2	Residual (disease fails to clear after even three months of completion of radiotherapy; this parameter does not apply to patients who underwent initial surgery)	Label 1: Residual present Label 0: Residual Absent
3	Locoregional recurrence (disease recurs in the irradiated site itself sometime after completion of treatment)	Label 1: Locoregional Recurrence Label 0: No Locoregional Recurrence
4	Distant recurrence (the recurrence of the disease at a site away from its origin, typically by spreading through the blood)	Label 1: Distant Recurrence present Label 0- No Distant Recurrence
5	New Primary (new cancer, unrelated to the treated cancer, developing in the head and neck region)	Label 1: New Primary present Label 0: New Primary absent

prediction of its respective output vector. For example, the events occurring after the start of treatment wouldn't be used to predict initial treatment, and late toxicities wouldn't be used to predict acute toxicity. Hence, the sample size varied while designing each model.

Figure 1 illustrates the workflow of our research. The data collection step involved collecting clinical data from medical records and structuring it into non-mutually exclusive columns. The data pre-processing step involved encoding of data, missing value imputation, and checking for class imbalance. If there was a class imbalance, then minority oversampling was performed. Further, the dataset was split into training and testing datasets, and feature scaling and feature selection were performed. The training data was made to fit on the ML algorithms, tuned on the hyperparameters. Testing dataset was used to generate performance metrics for each algorithm. Also, Shapely analysis was used to identify the list of variables contributing the most to classification.

Data encoding

The recorded input variables were of quantitative and categorical form. Quantitative attributes were recorded according to their respective measuring units. Ordinal attributes were encoded into an orderly numeric form using Label Encoder. Nominal attributes were encoded using OnehotEncoder [15, 23].

Missing value imputation using imputation algorithms

A retrospective data will have missing values by its nature because of various reasons, such as corrupt data, failure to retrieve the information, or incomplete extraction. If we attempt to delete the rows consisting of missing data, then sample size would reduce drastically, resulting in a poor machine learning classifier. The missing input dataset was imputed using Statistical Imputation, KNN Imputation, and Multiple imputation by chained equations (MICE) [16, 24]. After imputation from each algorithm, the imputed dataset was made to fit on the RF algorithm [23], and the accuracy generated was calculated as shown in Table 5, appendix. The model was evaluated using ten splits of k-fold cross-validation so that each sample would be a part of the training and testing dataset. This ensures the effectiveness of ML models in limited data. The best-performing missing value imputation algorithm was chosen for the dataset.

Handling class imbalance using minority oversampling and feature scaling

ML algorithm assumes that an approximately equal number of samples are present in each class of a dataset. But in a retrospective dataset and variable medical outcomes, class imbalances are common. To deal with such imbalance, for each model, the number of samples belonging to respective classes was counted. If the number of samples in each class were approximately equal, then oversampling step was skipped. Otherwise, the class imbalance was handled using oversampling techniques, including Random Oversampling, SMOTE, Borderline SMOTE, SVM SMOTE, and ADASYN algorithms for oversampling of minority class [25]. Performance was evaluated using each algorithm, and the best-performing metric was chosen. The shape of the new dataset was recorded to determine the dimensions. Newly added synthetic data was stored in separate variables similarly minority and majority dataset concerning classes were stored in separate variables [26].

The dataset consisting of samples only from the majority class was split into training and testing datasets. The ratio of 67:33 was maintained for training and testing. The training dataset for the model was built by adding a training split majority class dataset with the synthetic dataset. The testing dataset was built by adding the test split of the majority dataset with the minority data from the original dataset (Fig. 2). In this improvised method as against conventional technique, after performing minority oversampling on the original dataset, training and testing variables were generated such that the training samples would contain train-split of majority class samples and synthetically generated samples. Also, it was ensured that the testing dataset would contain test-split of majority and original minority samples. This customization was performed to extract more reliable performance from the testing dataset.

