Exploring Machine Learning Classifiers for Breast Cancer Classification

Inayatul Haq[1], Tehseen Mazhar [2,*], Hinna Hafeez[3], Najib Ullah[4], Fatma Mallek[5], and Habib Hamam[5,6,7,8]

[1] School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China
[e-mail: inayatulhaq1@gmail.com]
[2] Department of Computer Science, Virtual University of Pakistan, Lahore 54000, Pakistan
[e-mail: tehseenmazhar719@gmail.com]
[3] Department of Computer Science Superior University Lahore 54000, Pakistan
[e-mail: hinnahafeezsh@gmail.com]
[4] Faculty of Pharmacy and Health Sciences, Department of Pharmacy, University of Balochistan,
Quetta 08770, Pakistan
[e-mail: najibullah9@yahoo.com]
[5] Faculty of Engineering, Université de Moncton, Moncton NBE1A3E9, Canada
[e-mail: Habib.Hamam@umoncton.ca; fatma.mallek.benamar@gmail.com]
[6] Bridges for Academic Excellence, Tunis 1001, Centre-Ville, Tunisia
[7] Hodmas University College, Taleh Area, Mogadishu 252, Somalia
[8] School of Electrical Engineering, University of Johannesburg, Johannesburg 2006, South Africa
*Corresponding author: Tehseen Mazhar

Abstract

Breast cancer is a major health concern affecting women and men globally. Early detection and accurate classification of breast cancer are vital for effective treatment and survival of patients. This study addresses the challenge of accurately classifying breast tumors using machine learning classifiers such as MLP, AdaBoostM1, logit Boost, Bayes Net, and the J48 decision tree. The research uses a dataset available publicly on GitHub to assess the classifiers' performance and differentiate between the occurrence and non-occurrence of breast cancer. The study compares the 10-fold and 5-fold cross-validation effectiveness, showing that 10-fold cross-validation provides superior results. Also, it examines the impact of varying split percentages, with a 66% split yielding the best performance. This shows the importance of selecting appropriate validation techniques for machine learning-based breast tumor classification. The results also indicate that the J48 decision tree method is the most accurate classifier, providing valuable insights for developing predictive models for cancer diagnosis and advancing computational medical research.

1. Introduction

Despite years of research, more women are being diagnosed with breast cancer. Validated risk assessment models can use mammographic density and polygenic risk to predict a woman's risk of breast cancer more accurately [1]. Breast cancer remains the dominant type affecting women, encompassing various pathological presentations, clinical characteristics, and outcomes. In the United States, it ranks as the second highest cause of cancer-related deaths [2]. **Fig. 1** depicts the breast cancer illustration.

Multiple observational studies have shown that regular mammography screening significantly decreases mortality rates associated with breast cancer [3]. Early diagnosis of breast cancer tumors can increase the chances of survival. In the domain of Machine Learning (ML) and Deep Learning (DL), Convolutional neural networks (CNNs) have emerged as effective tools for classifying breast cancer tumors in medical images. Ensemble learning methods such as Random Forest (RF) and gradient boosting can help feature engineering to improve accuracy. Radiomics is a method that extracts detailed features from medical images to help classify breast tumors more effectively. Classifying breast cancer involves examining genes, tissues, and images from scans like MRI and ultrasound. Combining data from various sources and using explainable AI and transfer learning can improve classification models. Accuracy can also be increased using some strategies, i.e., synthetic data generation, quantitative image marker identification, and data augmentation [4-7].

Some challenges observed in breast cancer classification techniques are unbalanced data, interoperability problems, scarcity of knowledge in the health domain, and confusion in annotations. Similarly, challenges may occur with robust generalization, cost considerations, computational requirements, the dynamic nature of breast cancer, and adaptation to different patient populations [5, 8, 9]. A multifaceted approach is required to resolve these issues occurring in breast cancer classification. Data augmentation and collaborative databases can boost the size and diversity of datasets [10]. Attaining interoperability in healthcare relies on standardization and the advancement of interoperable systems [11]. Data security and privacy can be maintained if encryption and access control are combined with privacy-preserving AI methods [12]. Options include crowdsourcing and semi-supervised learning to enhance annotation quality and quantity. Model interpretability is facilitated through Explainable AI (XAI) and external interpretation tools. Generalization is improved with regularization techniques and cross-validation. Clinical validation necessitates rigorous trials and collaboration with regulatory bodies. Computational resource challenges are met with cloud computing and model optimization. Given the dynamic nature of breast cancer, models must incorporate continuous learning [13, 14]. Linear Discriminant Analysis (LDA) is a method in ML that helps to separate and classify different groups by finding the most important features. It's often used in pattern recognition and ML to help classify objects or predict categories [15].

Comparing multiple ML classifiers is essential for optimizing their performance in cancer diagnosis. This analysis helps pinpoint the most effective model by assessing metrics such as accuracy and precision. It also provides valuable insights into the reliability of classifiers across various datasets, guiding the selection of robust models. Fine-tuning hyperparameters based on their impact ensures optimal model performance and considers adaptability to diverse datasets [16, 17].

The problem focused in this study is the accurate identification and classification of breast tumors using ML classifiers. This study aims to explore an efficient classifier among various classifiers for accurately classifying breast tumors. Also, the optimal splitting percentage and folding values should be determined to increase the accuracy of the classification model. **Table**

1 comprises the enumeration of acronyms.
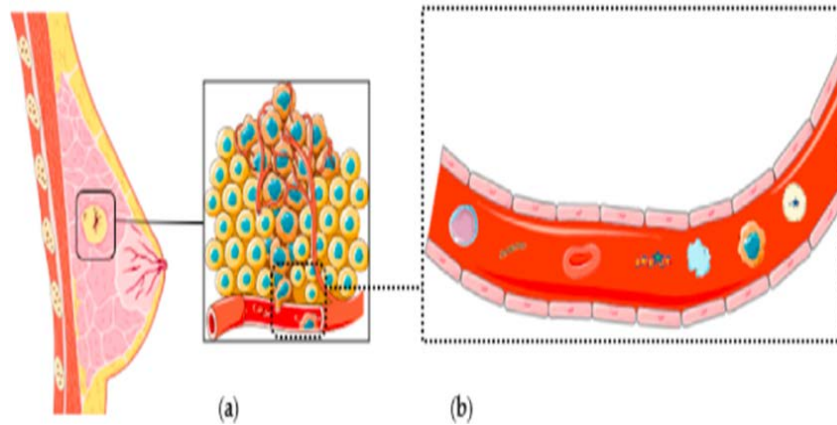
Fig. 1. Breast Cancer illustration: (a) breast cancer cells relation with the vascular system. (b) Major components present in blood [18].

