Breast Cancer Classification Using an Extreme Gradient Boosting Model with F-Score Feature Selection Technique

Tina Elizabeth Mathew

Government College Kariavattom, Thiruvananthapuram, Kerala, India
Email: tinamathew04@gmail.com

***Abstract*—Breast cancer is considered the most problematic of all cancers affecting women. With high incidence and mortality rates, it is ranked as the primary and most significant health hazard for women globally. Early detection of the disease is the key to ensure the survival of the patient. Several medical techniques comprising of Mammography, Magnetic Resonance Imaging, Thermography and many more are available to detect the disease. But these techniques create much stress and pain, besides employing harmful rays for detection, to the patient undergoing them. Hence for early detection other categories of techniques can be implemented. Machine- learning assisted detection and classification is one such alternative. In this paper a hyper parameter optimized extreme gradient boosting model implemented along with F-Score feature selection is proposed and the model is used for classification of the breast tumor as either malignant or benign on the Wisconsin Breast Cancer dataset. The implementation of feature importance is investigated using F-Score and this is used for selecting the most relevant features that influence the target variable and classification is based on this. Experimentation is done using different training-testing partitions and the best performance of 99.27% accuracy score was shown by the 80−20 partition by the proposed XGBoost and F-Score Model.**

***Keywords*—breast cancer, classification, extreme gradient boost, feature importance, F-score**

I. Introduction

Cancer is presently the foremost or second most contributor to premature mortality in almost all countries of the world. Considering the current trends and statistics, the incidence of all cancers combined, is presumed to double by 2070 relative to 2020 [1]. Hence it is critical that, countries instigate prevention methods and programmes through urgent action and advocacy. Study on prediction of the breast cancer burden is taken on so that a snapshot of the magnitude and distribution of the key cancer categories will be obtainable and thus will help to play a major role in the design of plans and means for supporting future Health-care Programmes. A key issue is detection of cancer at the earliest. Several medical technologies and modalities exist for detection of cancer yet, each have their own pros and cons. Availability of more hassle-free solutions will help the medical community in early diagnosis. The motivation for this study is the alarming rate at which new cancer cases are increasing worldwide [2].

Designing alternate techniques will provide additional support to the existing medical modalities. Determining the appropriate techniques and methodologies for the early detection of cancer still remains, among the scientific community, as an unresolved and open research problem [3]. State of art disciplines providing support to medical diagnosis, prediction and classification are Machine Learning (ML) [4] and Data Mining (DM). ML and DM techniques have found widespread use in the healthcare field [5, 6]. Several Machine Learning techniques are seen to be implemented and applied for disease diagnosis [7, 8]. These disciplines are part of a broader domain Artificial Intelligence (AI). AI is a ubiquitous, omnipresent and advancing technology in our present day lives. Artificial Intelligence can play a pivotal role in Oncology and in the near future it may be considered as the sixth sense for an oncologist [9].

A major concern in cancers affecting women is Breast cancer. Breast cancer is a non-communicable. predominant type of cancer in women and currently the first in incidence and mortality in almost all countries of the globe. Primarily affecting women, it is curable, and survivability can be ensured if detected at the earliest. As an assistive practice, machine learning techniques, specifically supervised learning methods, are seen to be suitable for breast cancer detection, prediction [10] and classification process [11]. Many techniques like Support Vector Machines [12], Logistic Regression [13], Artificial Neural Networks, *k*-NN [14], Decision Trees [15, 16] and many more have been applied for Breast cancer classification. These techniques have generally seen to be capable in distinguishing the benign and malignant breast tumours, Hence AI assisted techniques such as Machine Learning and Data Mining tools can constitute a technological armamentarium for medical practitioners. These techniques are capable and central in formulating clinical decision support systems.

Manuscript received July 1, 2022; revised October 21, 2022.; accepted December 14, 2022; published April 26, 2023.

Even though there are many cases which implement machine learning classifiers and they provide good classification, they are still not necessarily adopted by medical professionals. Hence it is critical that, new, intuitive, yet simpler techniques be rolled out so that they are easily adoptable and usable with the daily medical workflow.

A recently developed supervised learning technique is eXtreme Gradient Boosting or XGBoost (XGB) in short. It is a technique which is rapidly gaining importance in the machine learning field owing to its exceptional performance in numerous domains. XGB utilizes the gradient descent algorithm and this helps in performance improvement when compared to other ML classifiers such as SVM, LR and so in. Literature shows it has a stable performance when compared to models such as SVM. It is applicable to large and small datasets equally which is not the case in most ML classifiers. Clinical decision-making systems need to be as precise as possible in their predictions. Hence XGB is chosen in this study to investigate its suitability for Breast Cancer classification

In prediction problems involving unstructured data such as images, text, etc., supervised learners like artificial neural networks tend to outperform all other algorithms or frameworks applied. However, when it comes to small-to-medium structured or tabular data, decision tree-based algorithms are considered best-in-class as of now. It is seen to have shown outstanding results across different problems such as motion detection [17], malware classification [18], customer behaviour analysis [19], sales predictions [20] and many more. XGB, implemented in this study, is a boosting ensemble decision tree classifier. The suitability of the classifier is examined and a model with F Score feature selection and hyperparameter optimization using log loss is developed in this study.

The major contributions of this paper are:

- The study proposes an XGB model for Breast cancer classification combined with F-Score feature selection technique.
- The model performance is evaluated on various training- testing sets.
- Selection of important features for the model is done and the most important attributes are selected based on F-Score and used to identify features influencing the target class.

The organization of the remaining part of the study is as follows Section II discusses related work. The materials and methods proposed are in Section III. Results and discussions done are provided in Section IV and Section V contains conclusions and recommendations made.