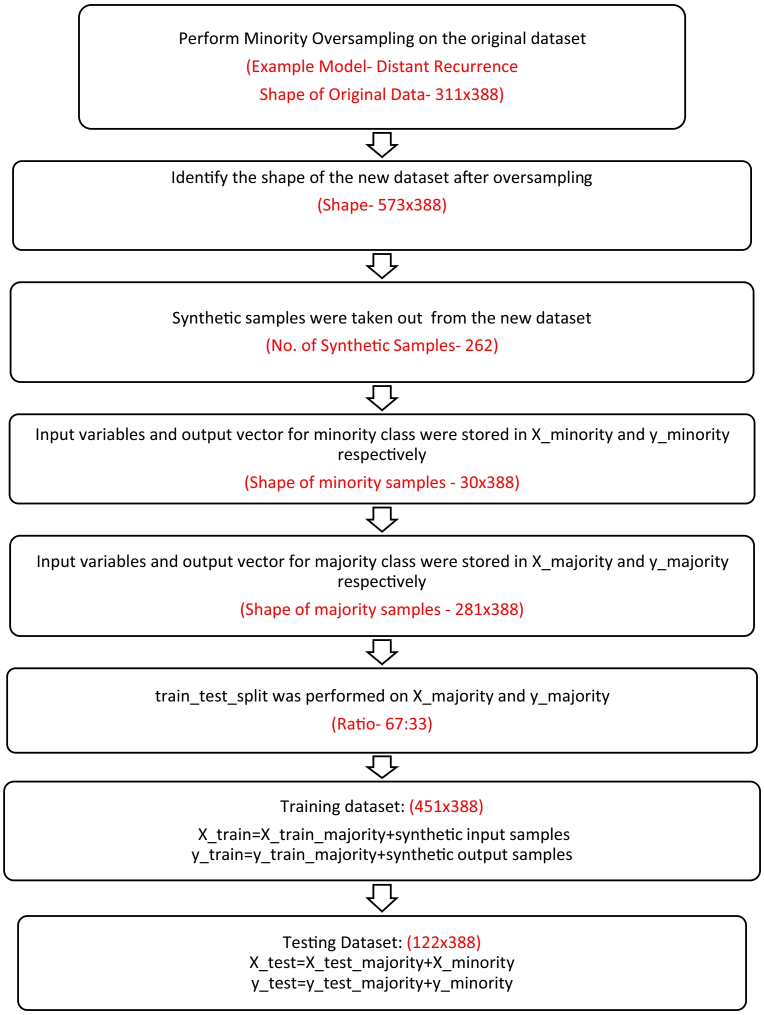
Perform Minority Oversampling on the original dataset
(Example Model- Distant Recurrence
Shape of Original Data- 311x388)
Identify the shape of the new dataset after oversampling
(Shape- 573x388)
Synthetic samples were taken out from the new dataset
(No. of Synthetic Samples- 262)
Input variables and output vector for minority class were stored in X_minority and y_minority respectively
(Shape of minority samples - 30x388)
Input variables and output vector for majority class were stored in X_majority and y_majority respectively
(Shape of majority samples - 281x388)
train_test_split was performed on X_majority and y_majority
(Ratio- 67:33)
Training dataset: (451x388)
X_train=X_train_majority+synthetic input samples
y_train=y_train_majority+synthetic output samples
Testing Dataset: (122x388)
X_test=X_test_majority+X_minority
y_test=y_test_majority+y_minority

Fig. 2 Flow Diagram for minority oversampling using an example of distant recurrence

The input variables present in the dataset were of different measuring scales and had different ranges of magnitude. Hence, the dataset was scaled using standard scalar function [23], such that all variables have zero mean and unit variance.

Feature selection

The collected dataset consisted of more features than the number of samples, thereby falling under the curse of dimensionality [27]. Therefore, feature selection was performed on the dataset. In this process, the variables contributing most to classification were selected. Boruta [28, 29] and Sequential forward floating selection (SFFS) [30, 31] methods were used for feature selection. Each method was evaluated using the performance metrics of three Machine Learning models- RF, KSVM [23], and XGBoost [32]. The performance results from SFFS were better and were therefore chosen as a feature selection method for the dataset. For the Boruta method, the RF algorithm was used as the base algorithm. For the SFFS method, the same algorithm was chosen to run as its base, on which the training dataset was fit. Both Boruta and SFFS were made to run in a setting where it will select k-best features out of all the features in the dataset, and precedence was given to the features having the highest accuracy.

Optimal hyperparameters for ML algorithms

The ML algorithms were made to run optimally for finding the best set of hyperparameters by fitting them on the given dataset. Using Gridsearch [23] approach, optimal hyperparameters were found for each RF, KSVM, and XGBoost model. These hyperparameters were used, both in the base algorithm while performing feature selection and on the ML algorithms while performing classification (Table 11, appendix).

The training dataset was fit on the ML model using the optimal hyperparameters derived using grid search. The new dataset with only the selected features was chosen, and the ML model was fit on the training dataset.

Testing the designed model

Testing of the designed model was done on the hold-out dataset, and its efficiency and reliability were determined using evaluation of performance metrics. This was the dataset that was not included while training the model. The results of classification prediction were compared with the output vector of the testing dataset to generate performance metrics [23]. We calculated Training and Testing Accuracy, Sensitivity, Specificity, F1 score [33] for both classes, and AUC of Receiver Operating Curve [34]. Each performance measure was recorded five times, and the average value was reported. Finally, the best performing ML algorithm for each model was made to run ten times, and the average performance measure value was reported. The most important features contributing to the classification were fetched using Shapely values [35]. The higher value on the labels yes or no indicates the stronger positive or negative correlation of a particular variable with respect to outcome. The mean AUC for each model was plotted, and the data was divided using stratified k fold cross-validation [36] to determine the best performance of the models. A higher AUC score represents that a classifier has a better performance in distinguishing between the two classes.

Results and analysis

A total of 311 patients with HNSCC treated with radiotherapy and having a follow-up of more than three months were found suitable for the study. Males constituted 257 patients, and the mean age of patients was 56.5 years. The average follow-up duration was 23 months.

During the various pre-processing steps, compared to statistical Imputation and KNN Imputation, the Iterative Imputation (MICE) algorithm gave the highest accuracy of 68.6% while running the algorithm for four iterations, having 'ascending' hyper-parameter with RF (Table 5; appendix).