Table 1. Enumeration of acronyms.

Abbreviation	Full Form	Abbreviation	Full Form
SVM	Support Vector Machine	PCA	Principal Component Analysis
DT	Decision Tree	RF	Random Forest
MLP	Multi-Layer Perceptron	K-NN	K-Nearest Neighbors
ANN	Artificial Neural Networks	DFA	Discriminant Function Analysis
LR	Logistic Regression	NDA	Normal Discriminant Analysis
ML	Machine Learning	DTC	Decision Tree Classifier
MCC	Matthews Correlation Coefficient	IGA	Improved Genetic Algorithm
LDA	Linear Discriminant Analysis	NCA	Neighbourhood Components Analysis
MAE	Random Absolute Error	ROI	Region of Interest
RAE	Relative Absolute Error	ARFF	Attribute-Relation File Format
RMSE	Root Mean Squared Error	AUC-ROC	Area Under the Receiver Operating Characteristic Curve
Weka	Waikato Environment for Knowledge Analysis	ITK	Insight Segmentation and Registration Toolkit
RRSE	Root Relative Squared Error	PSO	Particle Swarm Optimization
GS	Genetic Search	BCW	Breast Cancer Wisconsin
RAE	Relative Absolute Error	AUC	The area under the ROC Curve
SL	Supervised Learning		

2. Literature Review

Among K-NN, ANNs, LR, and RF, SVM was found to be the most accurate ML classifier for predicting breast cancer. On the other hand, ANNs outperformed other approaches with the highest accuracy of 98.57% [19]. A hybrid approach was created for feature selection that combines the advantages of feature selection methods with an enhanced GA (improved Genetic Algorithm). The findings showed that when choosing the best features, the hybrid feature selection approach is better for both single filter methods and PCA [20]. Similarly, genetic programming and ML techniques were used to create a system differentiating between benign and malignant breast malignancies. The objective of the research was to improve the learning algorithm. This study highlights the potential of genetic programming to automatically select the optimal model by combining feature pre-processing strategies and classifier algorithms [21].

A new integration method combining ML with specific selection and survival analysis based on Cox regression was presented in a study. The study aimed to identify the most useful miRNA biomarkers in different types of breast cancer [22]. The wrapper-based feature selection strategy uses PSO, GS, and a greedy step algorithm. The J48 (DT) estimator is the most accurate predictor of breast cancer using ML [23]. Diagnostics of an IoT environment based on machine learning aims to distinguish between normal and malignant tumors. To develop the classification of this method, an iterative feature selection strategy was used to identify the most important features in breast cancer data [24].

Four ML classifiers (kNN, DT, binary SVM, and Adaboost) were compared and contrasted regarding performance on the BCW dataset. The feature selection model used NCA to select and reduce the number of relevant features to reduce model complexity [25]. Using symmetrical CT scan data, several ML classifiers differentiate between images of healthy and tuberculosis-infected lungs. The MLP classifier outperforms other classifiers with 98.83% accuracy and fast execution time [26].

Naive Bayes and KNN were used to classify breast cancer. The findings indicated that the KNN method performed better and achieved high accuracy, 97.51%, and a lower error rate. On the other hand, the Naive Bayes method also showed good results, with an accuracy of 96.19%. Similarly, CNN was used to detect nodules from large numbers of images and has been evaluated to help radiologists diagnose cancer early [27].

Likewise, public data was used to build a DL model for breast cancer diagnosis and classification. The high accuracy highlights the DL model's effectiveness in accurately detecting and classifying breast cancer [28]. A unique hybrid method that integrates traditional handcrafted features with CNNs to improve the effectiveness of segmenting brain tumors [29]. Decision tree (DT) methodologies offer several advantages in medical image analysis. Firstly, their interpretability is a key strength, allowing clinicians and researchers to understand the reasoning behind each decision [30]. DT handles non-linear relationships effectively [31]. DT methods are robust to outliers, which is common in medical datasets [32]. Moreover, DT methods perform implicit feature selection, prioritizing the most informative features [33].

3. Methodology and Techniques

This section presents the data collection, data preparation, and proposed methodology.

3.1 Dataset Collection and Preparation

This study uses a publicly available dataset from GitHub [34, 35]. ARFF format files are

used in this study because they are compatible with the Weka software. This dataset was sourced from the Institute of Oncology at the University Medical Centre, provided by physicians Matjaz Zwitter and Milan Soklic. The dataset was donated by Jeff Schlimmer and Ming Tan [35]. The dataset typically contains several hundred instances, each representing a case with a set of features and a class label indicating the presence or absence of breast cancer. This data includes demographic information, tumor characteristics, and medical history details. The pre-processing of the dataset involves handling missing values and encoding categorical variables. The dataset consists of 286 instances, each characterized by 10 attributes. It is noted that there are missing values present within the dataset. As per the class distribution, 201 instances are labeled as 'no-recurrence-events,' while 85 instances are labeled as 'recurrence-events.' The data is divided into 80% for training and 20% for testing of the model.

3.2 Proposed Classification Model

This proposed model presents a detailed examination to classify breast tumor recurrence. This begins with dataset evaluation and resolving data quality issues via pre-processing. We utilized feature selection methods to pinpoint relevant attributes effectively. Weka classifiers, like MLP, Bayesian Networks, J48, AdaBoostM1, and LogitBoost, were employed. To ensure model robustness, we used both 10-fold and 5-fold cross-validation, along with testing different percentage splits. The model evaluation used metrics like precision, recall, F-measure, and ROC curve analysis, followed by parameter optimization in Weka to enhance performance. The refined model was then deployed for prediction on seen datasets. **Fig. 2** depicts the proposed model. The version 3.8.3 of WEKA software [36-38] is employed to generate results.
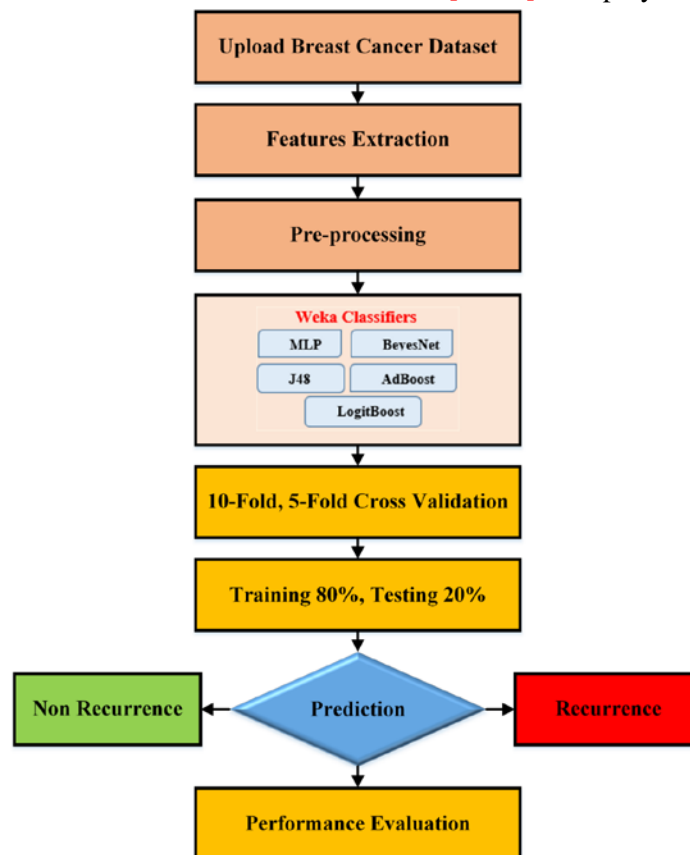

