Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset

Md. Murad Hossin[1], F.M. Javed Mehedi Shamrat[2], Md Rifat Bhuiyan[1], Rabea Akter Hira[3], Tamim Khan[1], Shourav Molla[1]

[1]Department of Computer Science and Engineering, Faculty of Science and Information Technology, Daffodil International University, Dhaka, Bangladesh
[2]Department of Software Engineering, Faculty of Science and Information Technology, Daffodil International University, Dhaka, Bangladesh
[3]Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh

Article Info

Article history:

Received Jul 20, 2022
Revised Dec 16, 2022
Accepted Feb 3, 2023

Keywords:

Breast cancer detection
Breast cancer Wisconsin
Explainable AI
Feature selection
Machine learning algorithms

ABSTRACT

According to the American cancer society, breast cancer is one of the leading causes of women's mortality worldwide. Early identification and treatment are the most effective approaches to halt the spread of this cancer. The objective of this article is to give a comparison of eight machine learning algorithms, including logistic regression (LR), random forest (RF), K-nearest neighbors (KNN), decision tree (DT), ada boost (AB), support vector machine (SVM), gradient boosting (GB), and Gaussian Naive Bayes (GNB) for breast cancer detection. The breast cancer Wisconsin (diagnostic) dataset is being utilized to validate the findings of this study. The comparison was made using the following performance metrics: accuracy, sensitivity, false omission rate, specificity, false discovery rate and area under curve. The LR method achieved a maximum accuracy of 99.12% among all eight algorithms and was compared to other comparable studies in the literature. The five features chosen are used to calculate the model's fidelity-to-interpretability ratio (FIR), which indicates how much interpretability was sacrificed for performance. The uniqueness of this work is the explainability approach taken in the model's performance, which aims to make the model's outputs more understandable and interpretable to healthcare experts.

Corresponding Author:

F.M. Javed Mehedi Shamrat
Department of Software Engineering, Faculty of Science and Information Technology
Daffodil International University
Daffodil Road, Asulia, Dhaka 1341, Bangladesh
Email: javedmehedicom@gmail.com

1. INTRODUCTION

Cancer is the most dangerous disease in the world, especially breast cancer being the most dangerous for women. Breast cancer claims the lives of many women every year. The International Agency for Research on Cancer (IARC) reported in December 2020 that breast cancer has overtaken lung cancer among the most common chronic cancer in women worldwide. The number of cancer cases doubled from 10 million in 2000 to 19.3 million in the past two decades [1]. One in five people in the modern world will eventually get cancer. Breast cancer is difficult to diagnose and takes a long time to identify manually. Consequently, it is essential to diagnose cancer utilizing a variety of automated diagnostic techniques. Logistic regression (LR), decision tree (DT), random forest (RF), K-nearest neighbors (KNN), support vector machine (SVM), ada boost (AB), gradient boost (GB), Gaussian Naive Bayes (GNB) are some of the methods and algorithms available for identifying breast cancer.

In this study, training and testing are conducted using the UCI open database, which contains both benign and malignant tumor types. Malignant tumors are cancerous, while benign tumors are non-cancerous. Many researchers are still working on finding ways to detect and diagnose cancer at an early stage. Because early-stage cancer is less expensive and easier to treat, many researchers are still working on developing a proper diagnosis method. As a result, treatment can begin sooner, and the resolution rate may increase. The primary objective of this study is to evaluate several machine learning (ML) algorithms and identify the most efficient method for breast cancer detection.