All the results discussed below are reported as the mean of five iterations. The performance of RF, KSVM, and XGBoost were compared, and results of the best performing ML algorithm are reported.

For 'choice of initial treatment,' no oversampling methods were performed due to insignificant class imbalance. RF gave the best performance, using 34 k-best features selected by SFFS. The mean training and testing accuracies were 94.3% and 84.58%, respectively. The sensitivity and specificity were calculated to be 85.1% and 85.7%, respectively. The algorithm tended towards overfitting, but comparatively, testing accuracy was better than the other two algorithms. Similarly, the F1 score was calculated to be the highest among the algorithms (85% for label 0 and 84.2% for label 1). Finally, the AUC was calculated to be 0.8904, signifying it to be a good classifier for the given dataset (Table 6, appendix).

For the remaining four models, all five oversampling techniques were applied, and each technique was evaluated using the three ML algorithms. ADASYN minority oversampling gave the best performance for 'distant recurrence,' whereas the SMOTE minority oversampling technique performed best for locoregional recurrence, new primary and residual-disease predictions. The classification performance was determined using the mean accuracy of the testing dataset. The distant recurrence, locoregional recurrence, new-primary, and residual models had a testing accuracy (using KSVM) of 95.12%, 77.55%, 98.61%, and 92.25%, respectively. The respective sensitivity and specificity values were 90% and 98%, 95% and 89%, 100% and 98%, and 72% and 97%.

As a measure of precision and recall, F1 scores were calculated for both classes of each model. For distant recurrence, the mean F1-score was calculated to be 91% and 96.8% for minority and majority classes. For locoregional recurrence, they were 68.6% and 82.4%; for new primary, they were 95.4% and 98.8%; for residual disease, they were 87.6% and 94.6%.

The AUC of ROC curves for distant recurrence, locoregional recurrence, new-primary, and residual disease prediction models were 0.998, 0.9453, 0.9994, and 0.9948, respectively (Table 7, 8, 9, 10; appendix).

Finally, the best performing ML algorithm was iterated ten times to ensure the reliability of performance that is consistent for the complete dataset. The mean AUC scores were 0.977, 0.734, 0.983 and 0.993, respectively (Table 3). Also, recurring features identified after ten iterations of Shapely analysis provide the list of most contributing predictors for each model (Table 4).

For visualizing the best performing ML algorithm for each model, ten splits of k-fold cross-validation were applied. The mean AUC and AUC for each iteration are presented (Fig. 3). The mean AUC for Choice of Initial Treatment, Distant Recurrence, Locoregional Recurrence, New Primary and Residual were 0.93 ± 0.07, 0.99 ± 0.00, 0.96 ± 0.02, 0.99 ± 0.00, and 0.98 ± 0.03, respectively.

Discussion

HNSCC merits significant advances in the way it is presently managed, and one of the potential strategies to achieve this is the ability to accurately predict outcomes. This study highlights the possible clinical utility of ML in the management of HNSCC. This approach aids, in addition to predicting the outcomes themselves, to also identifying the factors responsible for a particular outcome (event) occurring. For instance, fore-knowledge of whether a patient is likely to suffer a residual disease, locoregional, and/or a distant recurrence carries tremendous importance. HNSCCs are aggressive tumors, and

Table 3 Results from test dataset for distant recurrence, locoregional recurrence, new primary and residual models using KSVM algorithm using its optimal hyperparameters

Model name	**Distance recurrence**	**Locoregional recurrence**	**New primary**	**Residual**
Particular	**Value**	**Value**	**Value**	**Value**
No. of original samples (rows)	311	311	311	152
Total number of independent variables (columns)	388	384	388	354
Feature selection method used	SFFS	SFFS	SFFS	SFFS
ML algorithm	KSVM	KSVM	KSVM	KSVM
The minority oversampling method used	ADASYN	SMOTE	SMOTE	SMOTE
No. of samples after oversampling	573	514	588	270
No. of synthetic samples	262	203	277	118
No. of features selected by SFFS	18	24	42	312
No. of original samples for class-0	281	257	294	135
No. of original samples for class-1	30	54	17	17
Mean train accuracy	0.99	0.96	1	1
Mean test accuracy	0.94	0.73	0.96	0.91
Sensitivity	0.87	0.83	0.94	0.89
Specificity	0.96	0.73	0.98	1.00
Mean training F1 score class label 0	0.99	0.96	1	1
Mean training F1 score class label 1	0.99	0.97	1	1
Mean testing F1 score class label 0	0.96	0.78	0.97	0.93
Mean testing F1 score class label 1	0.89	0.64	0.89	0.87
Base algorithm	KSVM	KSVM	KSVM	KSVM
Mean AUC_ROC	0.97	0.73	0.98	0.99

Table 4 List of common features selected by Shapely analysis

S. No.	**Model name**	**Feature name**	**Label: yes**	**Label: no**
1	Distant recurrence	No. of times recurrences/residual	High	Low
		Type of recurrence LocoRegional	Low	High
		New primary	High	Low
2	Locoregional recurrence	Alcohol consumption	High	Low
		Pain in the oral cavity	No Impact	High
		Mean dose larynx	High	Low
3	New primary	No. of recurrences	High	Low
		Size of lymph nodes	Low	High
		PTV V95%	Low	High
		Age in years	Low	High
4	Residual	Number of lymph nodes	High	Low
		Age	Low	High
		Vegetarian diet	No Impact	High
		Site of cancer	High	No Impact
		Poor orodental hygiene	Low	High