Fig. 2. Proposed model block diagram.

3.3 Performance Evaluation of the Model

The performance evaluation of the proposed classifiers was measured using the following metrics. These parameters were also used in the previous study [26].

$$TP - rate = \frac{TP}{(TP + FN)},$$
(1)

$$TN - rate = \frac{TN}{(TN + FP)},$$
(2)

$$FP - rate = \frac{FP}{(FP + TN)},$$
(3)

$$FN - rate = 1 - TP\ rate,$$
(4)

$$Accuracy = \left(\frac{correctly\ predicted\ class}{total\ testing\ class\ instance}\right) \times 100\%,$$
(5)

$$Precission = \frac{TP}{(TP + FP)},$$
(6)

$$Recall = \frac{TP}{(TP + FN)},$$
(7)

$$F - measure = \frac{2 \times Precission \times Recall}{Precision + Recall}.$$
(8)

In these equations, TP means true positive, TN means true negative, FP means false positive, and FN is false negative. The ROC Area, also known as the AUC, is a performance metric that assesses the accuracy of a binary classification algorithm. Two classes, "No recurrence" and "recurrence events," have been classified. In the context of this study, no recurrence means normal breast tumors. In contrast, recurrence events mean malignant breast tumors. **Fig. 3** depicts the confusion matrix for this analysis. In classifying breast cancer cases into two classes of non-recurrence and recurrence events, "True A" denotes the number of non-recurrence marked right cases, and "True B" gives the cases where recurrence events have been mistakenly marked as non-recurrence. False-A counts the class of non-recurrence instances that the machine mistakenly classified. Meanwhile, False-B counts the instances the machine classified to the class of non-recurrence, but that belonged to the class of recurrence.

Fig. 3. Proposed confusion matrix.

4. Results and Discussion

The following are the results of classifiers used for breast tumor classification.

4.1 Performance of MLP Classifier

Table 2 summarizes the MLP classifier, displaying a runtime of 0.89 seconds and utilizing 10 cross-validation folds. The count of instances is 286, where correctly classified instances are 185 and incorrectly classified instances are 101.

Table 2. Summary of MLP classifier.

Metrics	Values	
The correctly classified instances	185	64.69%
The incorrectly classified instances	101	35.31%
Kappa Statistic	0.1575	
MAE	0.3552	
RMSE	0.5423	
RAE	84.88%	
RRSE	118.65%	
The total number of instances	286	

Table 3 presents the results of a classification model's performance in distinguishing between "no-recurrence-events" and "recurrence-events." It comprehensively assesses the model's accuracy, precision, recall, and other key metrics. The model demonstrates reasonably good performance, with an overall weighted average accuracy of 0.647, indicating its ability to classify instances into these two classes correctly. Additionally, the MCC suggests moderate overall model quality. Collectively, these metrics show that the model has the potential for identifying instances related to breast cancer recurrence, providing valuable insights for medical decision-making and treatment strategies.

Table 3. Detailed accuracy of MLP classifier.

Metrics	no-recurrence-events	recurrence-events	Weighted Average
TP-rate	0.746	0.412	0.647
FP-rate	0.588	0.254	0.489
Precision	0.75	0.407	0.648
Recall	0.746	0.412	0.647
F-measure	0.748	0.409	0.647
MCC	0.158	0.158	0.158
ROC-Area	0.623	0.623	0.623
PRC-Area	0.79	0.41	0.677

Fig. 4 depicts the confusion matrix and events classification of the MLP classifier. It indicates that out of 150 instances of "no-recurrence-events," the model correctly classified 150 (TP), but it misclassified 51 as "recurrence-events" (FP). Similarly, out of 50 instances of "recurrence-events," the model correctly classified 35 (TP) but misclassified 50 as "no-recurrence-events" (FN).
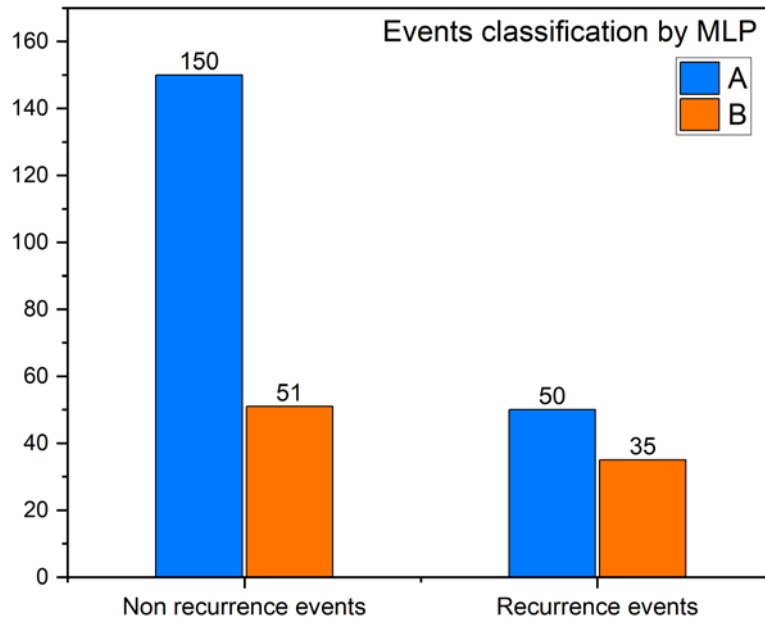
Fig. 4. Events Classification by MLP Classifier.