Various modern strategies for breast cancer prediction have grown with the advancement of technology. The following is a summary of the work done in this field: Bazazeh and Shubair [2] used SVM, RF, and bayesian networks to diagnose breast cancer using the Wisconsin breast cancer dataset (WBCD). Among the performance, measures are accuracy, recall, and precision. The final result was 97% accurate. Breast cancer screening strategies using SVM and KNN are proposed in [3]. They are calculated as a performance matrix. The final results indicated that SVM and K-NN had 98.57% and 97.14% accuracy, respectively. According to Naji *et al.* [4], SVM, RF, LR, DT, and KNN are used to detect breast cancer. This machine's accuracy was 97.2%. Sharma *et al.* [5] used SVM, RF, and NB to detect breast cancer using the WBCD. RF, KNN, and NB accuracy was 94.7%, 95.9%, and 94.47%, respectively. Research by Sengar *et al.* [6] present breast cancer detection technique using LR and DT. The study found that LR and DT were 94.40% and 95.10% accurate, respectively. Agarap [7] advise employing gated recurrent unit (GRU)-SVM, LR, multi-layer perceptron (MLP), L1-NN, L2-NN, SoftMax regression, and SVM. Among these, MLP had the highest accuracy of 99.03% on the breast cancer Wisconsin (diagnostic) dataset.

2. PROCESS FLOW DIAGRAM

The study's principal objective is to identify the best accurate and predicting algorithm for diagnosing breast cancer. The suggested structure is depicted in detail in Figure 1. The work begins with data acquisition followed by pre-processing, which includes four steps: data cleaning, attribute selection, target role selection, and feature extraction. ML algorithms are constructed utilizing the prepared data in order to identify breast cancer depending on a new set of measurements. For evaluating the algorithms, the model gets labelled data. This is often accomplished by using the train_test_split method to split the labelled data that are gathered into two pieces. The training set, sometimes referred to as the training data, contains 80% of the data used to build our ML algorithm. The 20% of the data used to assess the model's performance is known as the test data or test set. In order to decide which algorithm is the most accurate, the results are compared. Univariate feature selection (UFS) and recursive feature elimination (RFE) are used to compare the prior models accuracy. Finally, the best algorithm for detecting breast cancer has been identified.

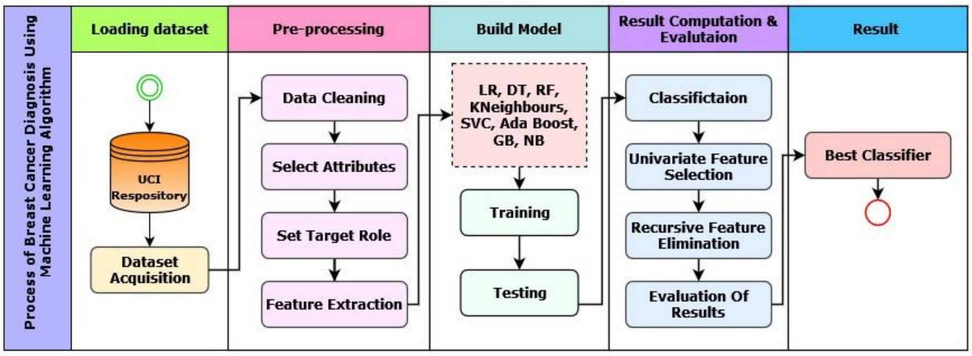

Figure 1. The proposed architecture of the breast cancer detection

3. METHOD

3.1. Machine learning library

The Scikitlearn library and the Python programming language were used to conduct all of the experiments on the ML techniques discussed in this study. Scikit-learn, commonly referred to as sklearn, is a free Python package for ML [8]. With the support of various scientific computing libraries such as Scikit-learn [9], NumPy [10], matplotlib [11], pandas [12], and seaborn [13], [14], are employed in this study.

3.2. Dataset

The WBCD [15] is used in this study. This data was obtained by Dr. William H. Wolberg, a physician at University of Wisconsin Hospital in Madison, Wisconsin, United States of America. To create the dataset, Wolberg obtained fluid samples from patients who had solid breast masses and used Xcyt, a user-friendly graphical computer application that can analyze cytological traits based on a digital scan. Using a curve-fitting method, the computer program computes ten features for each cell in the sample, then calculates the i) mean value, ii) standard error, and iii) extreme value of each feature for the image, returning a 30 real-valued vector. Each characteristic is ranked from 1 to 10, with 1 being the most benign and 10 denoting the most malignant. The ten features are as follows: i) radius: mean of distances from center to points on the perimeter, ii) texture: standard deviation of gray-scale values, iii) perimeter, iv) area, v) smoothness: local variation in radius lengths, vi) compactness: perimeter 2/area-1.0, vii) concavity: severity of concave portions of the contour, viii) concave points: number of concave portions of the contour, ix) symmetry, and x) fractal dimension: “coastline approximation”-1. The dataset has 569 data items (rows) and 33 characteristics (features), with 357 (62.7%) benign (non-cancerous) and 212 (37.3%) malignant (cancerous) features (cancerous).