II. Literature Review

Artificial Intelligence (AI) and its subdomains, data mining, machine learning, and deep learning have penetrated deeply into all arenas of our day-to-day life and are presently the most rapidly evolving areas. AI and its allied techniques are seen to possess the potential in identifying and diagnosing diseases, as suggested by Liew and Hameed *et al.* [21]. Several classification algorithms have been applied in disease diagnosis in general, including Breast cancer classification too. The applications of a few of these techniques are described below. Gao [19] identified XGB as a powerful prediction method for breast cancer image classification. They proposed a XGB and Deep Learning (DL) technique for binary classification of Breast cancer into malignant and benign and also used the same for a multiclass classification, identifying the category of the malignancy. They used DenseNet201 a CNN model and replaced the fully connected layer by XGB. Histopathological images were used as data. The model performed with an accuracy of 97% but parameter optimization was not done and the time taken and memory efficiency of the model was not taken into consideration and it needs more exploration. Also, a common issue with Deep learning models is that they perform well with large training data and is not suitable for small datasets. Abdulkareem *et al*. [22], used RFE feature selection with XGB for breast cancer classification and obtained an accuracy of 99.02% for a five reduced feature set. When RFE is used with tree models the correlation of features is to be considered and highly correlated features are to be avoided. RFE uses as parameters a subset of the features and this subset size has to be provided. Besides RFE is a greedy method. In their proposed work, Bhattacharya *et al.* [23] proposed a hybrid Principal Component Analysis (PCA)-firefly based machine learning model to classify Intrusion Detection System (IDS) datasets. The model first performed One-Hot encoding for transforming the IDS datasets. The hybrid PCA-firefly algorithm is then used for dimensionality reduction. Classification is then implemented by the XGB algorithm on this reduced dataset. A comprehensive evaluation was done with the state-of-the-art machine learning techniques and it justified the superiority of their proposed approach. Desdhanty and Rustam [24] used Genetic Algorithm as feature selection with 2 classifiers Random Forest and XGB for liver cancer classification. The result using 20% testing data, illustrated that XGB with Genetic Algorithm gave the highest accuracy of 82%. Though, genetic algorithms are good optimizers obtaining the appropriate objective function and correct operators are important and this is computationally expensive and as data grows scaling will be complex. Hou *et al.* [25] compared the performance of four machine learning algorithms- LR, Random Forest, DNN and XGB- on predicting breast cancer among Chinese women using 10 breast cancer risk factors. XGB was seen to be the most suitable classifier with better performance measures on 7127 control and experimental cases. They concluded that XGB was a suitable classifier for breast cancer prediction. They implemented the models with hyperparameter optimization using dropout and regularization techniques. The issue they faced was the imbalance of the dataset which affects the performance of the model. Kabiraj *et al.* [26] compared two ensemble classifiers — Random Forest and XGB to predict breast cancer. A total of 275 instances with 12 features were used for this analysis. Random forest algorithm gave an accuracy of 74.73% accuracy and XGB produced an accuracy of

73.63%. The study was done on a small sample and requires more exploration. Hyper parameter tuning and optimization was not done. Likitha *et al*. [27] compared various machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, Logistic Regression, Gaussian Naive Bayes, Random Forest and XGB Classifier algorithms for breast cancer classification using the Wisconsin dataset. They used ANOVA f test to identify the best subsets of features that influenced the target and obtained the highest accuracy of 98.25% with the XGB classifier. Mangukiya, *et al.* [28] compared different machine learning algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB), k Nearest Neighbours (k-NN), Adaboost, XGB and Random Forest using Wisconsin breast cancer Dataset. The experimental results showed that XGB offered the highest accuracy (98.24%) with the lowest error rate. Michael *et al.* [29] proposed a Computer Aided Diagnosis (CAD) system to generate an optimized algorithm. Five machine learning classifiers were used to classify malignant versus benign tumours. Hyperparameter optimization was done by the Bayesian optimization using tree-structured Parzen estimator. The LightGBM classifier was seen to perform better than the other four classifiers used, achieving 99.86% accuracy, 100.0% precision, 99.60% recall, and 99.80% for the FI score. Ozmen and Ozcan [30] employed XGB and Artificial Neural Network (ANN) algorithms by hybridizing with Genetic Algorithm (GA). The performance analysis of the proposed approaches was performed using Wisconsin breast cancer dataset. The numerical results illustrated that the proposed hybrid XGB-GA approach significantly outperformed the classical prediction algorithms besides achieving the best classification accuracy. The issue with GA is its computational complexity. Phankokkruad [31] proposed the cost-sensitive XGB model, improved version of the XGB model in conjunction with cost-sensitive learning to classify four breast cancer datasets that contained imbalanced data. In the experiment, they determined the best parameters on each dataset by using hyperparameters optimization techniques by applying random search. The results indicated that the cost-sensitive XGB model improved classification accuracy in four datasets. Prastyo *et al.* [32] compared eight different machine learning algorithms: Gaussian Naïve Bayes (GNB), k-Nearest Neighbours (K-NN), Support Vector Machine (SVM), Random Forest (RF), AdaBoost, Gradient Boosting (GB), XGB, and Multi-Layer Perceptron (MLP). They used the Breast Cancer Wisconsin datasets, confusion matrix, and 5-folds cross-validation. The experimental results showed that XGBoost displayed the best performance with an accuracy of (97.19%), recall of (96.75%), precision of (97.8%), F1-score of (96.9%), and AUC of (99.61%). They concluded that XGB is the most effective method to predict breast cancer in the Breast Cancer Wisconsin dataset. Sinha *et al.* [33] used various machine learning classification techniques like Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), Adaboost Classifier and XGB Classifier for the classification of benign and malignant tumour and identified that the XGB classifier provided the best accuracy of 98% against all other classifiers.

The various studies reviewed in this study explore the possibility of ML and DM techniques in disease diagnosis and prediction. The various methods involved have different constraints such as some cannot tolerate missing data, irrelevant data others have issues when the data is skewed, some need datasets with large size whereas, some are appropriate for smaller datasets. Hence the models produced cannot be generalized and show varying performance with datasets from different domains The studies also highlight the suitability and potential of XGB over other ML methods, and XGB is being seen to perform better than DL methods for classification problems. XGB is found to have facilities for cross validation and identifying feature importance and since it makes use of the gradient descent algorithm, it has a correction mechanism that rectifies errors found in the model created. It handles large data sets well and is unaffected by multicollinearity. Above all, it is an ensemble method and hence instead of dealing with a single model a group of models are aggregated to produce the resultant output. Besides XGB is seen to showcase state of art results in many areas such as text classification [34] malware classification [35], online user purchase prediction [36] and so forth. The proposed study explores the XGB model with simple feature selection methods for better accuracy.

III. Materials and Methods

A. Dataset Used

The Wisconsin Breast Cancer dataset available in the UCI repository is used in this study. It comprises of 699 instances and 11 attributes. It has 16 instances with missing values. These are omitted. Hence 683 instances are used. The target variable is class which has two values, 2 for benign and 4 for malignant. One variable Id number is avoided as it provides no relevance to the study. The remaining 9 variables have values ranging between 1–10. The malignant class contains 239 instances and the benign class has 444 instances. The dataset which is freely available in the UCI repository was created by a physician Dr William H Wolberg of the University of Wisconsin Hospitals Madison. The dataset comprises of cytological features from breast tissues obtained from an FNA slide.