4.2 Performance of J48 (Decision Tree) Classifier

The DT classifier required 0.06 seconds for execution, and the cross-validation involved 10 folds. A summary of the J48 classifier can be found in **Table 4**. The total number of instances is 286, where correctly classified are 216 and incorrectly classified are 70.

Table 4. The summary of 48 (DT) classifier.

Metrics	Values	
The correctly classified instances	216	75.5245%
The incorrectly classified instances	70	24.48%
The Kappa statistic	0.2826	
MAE	0.3676	
RMSE	0.4324	
RAE	87.8635%	
RRSE	94.61%	
The total number of instances	286	

Table 5 presents the performance metrics for the J48 classifier and demonstrates the model's high accuracy in identifying "no-recurrence-events" but struggles with "recurrence-events." Finally, the model's quality is moderate, as indicated by the MCC of 0.339.

**Table 5.** J48 classifier detailed accuracy.

Metrics	no-recurrence-events	recurrence-events	Weighted Average
TP-rate	0.96	0.271	0.755
FP-rate	0.729	0.04	0.524
Precision	0.757	0.742	0.752

Recall	0.96	0.271	0.755
F-measure	0.846	0.397	0.713
MCC	0.339	0.339	0.339
ROC-Area	0.584	0.584	0.584
PRC-Area	0.736	0.436	0.647

Fig. 5 depicts the confusion matrix and events classification of the J48 classifier. The model accurately predicted 193 instances of "no-recurrence-events" and 23 instances of "recurrence-events." However, it made 8 FP predictions for "no-recurrence-events" and 62 FN predictions for "recurrence-events."
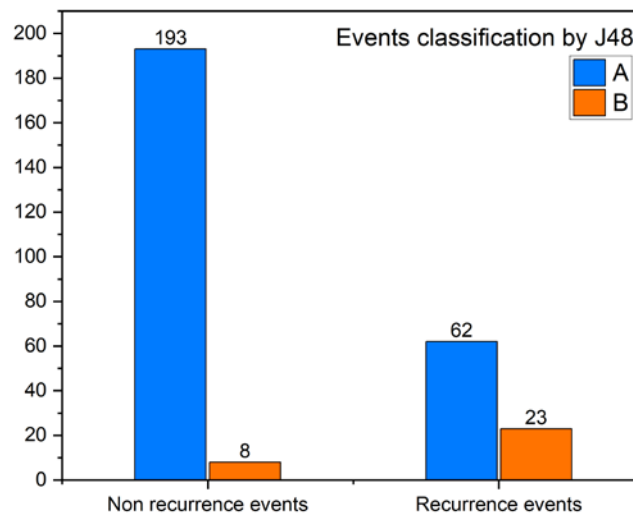
Fig. 5. Events classification by J48 classifier.

4.3 Performance of LogitBoost Classifier

The model was constructed with a 10-fold cross-validation with a building time of 0.03 seconds. Testing the model on the test split took 0 seconds. **Table 6** presents a comprehensive summary of the LogitBoot classifier, including various instances and errors.

Table 6. A Summary of LogitBoot Classifier.

Metrics	Values	
The correctly classified instances	207	72.38%
The incorrectly classified instances	79	27.62%
The Kappa statistic	0.2666	
MAE	0.3604	
RMSE	0.4409	
RAE	86.13%	
RRSE	96.46%	
The total number of instances	286	

Table 7 presents performance metrics for a LogitBoost classifier and shows strong performance in identifying "no-recurrence-events" with a good TP-rate and Precision. However, for "recurrence-events," the model's performance is comparatively weaker. Finally, all, the MCC suggests a moderate model quality.

Table 7. LogitBoost classifier detailed accuracy.

Metrics	no-recurrence-events	recurrence-events	Weighted Average
TP-rate	0.876	0.365	0.724
FP-rate	0.635	0.124	0.483
Precision	0.765	0.554	0.702
Recall	0.876	0.365	0.724
F-measure	0.817	0.44	0.705
MCC	0.277	0.277	0.277
ROC-Area	0.676	0.676	0.676
PRC-Area	0.816	0.475	0.715

Fig. 6 depicts the confusion matrix and events classification of a LogitBoost classifier. It shows that the model correctly predicted 176 instances of "no-recurrence-events" and 31 instances of "recurrence-events." However, it made 25 FP predictions for "no-recurrence-events" and 54 FN predictions for "recurrence-events."
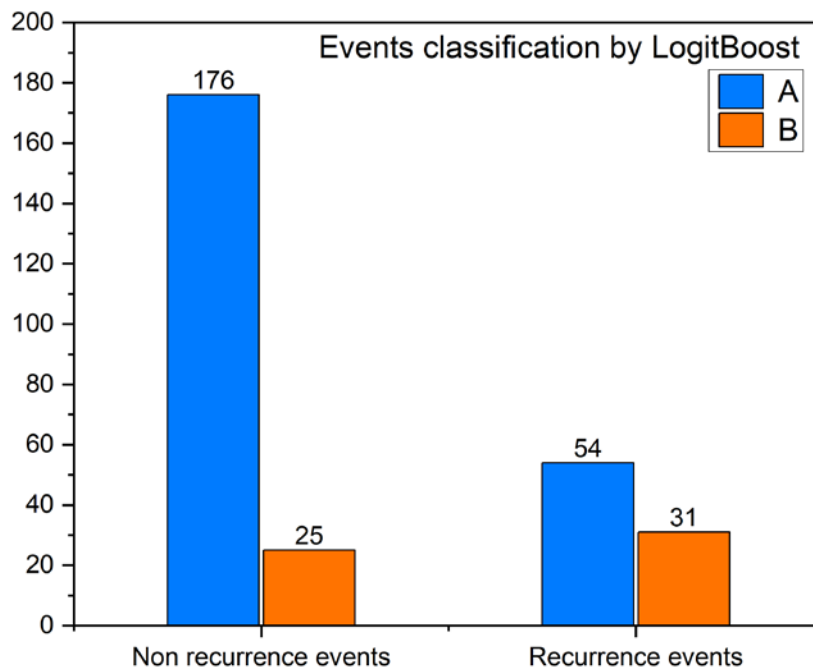
Fig. 6. Events classification by LogitBoost classifier.