3.3. Data pre-processing

The original dataset comprises 33 features, but the first (id) and last (unnamed) columns add nothing to our outcome (whether is benign or malignant?), so they are eliminated from the dataset. A character feature of the dataset, "diagnostic," was manually turned into a numeric feature using a python dictionary where M=1 and B=0. Finally, the dataset was standardized using the following equation to avoid improper relevance assignment:

$$z = (x - u) / s$$

The standardization feature's mean is given by u, and the standard deviation by s. Standardization uses StandardScalar().fit transform() for training and StandardScalar().transform() for testing. Standardization is conducted after splitting the train-test data to eliminate data leaks.

3.4. Features selection process

Feature selection is a significant approach for improving any classifier's performance. Feature selection has the potential to reduce the overall training and prediction time. During this step, just the most crucial characteristics from the original datasets were selected. Then, these features were examined for training and testing. These techniques have a substantial effect on classification results.

3.4.1. Univariate feature selection

UFS grabs the best features employing univariate statistical tests. Each attribute is compared to the dependent variable to see if a statistically significant relationship exists between them. It is sometimes referred to as an analysis of variance (ANOVA). While examining the connection among one feature and the predictor variables, all other aspects are ignored. This is why it is called univariate [16]. There is a test score for each feature. Finally, the results of all the tests are compared and the best five selected features are: ['perimeter_mean', 'area_mean', 'area_se', 'perimeter_worst', 'area_worst'].

3.4.2. Recursive feature elimination

RFE, sometimes referred to as the feature selection algorithm, chooses features by recursively taking into account progressively smaller sets of features while an external estimator provides feature weights. The significance of each feature is assessed utilizing coef_ or feature importances_ attributes once the estimator has been trained on the initial set of features. The least important features are then removed considering the current set of features. The trimmed set is subjected to this method repeatedly up until the required set of attributes to choose is attained [17]. A training dataset's strongest predictive features are identified quickly using RFE. For LR, DT, RF, KNN, SVM, AB, GB, and GNB, the optimal number of features is 19, 4, 20, 15, 14, 26, 22, and 17, respectively.

3.4.3. Correlation heatmap

Correlation is used to determine how closely two or more variables fluctuate. The stronger the relationship between a dependent variable and an independent variable, the more important that independent variable is in determining the dependent variable. The correlation value can be positive, negative, or zero based on the direction of change. A strong correlation between dependent and independent variables is beneficial, but a strong correlation between two independent variables is undesirable [18]. Variables are correlated in Figure 2. Highly correlated pairings of independent variables are deemed redundant, resulting in a waste of time and space.