B. eXtreme Gradient Boosting

eXtreme Gradient Boosting abbreviated XGB is a decision tree based boosting ensemble and supervised machine learning classifier. It can be used for classification and regression purposes [37]. Introduced in 2016, it is now gaining fast and wide popularity in classification problems in various domains. As in the name it practices an extreme approach with gradient boosting. Gradient boosting utilizes the concept of Additive Modelling where, a new decision tree is added one at a time to a model that shows minimum loss using gradient descent. Existing trees in the model remain intact and untouched and this slows down the overfitting rate. The output of the new tree is then combined with the output of

existing trees until either the loss is minimized below a threshold value or a specified limit of trees is reached. Tiwari *et al.* [38] suggested XGBoost as a best model for breast cancer classification.

XGB is a decision-tree-based boosting ensemble Machine Learning algorithm that uses a gradient boosting framework. Developed by Tianqi Chen, it is an ensemble tree method that applies the principle of boosting weak learners, CARTs in general, using the gradient descent architecture. It involves a depth-first approach that helps to improves the computational performance significantly. It is considered as an optimized gradient boosting algorithm as it implements parallel processing, tree pruning, handles missing values effectively, has inbuilt cross validation, takes care of outliers and uses regularization to avoid overfitting or bias. XGB uses the max_depth parameter instead of criterion first, and performs pruning trees backward. The advantage of XGB over other Machine Learning classifiers is that it handles large datasets effectively and hence a model created using small datasets have the potential to be scaled for larger datasets. Earlier studies show that it has significant upper hand in the case of execution speed and model performance using minimal quantity of resources [39]. Besides it permits regularization techniques to avoid overfitting,

Various studies using XGB indicate that the classifier helps in improved model performance and better execution speed. The classifier is seen to be best for prediction-classification, regression problems, fraud detection, customer prediction and a wide variety of data science challenges.

C. F-Score

Feature selection techniques can be categorized as supervised techniques and unsupervised techniques. A further classification done on Supervised techniques categorize them into as filter, wrapper and intrinsic types. Similarly, Unsupervised feature selection methods available are Variance, Mean Absolute Difference, Dispersion ratio, Laplacian Score and many combination techniques such as Laplacian score and distance-based entropy, Multi cluster Feature Selection and so on. The supervised feature selection technique F-Score fits into the category of filter-based techniques owing to its statistical nature. It is based on the F distribution. It makes use of Mutual Information criteria for ranking of features. Henceforth, F-Score is categorized as a univariate statistical based method which can be used for feature selection in binary problems [38]. It is calculated as shown in Eq. (1). The selected features have to satisfy the condition and the mean value of the F-Score of all the features (y) set as the threshold [40, 41].

$$F-Score=\frac{(x_{i+}-y)^2+(x_{i-}-y)^2}{\frac{1}{n+1}+\sum(x_{i+}-z)^2+\frac{1}{n-1}\sum(x_{i-}-a)^2} \quad (1)$$

where x, y, z, a represent the means of whole, the positive and negative instances and x_{ki} is the i^{th} feature of the k^{th} positive or negative instance of a vector x_k for $k = 1, 2, \ldots, m$ having $n+$ and $n-$ positive and negative instances. The positive and negative classes are represented by the plus (+) and minus (−) signs. The numerator gives the difference between the positive and negative class and the denominator gives the difference between each of the classes. If the features have F-Score values greater than the given threshold, then those features are considered as relevant features and are selected. Those features having a F-score value which is lower than the threshold, will represent the irrelevant features in classification with respect to the target variable class [35, 41]. Irrelevant features get discarded from the feature space and then the remaining feature set is used with the chosen classifier. Bigger F-Score values represent better discriminative quality of the corresponding feature. Advantage of filter methods is that they do not depend on the classifier used, albeit the choice of the optimal feature set is tricky and is to be done carefully [42].

D. Proposed Model

A model using XGB is proposed (Fig. 1). Initially the dataset is pre-processed and instances with missing values are removed. Parameter optimization is still needed even though XGB has its own default tuning mechanisms. The model has to adapt explicitly to the precise characteristics of each specific dataset. Hence, the model is optimized by using the log loss evaluation criteria for validation of data. Similarly, the best feature set is needed for producing an optimized model. To achieve this objective F-Score of the features are calculated and based on the obtained feature importance the model is evaluated. The performance of the dataset using various training and testing partitions is also examined. For this. four sets of training-testing partitions are chosen − 80−20, 70−30, 60−40 and 50−50. The performance is evaluated. Besides, in each set the importance of the features used are calculated using F score. Slight variations in the ranking of the features are noticed. Once the training data is modelled, it is evaluated using the test data to predict the outcomes. The main conundrum to be solved in the classification task with the XGB classifier is to predict the labels of the provided test data accurately. Performance metrics like Precision, Recall, F1-score, Accuracy are used to evaluate the classifier. The classification error and log loss plots are also used plotted and used for evaluation.

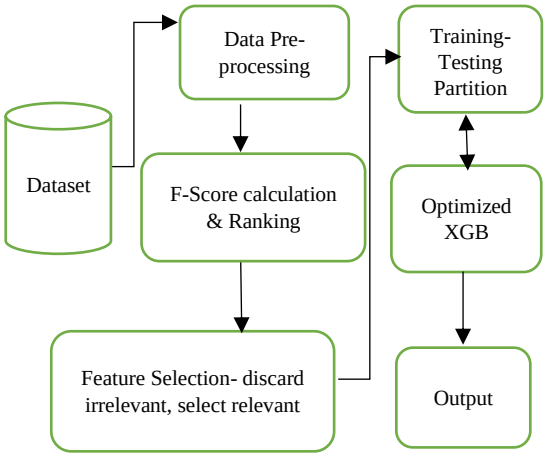


Figure 1. Proposed workflow.

IV. Results and Discussions

The XGB classifier is trained and tested on the Wisconsin Breast Cancer dataset. The performance of the classifier on different partition of datasets is examined. The relevant features are identified using F- Score which implements a simple feature selection technique Hyperparameter optimization is done using log loss. Optimization of the parameters is essential for better performance of the classifier. To achieve the best values for the parameters a grid search is applied. The optimal parameter values identified are then used with the XGB classifier. The results of training the dataset using XGB classifier with various training-testing partition sets −80−20, 70−30, 60−40, 50−50− are illustrated in Table I. The performance is evaluated using various performance metrics. Precision, Recall, F1 Score, Accuracy scores obtained for each set is shown in the tables.