4.4 Performance of AdaBoostM1 Classifier

The AdaBoostM1 classifier was executed in 0.02 seconds with 10 cross-validation folds. The testing of the model on the test split took 0 seconds. **Table 8** offers a comprehensive overview of the AdaBoostM1 classifier, encompassing various instances and associated errors.

Table 8. Summary of AdaBoostM1 classifier.

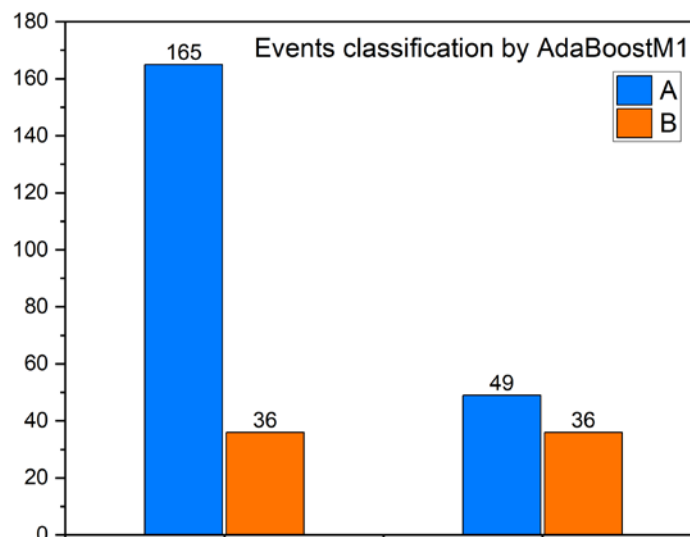| Metrics	Values	
The correctly classified instances	201	70.28%
The incorrectly classified instances	85	29.72%
The Kappa statistic	0.2557	
MAE	0.3526	
RMSE	0.4329	
RAE	84.27%	
RRSE	94.71%	
The total number of instances	286	

Table 9 presents performance metrics for the AdaBoostM1 classifier, and the model demonstrates moderate accuracy in both classes, as seen in TP-rate, Precision, and Recall. The MCC is 0.257, indicating moderate overall model quality.

Table 9. Accuracy of AdaBoostM1 in detail.

Metrics	no-recurrence-events	recurrence-events	Weighted Average
TP-rate	0.821	0.424	0.703
FP-rate	0.576	0.179	0.458
Precision	0.771	0.5	0.69
Recall	0.821	0.424	0.703
F-measure	0.795	0.459	0.695
MCC	0.257	0.257	0.257
ROC-Area	0.697	0.697	0.697
PRC-Area	0.833	0.494	0.732

Fig. 7 depicts the confusion matrix and events classification of the AdaBoostM1 classifier. The model correctly predicted 165 instances of "no-recurrence-events" and 36 instances of "recurrence-events." However, it made 36 FP predictions for "no-recurrence-events" and 49 FN predictions for "recurrence-events.

Fig. 7. Events classification of AdaBoostM1 classifier.

4.5 Performance of BayesNet Classifier

Table 10 presents the summary of the BayesNet classifier, including the time taken to build the model (0.04 seconds) and the number of cross-validation folds (10).
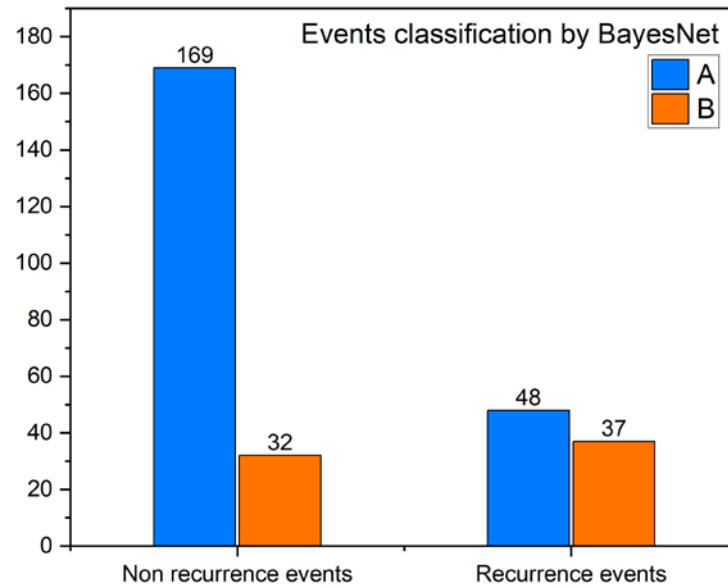
Table 10. BayesNet classifier summary.

Metrics	Values	
The correctly classified instances	206	72.03%
The incorrectly classified instances	80	27.97%
The Kappa statistic	0.2919	
MAE	0.3297	
RMSE	0.4566	
RAE	78.79%	
RRSE	99.90%	
The total number of instances	286	

Table 11 presents the performance metrics for the BayesNet classifier. Also, it indicates a moderate ability of the model to correctly classify instances in both classes, as shown by metrics like TP-rate, Precision, and Recall. The MCC of 0.295 suggests moderate overall model quality.

Table 11. Detailed accuracy assessment for the BayesNet classifier.

Metrics	no-recurrence-events	recurrence-events	Weighted Average
TP-rate	0.841	0.435	0.72
FP-rate	0.565	0.159	0.444
Precision	0.779	0.536	0.707
Recall	0.841	0.435	0.72
F-measure	0.809	0.481	0.711
MCC	0.295	0.295	0.295
ROC-Area	0.698	0.698	0.698
PRC-Area	0.833	0.51	0.737

Fig. 8 depicts the confusion matrix and events classification of the BayesNet classifier. It correctly predicted 169 instances of "no-recurrence-events" and 37 instances of "recurrence-events." However, it made 32 FP predictions for "no-recurrence-events" and 48 FN predictions for "recurrence-events."

Fig. 8. Events classification by BayesNet classifier.

4.6 Comparison of the Proposed Models

This study implemented various classifiers like MLP, J48, LogitBoost, AdBoostM1, and BeyesNet for pattern recognition. **Table 12** presents the performance results for various classifiers based on their F-measure and Accuracy scores. It demonstrates that the J48 (DT) classifier achieves the highest F-measure and accuracy at 0.713 and 71.3%, respectively, indicating its superior performance in classifying instances.

Table 12. A comparative analysis proposed classifiers.

Classifier	F-measure	Accuracy
MLP	0.647	64.7%
J48	0.713	71.3%
LogitBoost	0.705	70.5%
AdBoostM1	0.695	69.5%
BeyesNet	0.711	71.1%

4.7 Further Evaluation of the J48 Classifier

The J48 classifier is further evaluated regarding validation fold and percentage splitting to find better results. Regarding 5-fold validation, **Table 13** presents a breakdown of the accuracy metrics for the J48 classifier, considering various parameters and their respective classes.