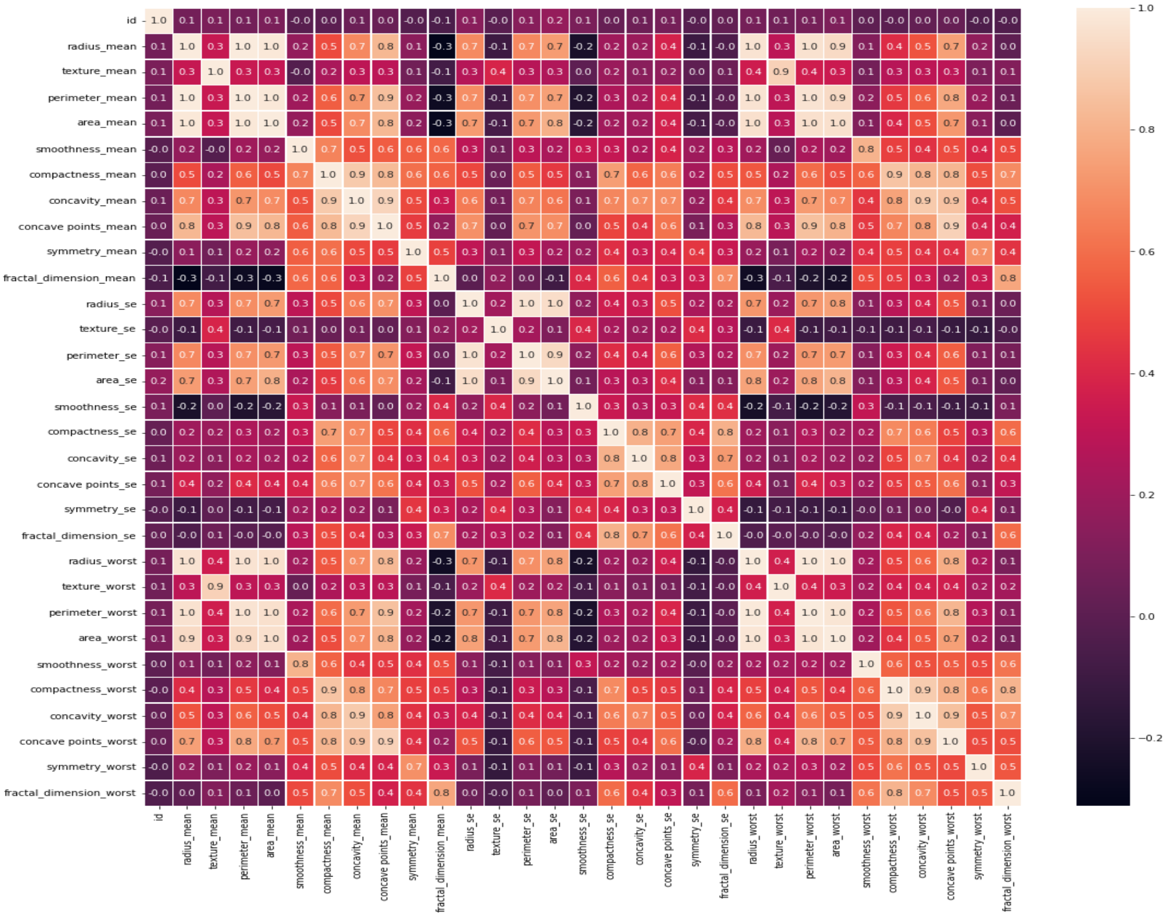

Figure 2. Heat map representing the correlation between all dependent and independent variable

3.5. Fold cross-validation

K-Folds produces a less biased model than other approaches and is suitable for small data sets. The data is randomly folded (the amount of K should not be too small or too large; ideally, 5 to 10 folds, depending on the data size, should be chosen). The model becomes less biased as K rises (although large variation may lead to over-fitting). The 5-fold cross-validation prevents overfitting and variance [19].

3.6. Machine learning algorithms

The study compares ML breast cancer detection approaches. Many algorithms are employed. It compares algorithms with and without feature selection.

3.6.1. Logistic regression

The classification algorithm LR is a supervised learning technique. A dichotomous target (dependent) variable's probability is predicted using this approach. It returns probabilistic values between 0 and 1 as a result of the calculation. The algorithm does the classification based on these probabilistic values. Where the method includes a sigmoid function (0=output>=1) linear regression and LR are quite similar. The hypothesis function is the only difference, on the other hand, it employs a regression function(-∞<=output>=∞). As a result, they have diverse applications, with the former being utilized for classification and the latter for regression. It is used in data analysis to comprehend dependent and independent variables [20].

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \tag{2}$$

Where P denotes the predicted output, a the biased or intercept term, and b the coefficient for a particular input value (X).