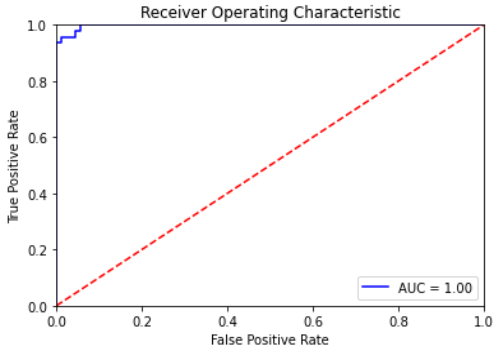
TABLE I. Performance Measures

Train-Test Ratio	80–20		70–30		60–40		50–50	
Class	2	4	2	4	2	4	2	4
Precision	1	0.98	0.99	0.93	0.99	0.95	1	0.95
Recall	0.99	1.0	0.96	0.99	0.97	0.98	0.97	0.99
F1-Score	0.99	0.99	0.98	0.96	0.98	0.96	0.98	0.97
Support	90	47	133	72	178	96	220	112
Accuracy	99.27		97.07		97.44		97.66	
Average Accuracy	97.86							

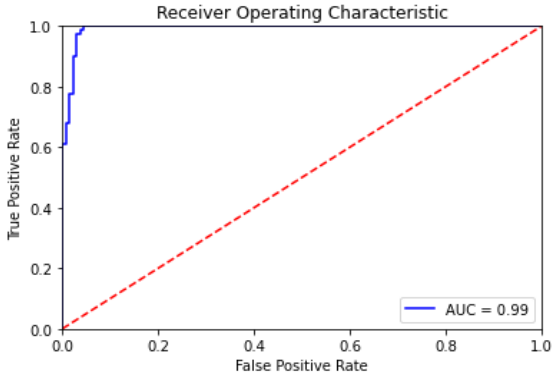
The 80−20 partition set displayed an accuracy of 99.27, an average score of 0.99 for Precision, Recall and F1 score. On the test set. The 70−30 partition, test set obtained an accuracy of 97.07% with average precision of 0.96, and 0.97 each for recall and F1 score. The 60−40 partition test set displayed an accuracy of 97.44% with average precision at 0.97, Recall at 0.98 and F1-Score at 0.97. The 50−50 partition test set illustrated an accuracy of 97.66% with precision and F1-Score at 0.97 each and recall at 0.98.

The ROC of the four training-testing partitions explored are depicted in the Fig. 2(a−d). The ROC AUC of the 80-partition is 1.00. The average accuracy obtained from all the partitions together is 97.86%.

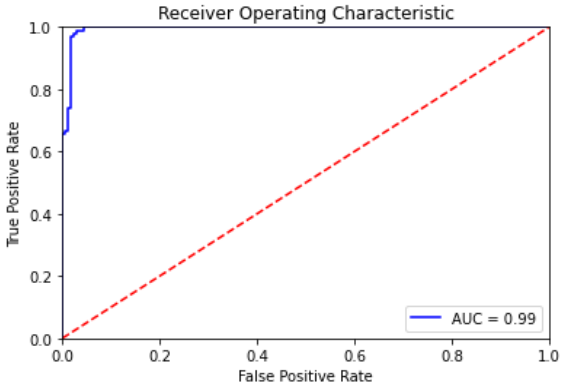
The ROC AUC of the 70−30, 60−40 and 50−50 sets are 0.99 each. The AUC indicates the classifiers capability to distinguish the two classes. It measures the degree of separability. The best value is displayed by the 80−20 partition set.



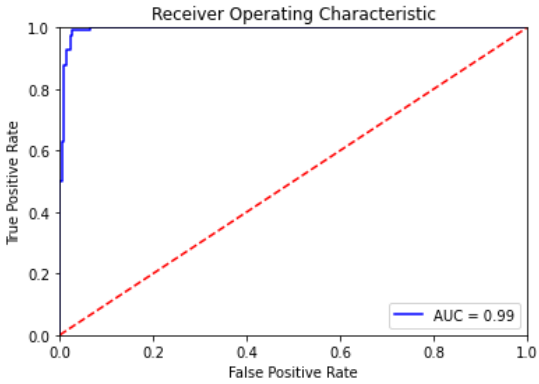
(a) ROC of train-test partition 80−20



(b) ROC of train-test partition 70−30

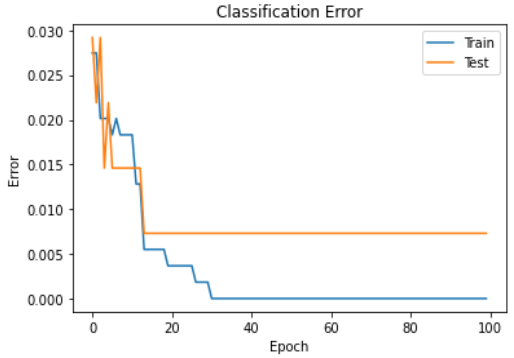


(c) ROC of train-test partition 60−40

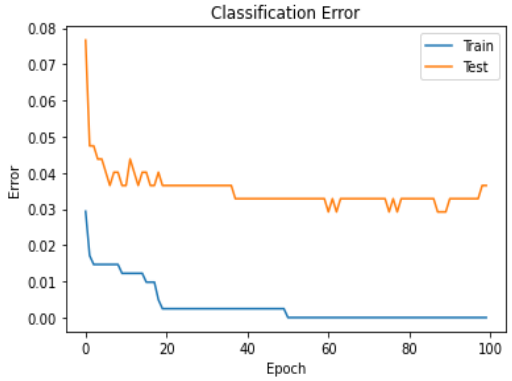


(d) ROC of train-test partition 50−50

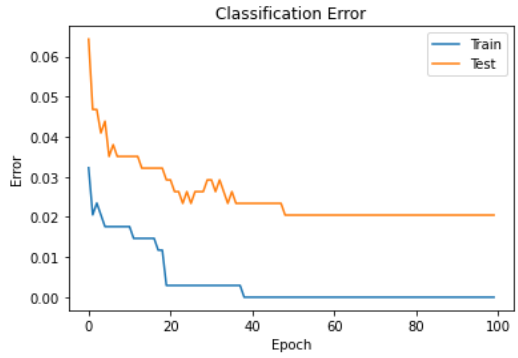
Figure 2. The ROC of the four training-testing partitions are explored.