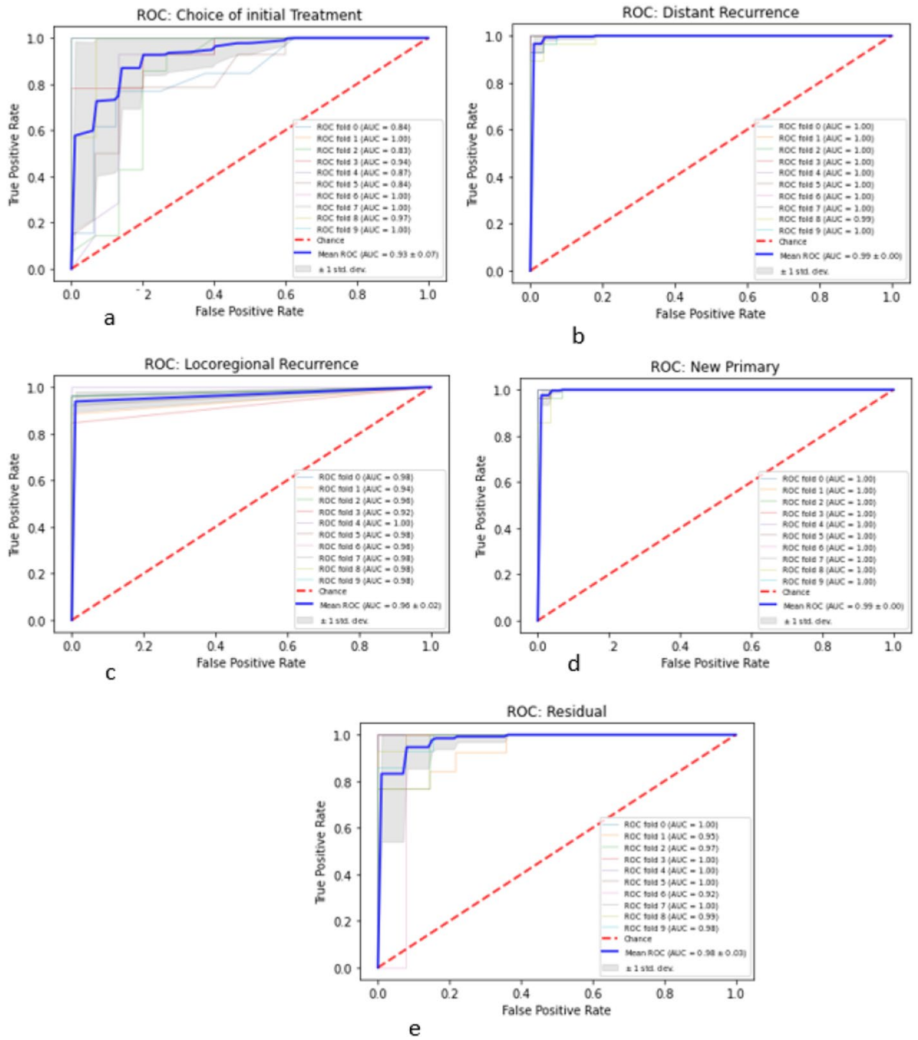

Fig. 3 Mean ROC plots for each model: **a** Choice of Initial Treatment, **b** Distant Recurrence, **c** Locoregional Recurrence, **d** New Primary, **e** Residual

especially in more advanced stages, recur in a substantial number of patients. Though locoregional recurrences constitute a significant proportion of recurrences, many present with distant recurrences too. It is currently difficult to predict recurrence patterns; the ability to foresee it could pave way for a greater understanding of the disease and help in initial treatment planning. For example, a patient inclined to recur at a distant site could be considered upfront for a more aggressive systemic therapy, and a patient at low risk of local recurrence could be considered for de-intensification strategies in locoregional treatment. Similarly, if the patient is expected to have residual disease following radiotherapy, alternative strategies, such as dose-escalation, chemotherapy intensification, or in some cases excluding radiotherapy completely, could greatly benefit the outcomes in terms of both tumor control and reduced toxicity. In our study, for example, the generated models could predict the risk of locoregional recurrence with reasonable accuracy of 73.45%.

Several researchers have looked into the benefit of ML in predicting treatment on a retrospective dataset of HNSCC patients, with very encouraging results. For instance, researchers from the University of Chicago reported that the ML algorithms successfully identified patients with intermediate-risk HNSCC who stood to benefit from concurrent chemotherapy from those who did not [37]. This study highlights the potential of ML in being a valuable tool for clinical use once the findings have been validated.