3.6.2. Decision tree classifier

Another supervised ML approach for classification is DT classifier. This algorithm generates a DT, a tree structure for classifying distinct classes, in which internal (decision) nodes reflect features of a dataset used to make any decision, branches indicate decision rules, and each leaf node represents the outcome (class). A DT basically asks a question and splits the tree into sub-trees based on the response (yes/no). Making decisions helps. DT are great for comparing options. They provide a robust framework for evaluating options. Entropy adequately refers to the volume of information necessary to describe a sample. So, if the sample is homogeneous, which means that all of the elements are comparable, the entropy is 0; otherwise, if the sample is evenly divided, the entropy is maximum 1 [21], [22].

$$Entropy = -\sum_{i=1}^{n} p_i \times log(p_i) \tag{3}$$

The Gini index measures sample inequality statistically. It is indeed a number scale of 0 to 1. A Gini value of 0 denotes that the sample is fully uniform and almost all of the elements are comparable, whereas a Gini index of 1 denotes that the population is very dissimilar. It's the square root of each class's probabilities.

$$Gini\ index = 1 - \sum_{i=1}^{n} {p_i}^2 \tag{4}$$

3.6.3. K-nearest neighbor

KNN is a supervised ML classification algorithm. This technique saves all of the training data and categorizes new data points based on how similar the nearest (with the shortest Euclidean distance) k training data points are. It's a non-parametric method, which means it doesn't make any predictions regarding the data. The algorithm does not instantly start learning from the training data provided; instead, it rests on the training data until new data for classification is provided. As a result, it's also known as a sluggish learning algorithm. The training data assumptions are nonparametric. It handles nonlinear data. Because KNN requires storing the training data, it can be time-consuming and space-consuming. In two-dimensional space, the Euclidean formula can be used to calculate the distance between two data points (x2, x1) and (y2, y1) [23].

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5}$$

P and Q denote the training and testing tuples, respectively, and n denotes the total number of observations. Fij normally has a value between 1 and -1.

$$f_{ij} = 1 - \frac{6\sum_{i=1}^{n}\left(rank(P_i) - rank(Q_j)\right)^2}{n(n^2-1)} \tag{6}$$

3.6.4. Support vector machine

SVM is another supervised ML technique for classification. This approach uses a hyperplane to partition the dataset into separate groups to ensure an optimum margin for accuracy. A hyperplane is a function that best separates classes, and a margin is the distance between the two nearest data points of distinct classes to the hyperplane. These models learn to categorize, forecast, and find outliers [24], [25].

$$L(\omega) = \sum_{i=1} max\ (0,1 - y_i\ [\omega^T x_i + b]) + \lambda||\ \omega||_2^2 \tag{7}$$

Here, $\sum_{i=1}$ max $(0,1 - y_i\ [\omega^T x_i + b])$ is loss function and $\lambda||\ \omega||_2^2$ is regularization.

3.6.5. Gaussian Naive Bayes classifier

GNB classifier is a supervised ML classification algorithm. This method calculates object probabilities using the Bayes Theorem. Hence it is a probabilistic classifier like LR. Because the approach posits that the presence of any feature in a class is independent (unrelated) to the presence of any other feature in that class, it is named the Naive Bayes classifier. It's typically used for loops. Attributes should have a Gaussian or normal distribution. The bayes theorem is the foundation of the naive bayes algorithm, a probabilistic ML system. The bayes theorem is a straightforward formula for calculating conditional probabilities [26].

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{8}$$

This tells us how likely A is if B occurs, denoted P(A|B), also known as posterior probability. When we know how often B occurs when A occurs, we write P(B|A) and P(A) denotes how likely A is on its own, and P(B) denotes how likely B is on its own (B).

3.6.6. Random forest classifier

RF classifier is a classification technique that uses supervised ML. This algorithm combines several DT classifiers on distinct subsets of the input dataset, as the name implies the algorithm increases its predictive accuracy by taking the average of their results to determine a single outcome. More trees improve accuracy and reduce the likelihood of overfitting. Ensemble learning is the name for this concept. The RF adds randomness to the model. Instead of the most important property, it randomly splits nodes. So, the model varies. RF is a technique for learning trees by aggregating bootstraps. Let X=x1, x2, x3, ..., xn) with replies Y=x1, x2, x3, ..., xn) with a lower limit of b=1 and an upper limit of B [27]:

$$j = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{9}$$

3.6.7. Ada boost (AB) classifier

AB is the most prominent ensemble method, and it has significantly improved the base learner's prediction accuracy. It's a learning algorithm that generates several classifiers and uses them to create the best classifier possible. This approach has the advantage of requiring fewer input parameters.and requires only a rudimentary understanding of the weak learner. It also offers a great degree of flexibility, making it ideal for combining with other methods for identifying weak hypotheses. The ABt approach displays self-reported confidence levels that measure the accuracy of its predictions [28].