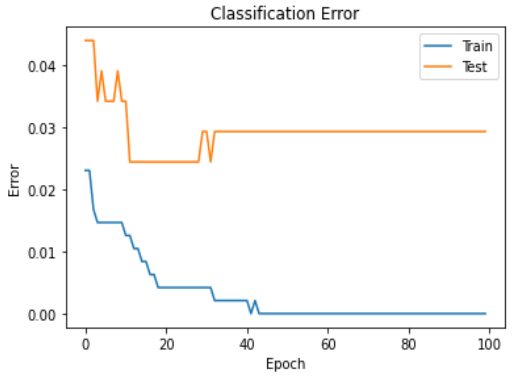
(a) Classification error vs epoch of train-test partition 80−20



(b) Classification error vs epoch of train-test partition 70−30



(c) Classification error vs epoch of train-test partition 60−40

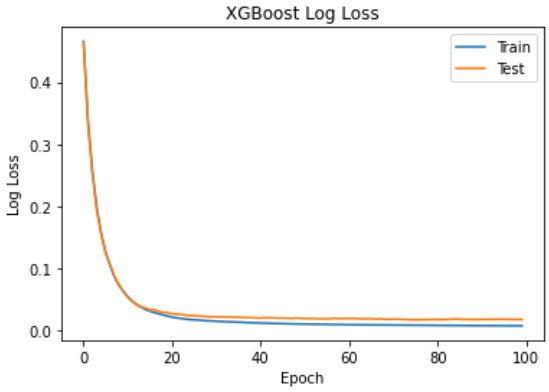


(d) Classification error vs epoch of train-test partition 50−50

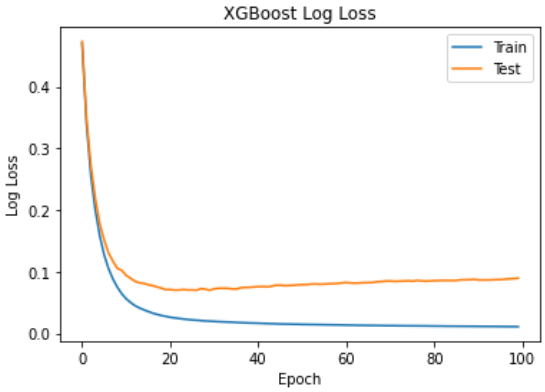
Figure 3. The classification error obtained during training and testing of the four partition sets.

The classification error obtained during training and testing of the four partition sets are also depicted in Fig. 3(a–d). It shows the classification error of the XGB model for each epoch on the different partitions of training and test datasets.

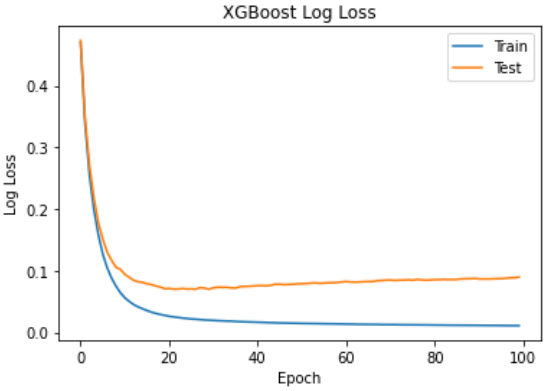
Log loss is plotted against the Epochs for each of the four partitions and the plots obtained are depicted in Fig. 4(a–d). The plots help to decide when pruning is needed. The log loss is an evaluation measure that checks the performance of a binary classification model. It is a measure of the amount of divergence of the predicted probability with the actual label. A lesser log loss value, represents a perfect model. The Log loss function helps to evaluate the performance according to the correct predictions besides penalizing the wrong predictions based on the predicted probabilities. By observing the log loss curve early stopping of training can be decided to prevent overfitting. The training is stopped between 10 to 20 epochs.



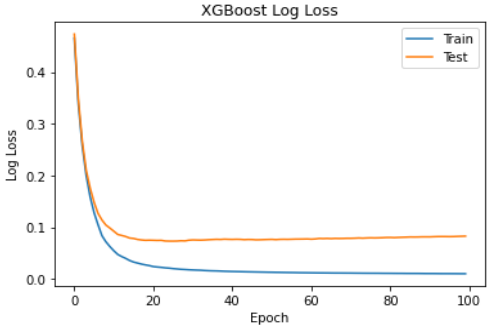
(a) Log loss of train-test partition 80−20



(b) Log loss of train-test partition 70−30

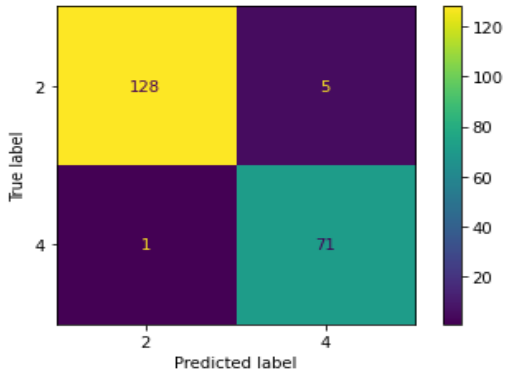


(c) Log loss of train-test partition 60−40

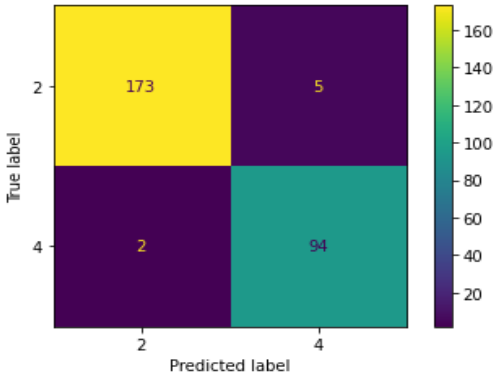


(d) Log loss of train-test partition 50−50

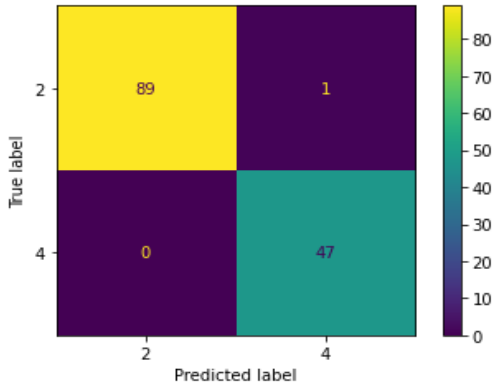
Figure 4. Log loss is plotted against the epochs for each of the four partitions and the plots obtained are depicted.