In addition, by looking at features that carried a higher weightage in a particular prediction, ML could also facilitate in developing a better understanding of the disease. As an example, for locoregional recurrence, among the feature weights, like presence and size of lymph nodes, presence of gross margin positive status following resection, the extent of primary and grade of tumor carried higher importance across multiple models. While these were expected factors, there were also some unusual factors having a bearing on local recurrence in some of the models, such as doses received by the organ at risk (normal structure) and use of chemotherapy- both of which could perhaps partially be explained by their likelihood of occurrence being high in more advanced disease. Similarly, most recurring factors weighing high in the prediction models for distant recurrence included the presence of residual disease, absence of locoregional recurrence, and development of new (metachronous) primary. Interestingly, in one of the models, the patients' address appeared to have a bearing on distant recurrence, with patients hailing from a particular district less prone to distant recurrence. This exercise highlights the potential of such an AI implementation in healthcare. Properly implemented, the possibilities are immense—all the way from the departmental level to the national and even international levels.

There is also the possibility of systematization of the practice of oncology. For example, multiple factors are taken into consideration in determining the initial choice of treatment for a patient with HNSCC; these include the site of primary, its locoregional extent, stage of the disease, the patient's overall health, and presence of comorbidities, etc. However, despite established guidelines, the choice of treatment for an individual patient can be variable and is usually individualized. Decision-making in the multi-disciplinary field of oncology is therefore fairly complex and is fraught with differing opinions and a significant lack of consensus among the treating team. The potential benefits of being able to accurately predict treatment allocation include systematizing the process, in addition to gaining a better understanding of the complex interaction of different factors that make up the eventual decision.

Our study has several limitations. It included multiple sub-sites of HNSCC, which could reduce clarity. Another glaring problem was the quality and quantity of data; it investigated a single-center retrospective data, with a small number of patients with relatively poor follow-up information on outcomes. In addition to improving the accuracy, a greater sample size could have also avoided the curse of dimensionality. Similarly, a prospective implementation of such data capturing across multiple centers could help gain a much better understanding of such cancers. This work needs additional validation before it can be implemented in the clinic. Additional variables, including radiomics and advanced pathological data could also be incorporated to make a robust tool that can be confidently utilized by clinicians in decision-making.

Conclusion

With the data available for analysis in our study, the iterative imputation algorithm gave the best accuracy compared to Statistical and KNN Imputation while evaluating with RF. SFFS feature selection algorithm was used since its performance metrics were better than Boruta while evaluating with RF, KSVM, and XGBoost for all models. SMOTE and ADASYN gave the best minority oversampling performance metrics using the three ML algorithms. Intrinsic pre-processing, as applied in this research, greatly facilitated the performance of the designed ML models.

Thus, ML was able to predict several clinically relevant outcomes of patients with HNSCC receiving radiotherapy, using clinicopathological data, with reasonable accuracy. This could significantly impact the way patients are managed, leading to a better understanding of disease and improved outcomes. However, such findings need to be validated prospectively and across multiple centers before they can be introduced into routine clinical use.

Appendix

See Tables 5, 6, 7, 8, 9, 10 and 11.

K Fold cross-validation as applied to the dataset using 10 splits, repeated 3 times. Evaluation of imputed dataset was performed using RF classifier, the mean and standard deviation of the accuracy of all the splits were reported. The comparison of the accuracy of different algorithms was performed to choose the best performing missing value imputation algorithm (Table 5).

The performance metrics from RF give the best result in comparison to KSVM and XGBoost. Results shown here are presented as the average of five iterations (Table 6).

The performance metrics from KSVM give the best results in comparison to RF and XGBoost. ADASYN minority oversampling technique was performed for this model. Results shown here are presented as an average of five iterations (Table 7).

Here performance metrics of KSVM show the best results in comparison to RF and XGBoost. SMOTE minority oversampling method was used for this model. Results shown here are presented as an average of five iterations (Tables 8, 9, 10).

Table 5 Comparison of results evaluated on RF for various missing value algorithms

S. No.	Missing value algorithm name	Hyperparameter	Mean accuracy score with RF	Std accuracy score with RF
1	Statistical imputation	Mean	0.67	0.06
		Median	0.66	0.06
		Most_frequent	0.67	0.05
		Constant	0.68	0.06
2	KNN imputation algorithm	No. of neighbour 1	0.67	0.06
		No. of neighbours 3	0.67	0.04
		No. of neighbours 5	0.66	0.07
		No. of neighbours 7	0.65	0.06
		No. of neighbours 9	0.66	0.06
		No. of neighbours 15	0.66	0.05
		No. of neighbours 18	0.66	0.07
3	Iterative imputation	Iterative imputation using Ascending hyperparameter	0.68	0.05
		Iterative imputation using Descending hyperparameter	0.67	0.07
		Iterative imputation using Roman hyperparameter	0.67	0.05
		Running iterative imputation for 1 iteration	0.68	0.06
		Running iterative imputation for 2 iterations	0.66	0.06
		Running iterative imputation for 3 iterations	0.67	0.06
		Running iterative imputation for 4 iterations	0.68	0.06
		Running iterative imputation for 5 iterations	0.66	0.09
		Running iterative imputation for 6 iterations	0.66	0.06
		Running iterative imputation for 7 iterations	0.67	0.08