Table 13. J48 classifier accuracy in detail.

Metrics	no-recurrence-events	recurrence-events	Weighted Average
TP-rate	0.96	0.224	0.741
FP-rate	0.776	0.04	0.558
Precision	0.745	0.704	0.733

Recall	0.96	0.224	0.741
F-measure	0.839	0.339	0.691
MCC	0.287	0.287	0.287
ROC-Area	0.582	0.582	0.582
PRC-Area	0.728	0.444	0.643

The accuracy of 10 fold is 71.3, whereas the accuracy of 5 folds is 69%, which is reduced. So, 10 folds of J48 are further evaluated with different percentage splits, as presented in **Table 14**.

Table 14. Evaluation of 10-fold J48 Classifier with different percentage splits.

Split (%)	Metrics	no-recurrence-events	recurrence-events	Weighted Average
50	TP-rate	0.938	0.191	0.692
	FP-rate	0.809	0.063	0.563
	Precision	0.703	0.6	0.669
	Recall	0.938	0.191	0.692
	F-measure	0.804	0.29	0.635
	MCC	0.198	0.198	0.198
	ROC-Area	0.656	0.656	0.656
	PRC-Area	0.754	0.466	0.66
90	TP-rate	0.895	0.2	0.655
	FP-rate	0.8	0.105	0.56
	Precision	0.68	0.5	0.618
	Recall	0.895	0.2	0.655
	F-measure	0.773	0.286	0.605
	MCC	0.131	0.131	0.131
	ROC-Area	0.626	0.626	0.626
	PRC-Area	0.734	0.419	0.625
35	TP-rate	0.89	0.373	0.726
	FP-rate	0.627	0.11	0.463
	Precision	0.753	0.611	0.708
	Recall	0.89	0.373	0.726
	F-measure	0.816	0.463	0.704
	MCC	0.309	0.309	0.309
	ROC-Area	0.637	0.637	0.637
	PRC-Area	0.761	0.429	0.656
73	TP-rate	0.959	0.214	0.688
	FP-rate	0.786	0.041	0.515

	Precision	0.681	0.75	0.706
	Recall	0.959	0.214	0.688
	F-measure	0.797	0.333	0.628
	MCC	0.273	0.273	0.273
	ROC-Area	0.577	0.577	0.577
	PRC-Area	0.678	0.48	0.606
40	TP-rate	0.872	0.327	0.698
	FP-rate	0.673	0.128	0.499
	Precision	0.734	0.545	0.674
	Recall	0.872	0.327	0.698
	F-measure	0.797	0.409	0.673
	MCC	0.236	0.236	0.236
	ROC-Area	0.646	0.646	0.646
	PRC-Area	0.759	0.416	0.649

Table 14 contains detailed accuracy and performance metrics of J48 split percentages 50, 90, 35, 73, and 40. As per split 50, the model excels in identifying "no-recurrence-events" with a high TP-rate, but it faces challenges in classifying "recurrence-events." Regarding the J48 split percentage 90, the model highlights a higher TP-rate for "no-recurrence-events" but challenges in classifying "recurrence-events." The Precision values are somewhat balanced. The MCC of 0.131 indicates a moderate overall model quality.

Regarding the J48 split percentage, the model achieves a relatively high TP-rate for "no-recurrence-events" but faces challenges with "recurrence-events." The Precision values show a reasonable balance between the classes. The MCC of 0.309 suggests moderate overall model quality. In evaluating J48 split percentage 73, the model identifies "no-recurrence-events" with a high TP-rate but struggles with "recurrence-events." The Precision values show a reasonable balance between the classes. The MCC of 0.273 indicates moderate overall model quality. Finally, the model is evaluated with J48 split percentage 40 and observed that the model is relatively proficient at identifying "no-recurrence-events" with a high TP-rate but faces challenges with "recurrence-events." The Precision values suggest a reasonable balance between the classes. The MCC of 0.236 indicates moderate overall model quality. These findings evaluate the model's ability to distinguish between the two classes, with room for improvement in some areas.

Table 15 compares different split percentages on the accuracy of a machine-learning model. It demonstrates that a split percentage of 66% yields the highest accuracy at 71%, indicating that this particular data split ratio is most effective for this model. Other split percentages result in varying levels of accuracy, suggesting the importance of selecting an appropriate data split strategy for optimal model performance.

Table 15. Accuracy analysis of J48 on different split percentages.

Split (%)	Accuracy
50	63.5%
90	60.5%
35	70.4%

66	71%
73	62.8%
40	67.3%

5. Discussion

This study implanted different classifiers, but the maximum accuracy (71%) is achieved using J48 (DT). Weka was used to implement the classifiers, and then different classifiers were tried. The results of these classifiers are presented, but the results of some classifiers are not presented if their accuracy is less than 69%. We mentioned some accuracies, but some are ignored, e.g., as the summary of Naïve Bayes classier is presented in **Table 16**. The detailed accuracy by class is presented in **Table 17**.

Table 16. Summary of Naïve Bayes classier.

Metrics	Values	
The correctly classified instances	19	65.5172%
The incorrectly classified instances	10	34.4828%
The Kappa Statistic	0.157	
MAE	0.3895	
RMSE	0.5399	
RAE	89.3597%	
RRSE	112.9399%	
The total number of instances	29	

Table 17. Detailed accuracy of Naïve Bayes.

Metrics	no-recurrence-events	recurrence-events	Weighted Average
TP-rate	0.842	0.300	0.655
FP-rate	0.700	0.158	0.513
Precision	0.696	0.500	0.628
Recall	0.842	0.300	0.655
F-measure	0.762	0.375	0.628
MCC	0.167	0.176	0.167
ROC-Area	0.518	0.518	0.518
PRC-Area	0.727	0.467	0.637

Weka provides the F-measure and ROC curves to analyze the accuracy of the model. The weighted F-measure of J48 is 0.713, and the ROC is 0.58, which indicates the performance of J48 on a given dataset. The accuracies of different classifiers are evaluated again J48 provided the maximum F-measure weighted average value of 71.3 compared to other classifiers.