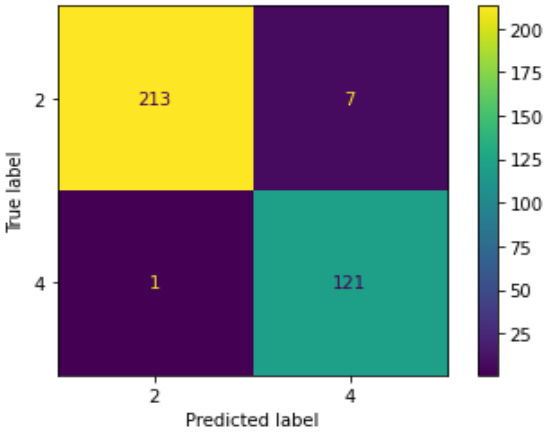
(a) Confusion matrix of train-test partition 80−20



(b) Confusion matrix of train-test partition 70−30



(c) Confusion matrix of train-test partition 60−40

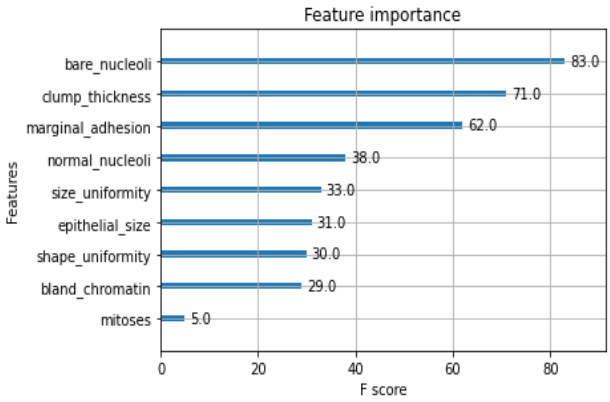


(d) Confusion matrix of train-test partition 50−50

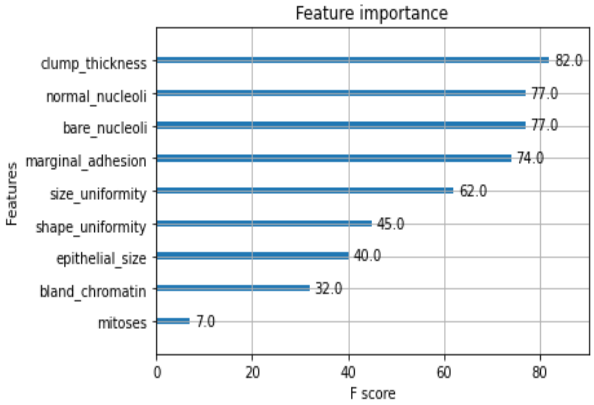
Figure 5. Confusion matrix for each of the four partition sets are depicted.

Confusion matrix for each of the four partition sets are depicted in Fig. 5(a−d). Misclassification was seen higher for the negative class in all the train-test partitions and comparatively the positive class was better classified. Misclassification of the positive class is considered more serious than vice versa. The 80−20 set had 0 instances of positive misclassification and 1 instance of negative class misclassification. The 50−50 partition illustrated more negative misclassifications among all four sets. The remaining two partitions had similar negative class misclassification instances but the positive class misclassification was more for the 60−40 set when compared with other partitions.

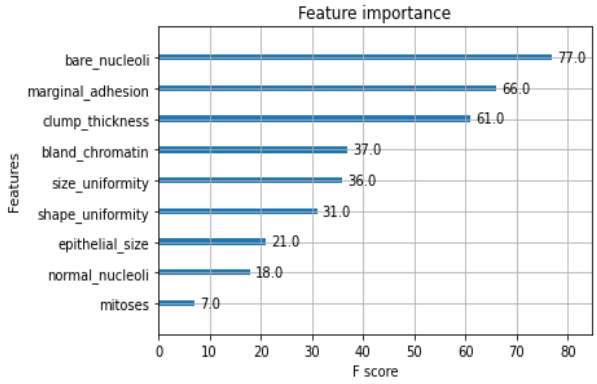
The importance of the 9 features used in classification are ranked on the basis of F1-score for the four partitions is depicted in Fig. 6(a−d). In three of the four training-testing partitions the feature bare nucleoli was ranked as the most important feature and mitosis was ranked as the least important feature in all training- testing sets. 10-fold stratified cross validation is being done to generate the feature importance using F score. Accuracy score was best for the 80−20 partition with 99.27% and here Clump thickness was ranked first followed by normal nucleoli, bare nucleoli as third important, marginal adhesion as fourth, size uniformity as fifth, shape uniformity as sixth epithelial size as seventh bland chromatin as 8th important and finally mitoses.



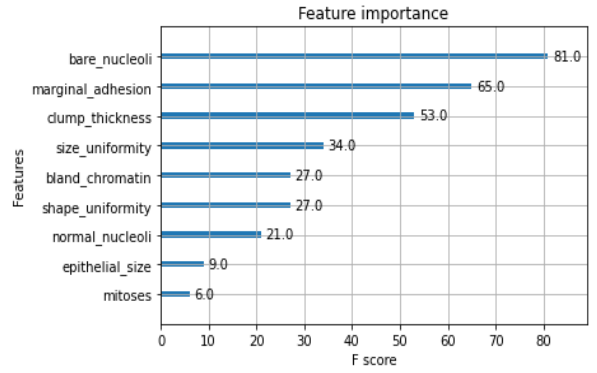
(a) Feature importance vs F-score of 80−20 train-test split



(b) Feature importance vs F-score of 70−30 train-test split



(c) Feature importance vs F-score of 60−40 train-test split



(d) Feature importance vs F-score of 50−50 train-test SPLIT

Figure 6. Feature importance vs F-score of the four train-test partition sets are depicted.

Features of the dataset are assessed individually and rated based on their F-Scores. The feature importance plots of the various training-Testing partition sets based on the F-Score feature ranking is consolidated in Table II. Higher F-Score represents higher importance for the feature. The discriminative capability of the feature is depicted by the F-Scores. The features of the dataset are — Bare Nucleoli, Normal Nucleoli, Marginal Adhesion, Clump thickness, Size Uniformity, Epithelial Size, Shape Uniformity, Bland Chromatin, and Mitoses. The feature importance score is illustrated in the figures and it depicts the influence it has on the target class. The results depict that F-Score can be implemented effectively for breast cancer classification.