$$H_k\ (p) = +/- \left(\sum_{k=1}^{k} a_k\, h_k(p)\right) \tag{10}$$

Where K is the total amount of weak classifiers, $h_k(p)$ is the result of a poor classifier t (this can be either −1 or 1), a_k is the classifier k's weight

3.6.8. Gradient boost classifier

The GB technique is among the more advanced methods in the ML field. In general, there are two sorts of ML technique problems: bias errors and variance errors. GB is one of the boosting strategies employed to decrease the method's bias error. The base estimator of this approach cannot be defined, which unlike adaboosting technique. The GB technique has a fixed base estimator. The method can be utilised to predict categorical as well as continuous target variables as a regressor and classifier. Mean squared error (MSE) is the cost function for regressors, and log loss is the cost function for classifiers [29], [30].

$$F_m(X) = F_{m-1}(X) + \eta \times f_m(X) \tag{11}$$

Where F is the ensemble model, f represents the weak learner, η is the learning rate, and X represents the input vector.

3.7. Performance metrics

3.7.1. Confusion matrix

The confusion matrix is a 2×2 matrix that comprises the letters true negative (TN), false positive (FP), false negative (FN), and true positive (TP) at locations 11, 12, 21, and 22, respectively. It divides the categorization model's expected outputs into four categories: TN (correctively rejected), FP (incorrectly accepted), FN (incorrectly rejected), and TP (correctly accepted).

3.7.2. Accuracy

The proportion of all predictions that are predicted correctly (accurately) is referred to as accuracy. It is the measurement of properly categorized subjects relative to the total number of subjects.

$$Accuracy = (TP + TN)\ /\ (TP + TN + FP + FN) \tag{12}$$

3.7.3. Sensitivity

The proportion of positive cases projected as positive is identified as sensitivity. It is sometimes referred to as recall.

$$Sensitivity = (TP)\ /\ (TP + FN) \tag{13}$$

3.7.4. Specificity

The proportion of actual negative cases projected as negative is known as sensitivity. Specificity denotes to the proportion of people that test negative for a given disease within a population of healthy people.

$$Specificity\ =\ (TN)\ /\ (TN + FP) \tag{14}$$

3.7.5. False discovery rate

The proportion of actual negative cases projected as positive is known as FDR. It's also known as FR rate.

$$False\ Discovery\ Rate\ =\ (FP)\ /\ (TN + FP) \tag{15}$$

The sum of the FDR and the specificity is 1.

3.7.6. False omission rate

The proportion of positive cases projected as negative is known as FDR. It's also known as FN rate.

$$False\ Omission\ Rate\ =\ (FN)\ /\ (TP + FN) \tag{16}$$

The sum of the false omission rate and sensitivity is 1.

3.7.7. Area under the curve-receiver operating characteristics curve

The plot of TPR vs FPR at various threshold values is identified as the ROC curve. The AUC is a measure of separability that shows how well a model can successfully categorize classes. The higher AUC, the more classes are predicted accurately.

3.7.8. Interpretability

Determined as the percentage of masked features that do not provide any information to the final classification result divided by the total number of features in the dataset. This metric will be used to express how interpretable a model is.

$$I = \frac{masked\ features}{total\ input\ features} \tag{17}$$

3.7.9. Fidelity

Basically, "fidelity" relates to how effectively a model fulfills its intended purpose. Which measures the relationship between the accuracy of an equivalent fully interpretable model (i.e., DT) and its uninterpretable model counterpart.

$$F = \frac{Fully\ Interpretable\ Model\ Accuracy}{UnInterpretable\ Model\ Accuracy} \tag{18}$$