6. Conclusion

This comprehensive study uses many machine learning classifiers to classify the occurrence or non-occurrence of breast tumors based on various features and data points related to individuals' medical history. The study evaluated several classification methods, including MLP, AdaBoostM1, logitBoost, BayesNet, and J48. The effectiveness of these classifiers is assessed using various performance metrics. The findings indicated that the J48 (DT) classifier outperformed the other classifiers, demonstrating the highest accuracy among the tested methods. With an accuracy of 71%, J48 demonstrated its effectiveness in accurately classifying instances into the appropriate class. It was also found that a split percentage of 66% provided the optimal balance for achieving the highest accuracy.

Furthermore, the impact of fold values on the model's accuracy is explored by modifying the fold value from 10 to 5. However, the results indicated that the 10-fold cross-validation produced the best accuracy results. This research highlights the potential of employing pattern recognition and DT-based classifiers, particularly J48, in accurately classifying cancer-related instances. These findings offer valuable insights for developing cancer assessment models and significantly contribute to the field of computational biology.

7. Limitations and Future Work

In future studies, the following limitations should be considered:
- The class distribution is imbalanced, which may lead to biased model performance towards the majority class.
- The presence of missing values in the dataset can affect the accuracy of the classifiers if not properly handled.
- Limited feature selection may result in models not capturing all relevant patterns in the data.
- The models trained on this dataset may not generalize well to other datasets or populations.
- Using a publicly available dataset instead of real-time clinical data may limit the generalizability of the classification results to real-world scenarios. It may not capture the full range of variability and complexities present in clinical settings.
- These models are suitable for smaller datasets with fewer features. However, deep learning (DL) models better classify complex patterns and high-dimensional data.

Funding

Competing Interest

The authors declare there are no competing interests.

References

[1] K. L. Britt, J. Cuzick, and K.-A. Phillips, "Key steps for effective breast cancer prevention," *Nature Reviews Cancer*, vol. 20, no. 8, pp. 417-436, 2020. Article (CrossRef Link).

[2] N. Bilani, E. C. Zabor, L. Elson, E. B. Elimimian, and Z. Nahleh, "Breast cancer in the United States: a cross-sectional overview," *Journal of cancer epidemiology*, vol. 2020, 2020. Article (CrossRef Link).

[3] R. M. Mann, R. Hooley, R. G. Barr, and L. Moy, "Novel approaches to screening for breast cancer," *Radiology*, vol. 297, no. 2, pp. 266-285, 2020. Article (CrossRef Link).

[4] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast tumor classification using an ensemble machine learning method," *Journal of Imaging*, vol. 6, no. 6, p. 39, 2020. Article (CrossRef Link).

[5] G. Murtaza et al., "Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges," *Artificial Intelligence Review*, vol. 53, pp. 1655-1720, 2020. Article (CrossRef Link).

[6] X.-X. Yin, L. Yin, and S. Hadjiloucas, "Pattern classification approaches for breast cancer identification via MRI: state-of-the-art and vision for the future," *Applied Sciences*, vol. 10, no. 20, p. 7201, 2020. Article (CrossRef Link).

[7] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi, "Medical image based breast cancer diagnosis: State of the art and future directions," *Expert Systems with Applications*, vol. 167, p. 114095, 2021. Article (CrossRef Link).

[8] A. Kalantari, A. Kamsin, S. Shamshirband, A. Gani, H. Alinejad-Rokny, and A. T. Chronopoulos, "Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions," *Neurocomputing*, vol. 276, pp. 2-22, 2018. Article (CrossRef Link).

[9] A. A. Abdul Halim et al., "Existing and emerging breast cancer detection technologies and its challenges: a review," *Applied Sciences*, vol. 11, no. 22, p. 10753, 2021. Article (CrossRef Link).

[10] A. N. Cobb, H. M. Janjua, and P. C. Kuo, "Big data solutions for controversies in breast cancer treatment," *Clinical breast cancer*, vol. 21, no. 3, pp. e199-e203, 2021. Article (CrossRef Link).

[11] C. Chakraborty, S. Barbosa, and L. Garg, "Preface to Special Issue on Scientific Computing and Learning Analytics for Smart Healthcare Systems (Part I)," *Computer Assisted Methods in Engineering and Science*, vol. 30, no. 2, pp. 107-109, 2023. Article (CrossRef Link).

[12] R. Kumar et al., "An integration of blockchain and AI for secure data sharing and detection of CT images for the hospitals," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101812, 2021. Article (CrossRef Link).

[13] A. Su et al., "A deep learning model for molecular label transfer that enables cancer cell identification from histopathology images," *NPJ precision oncology*, vol. 6, no. 1, p. 14, 2022. Article (CrossRef Link).

[14] C. H. Barrios, "Global challenges in breast cancer detection and treatment," *The Breast*, vol. 62, pp. S3-S6, 2022. Article (CrossRef Link).

[15] F. Zhu, J. Gao, J. Yang, and N. Ye, "Neighborhood linear discriminant analysis," *Pattern Recognition*, vol. 123, p. 108422, 2022. Article (CrossRef Link).

[16] F. Teixeira, J. L. Z. Montenegro, C. A. da Costa, and R. da Rosa Righi, "An analysis of machine learning classifiers in breast cancer diagnosis," in *Proc. of 2019 XLV Latin American computing conference (CLEI)*, pp. 1-10, 2019. Article (CrossRef Link).

[17] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of breast cancer detection using different machine learning techniques," in *Proc. of Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia*, pp. 108-117, 2020. Article (CrossRef Link).

[18] M. Sant, A. Bernat-Peguera, E. Felip, and M. Margelí, "Role of ctDNA in breast cancer," *Cancers*, vol. 14, no. 2, p. 310, 2022. Article (CrossRef Link).

[19] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast cancer prediction: a comparative study using machine learning techniques," *SN Computer Science*, vol. 1, pp. 1-14, 2020. Article (CrossRef Link).

[20] A. A. Farid, G. Selim, and H. Khater, "A Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm For Breast Cancer Detection," in *Proc. of The 2nd International Conference on Advanced Research in Applied Science and Engineering*, 2020. Article (CrossRef Link).

[21] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani, and M. Faisal Nagi, "Automated breast cancer diagnosis based on machine learning algorithms," *Journal of healthcare engineering*, vol. 2019, 2019. Article (CrossRef Link).

[22] J. P. Sarkar, I. Saha, A. Sarkar, and U. Maulik, "Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers," *Computers in Biology and Medicine*, vol. 131, p. 104244, 2021. Article (CrossRef Link).

[23] Y. S. Solanki et al., "A hybrid supervised machine learning classifier system for breast cancer prognosis using feature selection and data imbalance handling approaches," *Electronics*, vol. 10, no. 6, p. 699, 2021. Article (CrossRef Link).