TABLE II. FEATURE RANKING

Features	Training-Testing partition			
	80−20	70−30	60−40	50−50
Bare Nucleoli	3	1	1	1
Normal Nucleoli	2	4	8	7
Marginal Adhesion	4	3	2	2
Clump Thickness	1	2	3	3
Size Uniformity	5	5	5	4
Epithelial Size	7	6	7	8
Shape Uniformity	6	7	6	6
Bland Chromatin	8	8	4	5
Mitoses	9	9	9	9

Table III illustrates the values of the hyperparameters of the XGB classifier that is being used in the study.

TABLE III. HYPERPARAMETERS USED

Hyperparameters	Description	Values
max_depth	Maximum tree depth	6
n_estimators	No. of trees	100
learning_rate		0.300000012
eval_metric	Metric for data validation	LogLoss
lambda	L2 Regularization parameter	1
scale_pos_weight	To control +ve and -ve samples	1

The proposed model is compared with state of art technologies in literature as in Table IV and the proposed model was seen to outperform all models with the best accuracy score. An XGB classifier with F-test feature selection was implemented on the WBCD dataset by Kabiraj *et al.* [26] and they obtained an accuracy of 98.25%. Tiwari *et al.* [38] used the XGB classifier and CNN on varying environments and cores using the Higgs 1 M dataset and the best performance obtained was 97.4%. Song, Li, and Wang [43] used a deep learning model with DCNN and obtained an accuracy of 92.8% on the DDSM dataset. Thongsuwan *et al.* [44] used a Deep learning CNN model with XGB on the WBC original dataset and obtained an accuracy of 97.4%. Mathew [45] used an improved random forest model on the WBC Dataset and obtained an accuracy of 97.9%. In their proposed models [31, 32] used several ML classifiers and each studies obtained XGB as the best classification model with 97.19% and 98% respectively. Compared to the ML and deep learning classifiers the proposed XGB model was seen to outperform in all cases.

TABLE IV. COMPARISON WITH LITERATURE

Author	Model	Dataset Used	Accuracy
Likitha [27]	XGB +Ftest Feature selection	WBCD	98.25%
Prastyo [32]	ML models	WBCD	97.19
Sinha [33]	Web based XGB prediction system	WBCD	98%
Chen [39]	Deep Learning CNN+ XGB	Higgs 1M	97.4%
Song [43]	Deep Learning model with DCNN	DDSM	92.8%
Thonguswan [44]	Deep Learning CNN and XGB	WBDC	97.4%
Mathew [45]	Random Forest Model	WBCD	97.9
Proposed Model	XGB+F1 score feature selection	WBCD	99.27%

V. Conclusion

In this study, the proposed XGB and F-Score model was used to classify breast cancer tumours as malignant or benign. Use of F-Score helped to improve the accuracy of the model. Hyperparameter optimization was attained using log loss evaluation metric. Feature importance was taken into consideration. This was implemented on different train-test partitions with different combinations of features and the best performance of 99.27% was displayed by the Feature set on partition 80−20. The feature importance of each feature in relation with the target class was utilized for better classification. The proposed model was seen to have superior performance. However, further investigation is to be done on the grouping done on the feature sets and they need to be validated with medical practitioners. A limitation with F-Score is that it does not consider feature interaction, that is, the possibility of better performance when individually poor features interact with each other features or mutual information. Besides, F-Score is suitable only for binary classification The study highlights the prospects of implementing XGB classifier as a model for breast cancer classification. As future work different simple yet well performing feature selection techniques, as well as unsupervised heuristics for supervised models are to be explored so as to improve model feature selection performance. The model handled class imbalance well, albeit the misclassification of the negative class was seen to have a higher rate than vice versa. Class balancing techniques can be implemented to address this issue. The study has some limitations; for instance, the model is evaluated on small-size datasets only, and it is imperative to validate the model on considerably large-sized datasets. Also, the proposed approach has been implemented using the Wisconsin breast cancer dataset; for the sake of generalizability, the proposed model needs to be validated further on 2-class breast cancer as well as on other disease datasets and problems in other domains as well. In addition, combining XGB with different deep learning techniques is a probable area for investigation in future.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The author wishes to acknowledge Dr. William H. Wolberg of the University of Wisconsin Hospitals, Madison for the dataset that is used in this study.

References

[1] I. Soerjomataram and F. Bray, "Planning for tomorrow: Global cancer incidence and the role of prevention 2020–2070," *Nat. Rev. Clin. Oncol.*, vol. 18, no. 10, pp. 663–672, 2021.

[2] F. Islami, *et al.*, "Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States," *CA. Cancer J. Clin.*, vol. 68, no. 1, pp. 31–54, 2018.

[3] S. A. Amin, H. Al-Shanabari, R. Iqbal, and C. Karyotis, "An intelligent framework for automatic breast cancer classification using novel feature extraction and machine learning techniques," *J. Signal Process. Syst.*, pp. 1–11, 2022.

[4] B. Thakur and N. Kumar, "Prediction, detection and recurrence of breast cancer using machine learning based on image and gene datasets," *Recent Innov. Comput.*, pp. 263–273, 2022.

[5] S. Mall, A. Srivastava, B. D. Mazumdar, M. Mishra, S. L. Bangare, and A. Deepak, "Implementation of machine learning techniques for disease diagnosis," *Mater. Today Proc.*, vol. 51, pp. 2198–2201, 2022.

[6] T. E. Mathew and O. SugelAnandh, "Lung cancer classification using extreme Anfis with red fox optimization algorithm," *Neuro Quantology*, vol. 20, no. 6, pp. 1839–1846, 2022.

[7] S. Manimurugan, *et al.*, "Two-stage classification model for the prediction of heart disease using IoMT and artificial intelligence," *Sensors*, vol. 22, no. 2, p. 476, 2022.

[8] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," *Artif. Intell. Med.*, vol. 127, 102276, May 2022.

[9] V. Talwar, K. S. Chufal, and S. Joga, "Artificial intelligence: A new tool in oncologist's armamentarium," *Indian J. Med. Paediatr. Oncol.*, 2021.

[10] S. Gupta, M. K. Gupta, and R. Kumar, "A novel multi-neural ensemble approach for cancer diagnosis," *Appl. Artif. Intell.*, pp. 1–36, 2021.

[11] T. E. Mathew and K. A. Kumar, "A logistic regression based hybrid model for breast cancer classification," *Indian J. Comput. Sci. Eng.*, vol. 11, no. 6, pp. 899–906, 2020.

[12] T. E. Mathew, "A comparative study of the performance of different support vector machine kernels in breast cancer diagnosis," *Int. J. Inf. Comput. Sci.*, vol. 6, no. 6, pp. 432–441, 2019.