3.7.10. Fidelity-interpretability ratio

The ratio of a model's fidelity to the product of its fidelity and interpretability is known as the fidelity-to-interpretability ratio (FIR). This indicates how much interpretability is sacrificed for performance, with 0.5 being the best score.

$$FIR\ =\ F\ /\ (F + I) \tag{19}$$

4. RESULTS AND DISCUSSION

We proposed different ML algorithms for diagnosing and detecting breast cancer in this study. For pre-processing the breast cancer dataset, we employed the standardization approach, then we used different ML classification algorithms with the univariate features selection and RFE. Using all features, the LR and SVM had the highest accuracy prediction of 99.12%, while the DT and GNB had the lowest accuracy prediction of 92.11%. The accuracy of various models increased after using the univariate features selection technique, such as the DT from 92.11% to 94.74%, RF from 95.61% to 96.49%, GNB from 92.11% to 92.98%. After using the RFE technique, also increased the accuracy of various models such as DT from

92.11% to 92.98%, RF from 95.61% to 97.37%, GNB from 92.11% to 92.98%. When using AF, UFS, and RFE, we got RF, GB, and AB to have the highest sensitivity of 97.87%, while GB has the lowest sensitivity of 87.23%. The best specificity is 100% for LR, KNN, and SVM, while the worst specificity is 92.54% for LR, DT, and GNB as seen in Figure 3.

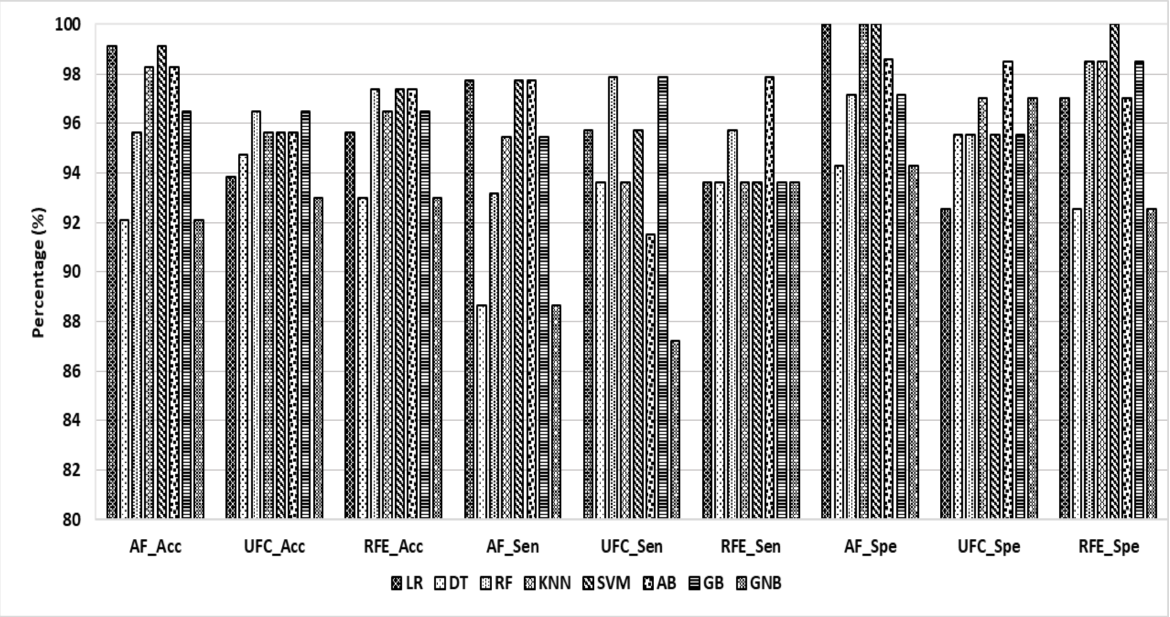

Figure 3. Comparison of classifiers in terms of accuracy, sensitivity, and specificity

The largest FDR was 7.46% for LR, DT, and GNB, while the lowest FDR was 0.00% for LR, KNN, and SVM. The AUC achieved with LR and SVM provides the highest results of 98.86% as seen in Figure 4. Table 1 summarizes the classification performance of the various ensemble tree algorithms following the training cross-validation module and their evaluation using the test set. The results demonstrate strong classification performance throughout the training phase, with 92.75 to 100% range for all parameters analyzed. This solid performance is maintained when using unseen data from the test, achieving more than 92% accuracy in all classifiers analyzed and less than 97% in the remaining classification metrics. Visual representation of the classifier's comparison is illustrated in Figure 5.