[24] M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, and W. Zhou, "Breast cancer detection in the IOT health environment using modified recursive feature selection," *wireless communications and mobile computing*, vol. 2019, pp. 1-19, 2019. Article (CrossRef Link).

[25] S. Laghmati, B. Cherradi, A. Tmiri, O. Daanouni, and S. Hamida, "Classification of patients with breast cancer using neighbourhood component analysis and supervised machine learning techniques," in *Proc. of 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*, pp. 1-6, 2020. Article (CrossRef Link).

[26] I. Haq et al., "Machine Vision Approach for Diagnosing Tuberculosis (TB) Based on Computerized Tomography (CT) Scan Images," *Symmetry*, vol. 14, no. 10, p. 1997, 2022. Article (CrossRef Link).

[27] I. Haq, N. Ullah, T. Mazhar, M. A. Malik, and I. Bano, "A Novel Brain Tumor Detection and Coloring Technique from 2D MRI Images," *Applied Sciences*, vol. 12, no. 11, p. 5744, 2022. Article (CrossRef Link).

[28] B. S. Abunasser, M. R. J. AL-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, "Breast cancer detection and classification using deep learning Xception algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022. Article (CrossRef Link).

[29] F. Ullah et al., "Brain Tumor Segmentation from MRI Images Using Handcrafted Convolutional Neural Network," *Diagnostics*, vol. 13, no. 16, p. 2650, 2023. Article (CrossRef Link).

[30] P. Karatza, K. Dalakleidi, M. Athanasiou, and K. S. Nikita, "Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis," in *Proc. of 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2310-2313, 2021. Article (CrossRef Link).

[31] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and E. Alba-Conejo, "A combined neural network and decision trees model for prognosis of breast cancer relapse," *Artificial intelligence in medicine*, vol. 27, no. 1, pp. 45-63, 2003. Article (CrossRef Link).

[32] C.-Y. Fan, P.-C. Chang, J.-J. Lin, and J. Hsieh, "A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification," *Applied Soft Computing*, vol. 11, no. 1, pp. 632-644, 2011. Article (CrossRef Link).

[33] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers-a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476-487, 2005. Article (CrossRef Link).

[34] github. Weka datasets/breast-cancer.arff [Online]. Available: https://github.com/tertiarycourses/Weka/blob/master/Weka%20datasets/breast-cancer.arff. https://doi.org/10.24432/C51P4M.

[35] M. Z. M. Soklic. datasets/breast-cancer [Online]. Available: https://github.com/datasets/breast-cancer/blob/master/README.md. https://doi.org/10.24432/C51P4M.

[36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10-18, 2009. Article (CrossRef Link).

[37] N. R. Pal and L. Jain, *Advanced techniques in data mining and knowledge discovery*, Springer, 2005. Article (CrossRef Link).

[38] S. Singhal and M. Jena, "A study on WEKA tool for data pre-processing, classification and clustering," *International Journal of Innovative technology and exploring engineering (IJItee)*, vol. 2, no. 6, pp. 250-253, 2013. Article (CrossRef Link).

Inayatul Haq is a PhD Scholar in the Department of Information and Communication Engineering at Zhengzhou University, China. He received a BS degree in telecommunication engineering from BUITMS Quetta, Pakistan, and an MS degree in telecommunication engineering from Hamdard University UIT Campus, Karachi, Pakistan. Also, he has obtained another master's (MS) degree in project management from VCOMSATS Islamabad, Pakistan. His research interests include cancer cell detection through AI techniques, medical imaging processing, deep learning, machine learning, and telecommunications. He has published more than 14 research papers in SCI and good-ranking journals. Also, he has 9 years of job experience in Pakistan's telecommunication sector in the radio frequency department.

Tehseen Mazhar received a B.Sc. degree in computer science from Bahaudin Zakaria University, Multan, Pakistan, an M.Sc. degree in computer science from Qauid-e-Azam University Islamabad, Pakistan, and an MSCS degree from the Virtual University of Pakistan, where he is currently pursuing the Ph.D. degree. He is also with SED and a Lecturer with GCUF. He has more than 21 publications in reputed journals. His research interests include machine learning, the Internet of Things, and computer networks.

Hinna Hafeez is a lecturer at the Department of Computer Science Superior University Lahore. She is PhD scholar at the Department of CS in the National College of Business Administration And Economics Lahore.

Dr Najib Ullah obtained a BS degree in Pharmacy from the University of Balochistan, Pakistan, in 2016. He is a registered member of the Pakistan Pharmacy Council. He has been working as a pharmacist in the pharmaceutical industry of Pakistan since 2016 to date. His research interest is early cancer diagnosis and drug delivery.

Fatma Mallek was born on November 19th, 1987, in Sfax, Tunisia. She received her Bachelor of Science (B.Sc) in the field of computer-applied sciences from the University of Sfax, Tunisia, in 2010. She completed her Master of Science (Msc), in the field of computer sciences from the University of Québec À Montréal (UQÀM, Canada) in 2017. She joined Université de Moncton, New-Brunswik, to continue her study for a Doctor of Philosophy (PhD) in the field of applied sciences. Since 2020, she has been a principal lecturer in Big Data and Artificial intelligence programs at Institut Élite de Montréal. Her research interests lie in information processing, big data, IoT, artificial intelligence, and deep learning.

Prof. Dr. Habib Hamam obtained the B.Eng. and M.Sc. degrees in information processing from the Technical University of Munich, Germany 1988 and 1992, and the Ph.D. degree in Physics and applications in telecommunications from Université de Rennes I conjointly with France Telecom Graduate School, France 1995. He also obtained a postdoctoral diploma, "Accreditation to Supervise Research in Signal Processing and Telecommunications", from Université de Rennes I in 2004. He was a Canada Research Chair holder in "Optics in Information and Communication Technologies," the most prestigious research position in Canada – which he held for a decade (2006-2016). The title is awarded by the Head of the Government of Canada after a selection by an international scientific jury in the related field. He is currently a full Professor in the Department of Electrical Engineering at Université de Moncton. He is an OSA senior member, an IEEE senior member, and a registered professional engineer in New Brunswick. He obtained several pedagogical and scientific awards. He is, among others, editor-in-chief and founder of CIT-Review, academic editor in Applied Sciences, and associate editor of the IEEE Canadian Review. He also served as a Guest editor in several journals. His research interests are in optical telecommunications, Wireless Communications, diffraction, fiber components, RFID, information processing, IoT, data protection, COVID-19, and Deep learning.