[13] T. Mathew, "A logistic regression with recursive feature elimination model for breast cancer diagnosis," *Int. J. Emerg. Technol.*, vol. 10, no. 3, pp. 55–63, 2019.

[14] T. E. Mathew and K. S. A. Kumar, "A modified-weighted-k-nearest neighbour and cuckoo search hybrid model for breast cancer classification," *Indian J. Comput. Sci. Eng.*, vol. 12, no. 1, pp. 166–177, 2021.

[15] T. E. Mathew, "Simple and ensemble decision tree classifier based detection of breast cancer," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 1628–1637, 2019.

[16] T. E. Mathew, "An optimized extremely randomized tree model for breast cancer classification," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 16, 2022.

[17] M. Alaziz, Z. Jia, R. Howard, X. Lin, and Y. Zhang, "In-bed body motion detection and classification system," *ACM Trans. Sens. Netw. TOSN*, vol. 16, no. 2, pp. 1–26, 2020.

[18] S. Euh, H. Lee, D. Kim, and D. Hwang, "Comparative analysis of low-dimensional features and tree-based ensembles for malware detection systems," *IEEE Access*, vol. 8, pp. 76796–76808, 2020.

[19] J. Gao, W. Sun, and X. Sui, "Research on default prediction for credit card users based on XGBoost-LSTM model," *Discrete Dyn. Nat. Soc.*, 2021.

[20] Z. Xia, S. Xue, L. Wu, J. Sun, Y. Chen, and R. Zhang, "ForeXGBoost: Passenger car sales prediction based on XGBoost," *Distrib. Parallel Databases*, vol. 38, no. 3, pp. 713–738, 2020.

[21] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Mach. Learn. Appl.*, vol. 6, 100154, 2021.

[22] S. A. Abdulkareem and Z. O. Abdulkareem, "An evaluation of the Wisconsin breast cancer dataset using ensemble classifiers and RFE feature selection," *Int. J. Sci. Basic Appl. Res.*, vol. 55, no. 2, pp. 67–80, 2021.

[23] S. Bhattacharya, *et al.*, "A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU," *Electronics*, vol. 9, no. 2, p. 219, 2020.

[24] V. S. Desdhanty and Z. Rustam, "Liver cancer classification using random forest and extreme gradient boosting (XGBoost) with genetic algorithm as feature selection," in *Proc. 2021 International Conference on Decision Aid Sciences and Application (DASA)*, 2021, pp. 716–719.

[25] C. Hou, *et al.*, "Predicting breast cancer in Chinese women using machine learning techniques: Algorithm development," *JMIR Med. Inform.*, vol. 8, no. 6, p. e17364, 2020.

[26] S. Kabiraj, *et al.*, "Breast cancer risk prediction using XGBoost and random forest algorithm," in *Proc. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–4.

[27] B. Likitha, J. Nakka, J. Verma, and N. S. Naik, "Prediction of breast cancer analysis using machine learning algorithms and XGBoost technique," in *Proc. International Conference on Information Processing*, 2021, pp. 298–313.
[28] M. Mangukiya, A. Vaghani, and M. Savani, "Breast cancer detection with machine learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 2, pp. 141–145, 2022.
[29] E. Michael, H. Ma, H. Li, and S. Qi, "An optimized framework for breast cancer classification using machine learning," *BioMed Res. Int.*, vol. 2022, 2022.
[30] P. Ozmen and T. Ozcan, *Diagnosis of Breast Cancer Using Novel Hybrid Approaches with Genetic Algorithm*, Springer Cham, 2021, pp. 589–595.
[31] M. Phankokkruad, "Cost-sensitive extreme gradient boosting for imbalanced classification of breast cancer diagnosis," in *Proc. 2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2020, pp. 46–51.
[32] P. H. Prastyo, I. G. Y. Paramartha, M. S. M. Pakpahan, and I. Ardiyanto, "Predicting breast cancer: A comparative analysis of machine learning algorithms," in *Proc. International Conference on Science and Engineering*, 2020, vol. 3, pp. 455–459.
[33] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, "Developing a web based system for breast cancer prediction using XGboost classifier," *Int. J. Eng. Res. Technol. IJERT*, vol. 9, 2020.
[34] H. El Rifai, L. Al Qadi, and A. Elnagar, "Arabic text classification: the need for multi-labeling systems," *Neural Comput. Appl.*, vol. 34, no. 2, pp. 1135–1159, Jan. 2022.
[35] S. D. Ashwini, M. Pai, and J. Sangeetha, "Android malware classification based on static features of an application," in *Advances in Computing and Network Communications*, Springer, 2021, pp. 567–581.
[36] J. Lee, O. Jung, Y. Lee, O. Kim, and C. Park, "A Comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 5, pp. 1472–1491, 2021.
[37] Y. Ono and Y. Mitani, "Evaluation of feature extraction methods with ensemble learning for breast cancer classification," in *Proc. 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, 2022, pp. 194–195.
[38] P. Tiwari, P. Bhardwaj, A. Keprate, and A. Tyagi, "Breast cancer survival prediction using machine learning," in *Computational Intelligence in Oncology*, Springer, 2022, pp. 143–158. doi: 10.1007/978-981-16-9221-5_8
[39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[40] S. Güneş, K. Polat, and Ş. Yosunkaya, "Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 998–1004, 2010.
[41] N. Sevani, I. Hermawan, and W. Jatmiko, "Feature selection based on F-score for Enhancing CTG data classification," in *Proc. 2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 2019, pp. 18–22.
[42] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.
[43] R. Song, T. Li, and Y. Wang, "Mammographic classification based on XGBoost and DCNN with multi features," *IEEE Access*, vol. 8, pp. 75011–75021, 2020.
[44] S. Thongsuwan, S. Jaiyen, A. Padcharoen, and P. Agarwal, "ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost," *Nucl. Eng. Technol.*, vol. 53, no. 2, pp. 522–531, 2021.
[45] T. E. Mathew, "An improvised random forest model for breast cancer classification," *NeuroQuantology*, vol. 20, no. 5, pp. 713–722, May 2022.

Tina Elizabeth Mathew is an assistant professor in computer science working at Government College Kariavattom, Thiruvananthapuram, Kerala. She completed her PG in Computer Science from Mahatma Gandhi University, Kottayam and is a PhD holder from University of Kerala, she has 18 years and 6 months of undergraduate teaching experience. Her area of specialization is data mining and machine learning and has published 7 papers in various journals.