Table 6 Results of test dataset for choice of the initial treatment model

Particular	Data		
Model name	Choice of initial treatment		
No. of samples (rows)	289	289	289
Total number of independent variables (columns)	268	268	268
ML algorithm used	Random Forest	KSVM	XGBoost
Feature selection technique	SFFS	SFFS	SFFS
No. of independent variables used in the dataset	34	53	46
Mean accuracy train	0.94	0.92	0.95
Mean accuracy test	0.84	0.75	0.77
Sensitivity	0.85	0.81	0.85
Specificity	0.85	0.69	0.78
Mean F-Score train label 0	0.95	0.93	0.96
Mean F-Score train label 1	0.94	0.92	0.95
Mean F-score test label 0	0.85	0.76	0.78
Mean F-score test label 1	0.84	0.73	0.75
No. of samples for class-0	152	152	152
No. of samples for class-1	137	137	137
Base algorithm	Random forest	KSVM	XGBoost
ROC_AUC_Score	0.89	0.83	0.88

Table 7 Results of test dataset for distant recurrence model

Particular	Data		
Model name	Distant recurrence		
No. of samples (rows)	311	311	311
Total number of independent variables (columns)	388	388	388
ML algorithm used	Random Forest	KSVM	XGBoost
Feature selection technique	SFFS	SFFS	SFFS
OverSampling method used	ADASYN	ADASYN	ADASYN
No. of samples after OverSampling	573	573	573
Number of synthetic samples	262	262	262
No. of independent variables used in the dataset	22	18	86
Mean accuracy train score	0.99	0.99	1
Mean accuracy test score	0.87	0.95	0.87
Sensitivity	0.92	0.90	0.75
Specificity	0.93	0.98	0.81
Mean F-Score train label 0	0.99	0.99	1
Mean F-Score train label 1	0.99	0.99	1
Mean F-score test label 0	0.92	0.96	0.92
Mean F-score test label 1	0.68	0.91	0.64
No. of samples for class-0	281	281	281
No. of samples for class-1	30	30	30
Base algorithm	Random forest	KSVM	XGBoost
ROC_AUC_Score	0.96	0.99	0.97

Table 8 Results of the test dataset for locoregional recurrence

Particular	Data		
Model name	locoregional Recurrence		
No. of samples	311	311	311
Total number of features	384	384	384
ML algorithm used	Random Forest	KSVM	XgBoost
Feature selection technique	SFFS	SFFS	SFFS
OverSampling method used	SMOTE	SMOTE	SMOTE
No. of samples after OverSampling	514	514	514
Number of synthetic samples	203	203	203
No. of features used in the dataset	216	24	169
Mean accuracy train	0.96	0.97	1
Mean accuracy test	0.66	0.77	0.62
Sensitivity	0.75	0.95	1.00
Specificity	0.63	0.89	0.62
Mean F-Score train label 0	0.96	0.97	1
Mean F-Score train label 1	0.96	0.97	1
Mean F-score test label 0	0.78	0.82	0.76
Mean F-score test label 1	0.26	0.68	0.07
No. of samples for class-0	257	257	257
No. of samples for class-1	54	54	54
Base algorithm	Random forest	KSVM	XGBoost
ROC_AUC_Score	0.89	0.94	0.84

Table 9 Results of the test dataset for the new primary model

Particular	Data		
Model name	New primary		
No. of samples	311	311	311
Total number of features	388	388	388
ML algorithm used	Random Forest	KSVM	XGBoost
Feature selection technique	SFFS	SFFS	SFFS
OverSampling method used	SMOTE	SMOTE	SMOTE
No. of samples after OverSampling	588	588	588
Number of synthetic samples	277	277	277
No. of features used in the dataset	20	42	18
Mean accuracy train	0.99	1	1
Mean accuracy test	0.91	0.98	0.90
Sensitivity	0.88	1.00	0.88
Specificity	0.91	0.98	0.91
Mean F-Score train label 0	1	1	1
Mean F-Score train label 1	1	1	1
Mean F-score test label 0	0.95	0.98	0.94
Mean F-score test label 1	0.63	0.95	0.53
No. of samples for class-0	294	294	294
No. of samples for class-1	17	17	17
Base algorithm	Random forest	KSVM	XGBoost
ROC_AUC_Score	0.95	0.99	0.95

Table 10 Results of test data for Residual model

Particular	Data		
Model name	Residual		
No. of samples	152	152	152
Total number of features	354	354	354
ML algorithm used	Random Forest	KSVM	XGBoost
Feature selection technique	SFFS	SFFS	SFFS
OverSampling method used	SMOTE	SMOTE	SMOTE
No. of samples after OverSampling	270	270	270
Number of synthetic samples	118	118	118
No. of features used in the dataset	93	312	91
Mean accuracy train	0.97	1	0.99
Mean accuracy test	0.84	0.92	0.74
Sensitivity	0.85	0.72	0.50
Specificity	0.89	0.97	0.74
Mean F-Score train label 0	0.97	1	0.99
Mean F-Score train label 1	0.97	1	1
Mean F-score test label 0	0.89	0.94	0.84
Mean F-score test label 1	0.68	0.87	0.25
No. of samples for class-0	135	135	135
No. of samples for class-1	17	17	17
Base algorithm	Random forest	KSVM	XGBoost
ROC_AUC_Score	0.94	0.99	0.91

Table 11 Optimal hyperparameters were calculated for each model with all three algorithms (RF, SVM, and XGBoost)

S. No.	Model name	Best ML model selected	Optimal hyperparameter
1	Choice of initial treatment	Random forest classifier	RandomForestClassifier(bootstrap=True, ccp_alpha=0.0,class_weight='balanced', criterion='gini', max_depth=None, max_features='log2', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100,n_jobs=− 1, oob_score=False, random_state=123, verbose=0,warm_start=False)
2	Distant recurrence	Kernel Support vector Machine(KSVM)	SVC(C=1, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=1, gamma=1, kernel='rbf', max_iter=− 1,probability=False, random_state=None, shrinking=True, tol=0.001,verbose=False)
3	Locoregional recurrence	Kernel Support Vector Machine(KSVM)	SVC(C=1, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=1, gamma=1, kernel='rbf', max_iter=− 1,probability=False, random_state=None, shrinking=True, tol=0.001,verbose=False)
4	New primary	Kernel Support Vector Machine(KSVM)	SVC(C=1, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=1, gamma=1, kernel='rbf', max_iter=-1,probability=False, random_state=None, shrinking=True, tol=0.001,verbose=False)
5	Residual	Kernel Support Vector Machine(KSVM)	SVC(C=10, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=1, gamma=0.001, kernel='rbf', max_iter=− 1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)

The optimal hyperparameters obtained using grid search were used with their respective ML model both for the purpose when implementing a Base ML algorithm while performing feature selection and while training the dataset (Table 11).

Acknowledgements

The authors would like to acknowledge Mr. Manjunatha Maiya, Project Manager, Philips Research India, Bangalore, Karnataka, Mr. Prasad R V, Program Manager Philips Research India, Bangalore, Karnataka, and Mr. Shrinidhi G. C., Chief Physicist, Dept. of Radiotherapy, KMC, Manipal for their help and support in the conduct of this research.

Authors' contributions

All authors contributed to the study design. TG, ABS and KS worked on collecting the data. TG, KS, BC and DR worked on framing the objectives of the study. Technical guidance was provided by KP and DR. Development of analytics model using python programming was done by TG and KP. KS and TG drafted the manuscript and was reviewed it by all authors. All authors read and approved the final manuscript.

Funding
This study was funded by the Manipal Academy of Higher Education, Karnataka, India, and Philips Research India.

Availability of data and materials
The datasets generated and analysed during the current study are not publicly available because the data are from the medical records of a private hospital but are available from the corresponding author on reasonable request.

Code availability
Custom code was developed for this study using python programming language version 3.7.12, and the associated ML packages-developed on google colaboratory platform.

Declarations

Ethics approval and consent to participate
Ethical clearance for collecting retrospective data was provided by Kasturba Medical College and Kasturba Hospital Institutional Ethics Committee (Registration No. ECR/146/Inst/KA/2013/RR-16). The Institutional Ethical clearance number for the study is IEC: 165/2018. The study was registered with the clinical trials registry of India (CTRI). CTRI Number: CTRI/2018/04/013517 (Registered on: 27/04/2018)—Trial Registered Prospectively. Type of study-Observational.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Radiotherapy and Oncology, Kasturba Medical College-Manipal, Manipal Academy of Higher Education, Manipal, Karnataka, India. [2]Manipal School of Information Sciences, Manipal Academy of Higher Education, Manipal, Karnataka, India. [3]Philips Research India, Bangalore, Karnataka, India.

Received: 26 October 2021 Accepted: 15 February 2022
Published online: 25 February 2022

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
2. Caudell JJ, Torres-Roca JF, Gillies RJ, Enderling H, Kim S, Rishi A, et al. The future of personalised radiotherapy for head and neck cancer. Lancet Oncol. 2017;18(5):e266–73. https://doi.org/10.1016/S1470-2045(17)30252-8.
3. Obermeyer Z, Ziad MDD, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1212–6.
4. Deig CR, Kanwar A, Thompson RF. Artificial intelligence in radiation oncology. Hematol Oncol Clin North Am. 2019;33(6):1095–104. https://doi.org/10.1016/j.hoc.2019.08.003.
5. Alkhadar H, Macluskey M, White S, Ellis I, Gardner A. Comparison of machine learning algorithms for the prediction of five-year survival in oral squamous cell carcinoma. J Oral Pathol Med. 2021;50(4):378–84.
6. Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. J Oral Pathol Med. 2020;49(10):977–85.
7. Karadaghy OA, Shew M, New J, Bur AM. Development and assessment of a machine learning model to help predict survival among patients with oral squamous cell carcinoma. JAMA Otolaryngol Head Neck Surg. 2019;145(12):1115–20.
8. Rosado P, Lequerica-Fernandez P, Villallain L, Pena I, Sanchez-Lasheras F, De Vicente JC. Survival model in oral squamous cell carcinoma based on clinicopathological parameters, molecular markers and support vector machines. Expert Syst Appl. 2013;40(12):4770–6. https://doi.org/10.1016/j.eswa.2013.02.032.
9. Bur AM, Holcomb A, Goodwin S, Woodroof J, Karadaghy O, Shnayder Y, et al. Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. Oral Oncol. 2019;92:20–5. https://doi.org/10.1016/j.oraloncology.2019.03.011.
10. Shan J, Jiang R, Chen X, Zhong Y, Zhang W, Xie L, et al. Machine learning predicts lymph node metastasis in early-stage oral tongue squamous cell carcinoma. J Oral Maxillofac Surg. 2020;78(12):2208–18. https://doi.org/10.1016/j.joms.2020.06.015.
11. Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, et al. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. Int J Med Inform. 2020;136:104068. https://doi.org/10.1016/j.ijmedinf.2019.104068.
12. Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, et al. Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool. Virchows Arch. 2019;475(4):489–97.
13. Mandal S, Gupta A, Chanu WP. Survival prediction of head and neck squamous cell carcinoma using machine learning models. 2021;1–8. Available from: http://arxiv.org/abs/2105.07390.

14. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ. Big data for health. IEEE J Biomed Heal Informatics. 2015;19(4):1193–208.
15. Lopez-Arevalo I, Aldana-Bobadilla E, Molina-Villegas A, Galeana-Zapién H, Muñiz-Sanchez V, Gausin-Valle S. A memory-efficient encoding method for processing mixed-type data on machine learning. Entropy. 2020;22(12):1–21.
16. Liu Y, Brown SD. Comparison of five iterative imputation methods for multivariate classification. Chemom Intell Lab Syst. 2013;120:106–15.
17. Arowolo MO, Adebiyi MO, Adebiyi AA, Aremu C. An ICA-ensemble learning approaches for prediction of RNAseq malaria vector gene expression data classification. Int J Electr Comput Eng. 2021;11(2):1561–9.
18. Arowolo MO, Adebiyi MO, Adebiyi AA, Okesola OJ. A hybrid heuristic dimensionality reduction methods for classifying malaria vector gene expression data. IEEE Access. 2020;8:182422–30.
19. Arowolo MO, Adebiyi MO, Aremu C, Adebiyi AA. A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector. J Big Data. 2021;8(1). https://doi.org/10.1186/s40537-021-00441-x.
20. Arowolo MO, Adebiyi MO, Adebiyi AA, Olugbara O. Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier. J Big Data. 2021;8(1). https://doi.org/10.1186/s40537-021-00415-z
21. Arowolo MO, Adebiyi MO, Adebiyi AA. Enhanced dimensionality reduction methods for classifying malaria vector dataset using decision tree. Sains Malaysiana. 2021;50(9):2579–89.
22. Saheed YK, Hambali MA, Arowolo MO, Olasupo YA. Application of GA feature selection on naive bayes, random forest and SVM for credit card fraud detection. Int Conf Decis Aid Sci Appl DASA. 2020;2020:1091–7.
23. Pedregosa F, Varoquaux S, Gramfort A, VincentMichel BT. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
24. Brownlee J, Sanderson M, Koshy A, Cheremskoy A, Halfyard J. Machine learning mastery with Python: Data Cleaning, Feature Selection, and Data Transforms in Python. 2020
25. Brownlee J. Imbalanced classification with Python. Mach Learn Mastery. 2020;463.
26 Kovács G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Appl Soft Comput J. 2019;83:105662. https://doi.org/10.1016/j.asoc.2019.105662.
27. Debie E, Shafi K. Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. Pattern Anal Appl. 2019;22(2):519–36.
28. Akmal C, Yahaya C, Firdaus A, Mohamad S, Ernawan F, Faizal M, et al. Automated feature selection using boruta algorithm to detect mobile malware. Int J Adv Trends Comput Sci Eng. 2020;9(5):9029–36.
29. Naik N, Mohan BR. Optimal feature selection of technical indicator and stock prediction using machine learning technique. In: Communications in computer and information science. vol. 985. Springer Singapore; 2019. p. 261–268. https://doi.org/10.1007/978-981-13-8300-7_22.
30. Shafiee S, Lied LM, Burud I, Dieseth JA, Alsheikh M, Lillemo M. Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. Comput Electron Agric. 2021;183(1432):106036. https://doi.org/10.1016/j.compag.2021.106036.
31. Tan M, Pu J, Zheng B. Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model. Int J Comput Assist Radiol Surg. 2014;9(6):1005–20.
32. Shi X, Li Q, Qi Y, Huang T, Li J. An accident prediction approach based on XGBoost. 20017;1–7. https://doi.org/10.1109/ISKE.2017.8258806.
33. Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. In: Calders T, Esposito F, Hüllermeier E, Meo R, editors. Machine learning and knowledge discovery in databases. Heidelberg: Springer; 2014. p. 225–39.
34. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145–59.
35. Messalas A, Kanellopoulos Y, Makris C. Model-agnostic interpretability with shapley values. In: 10th Int Conf Information, Intell Syst Appl IISA 2019. 2019;1–7.
36. Jung Y, Hu J. A K-fold averaging cross-validation procedure. J Nonparametr Stat. 2015;27(2):167–79.
37. Howard FM, Kochanny S, Koshy M, Spiotto M, Pearson AT. Machine learning-guided adjuvant treatment of head and neck cancer. JAMA Netw Open. 2020;3(11):1–13.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.