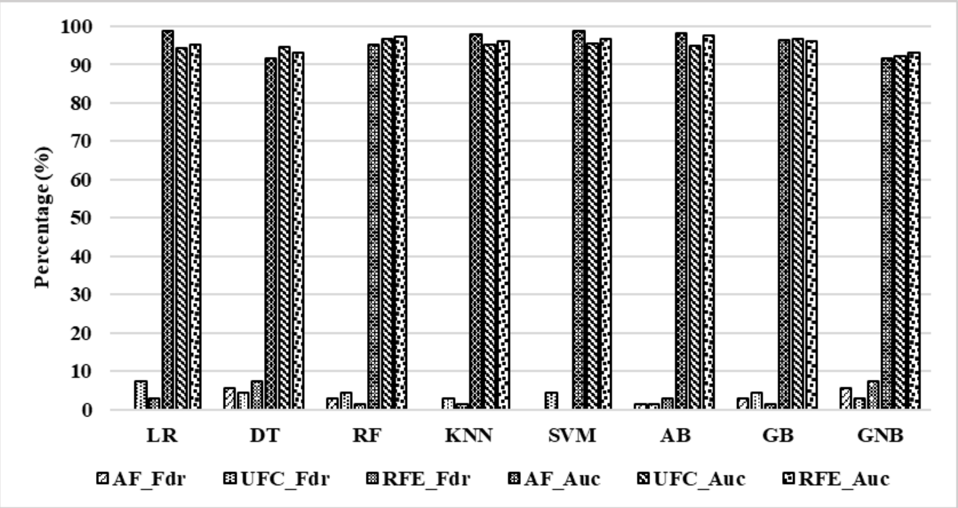

Figure 4. Comparison of classifiers in terms of false discovery rate and AUC

Table 1. Explainability metrics results

Classifier	Interpretability (%)	Fidelity (%)	FIR
Logistic regression	83	100	0.55
Decision tree	83	100	0.55
Random forest	83	98	0.54
K-neighbors	83	99	0.54
Support vector machine	83	99	0.54
Ada boost	83	99	0.54
Gaussian naive bayes	83	100	0.55
Gradient boost	83	100	55

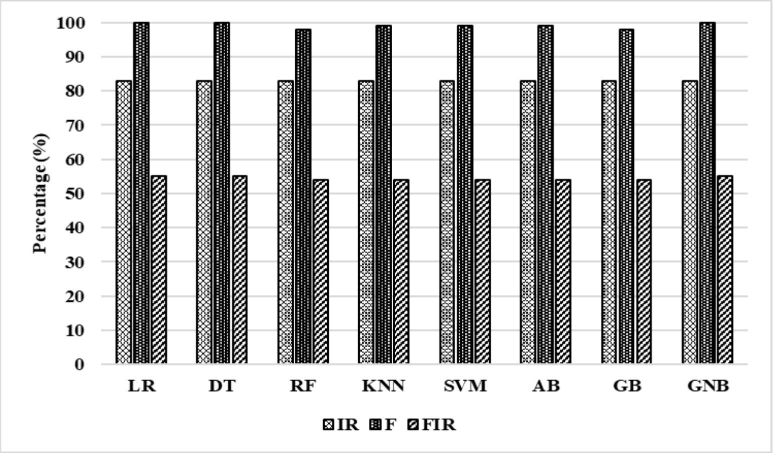

Figure 5. Visual representation of classifiers comparison in terms of interpretability, fidelity, and FIR

Once the top five features have been determined, explainability can be examined (shown in Table 1). A specific DT was constructed for each ensemble tree algorithm and its collection of selected characteristics to obtain the equivalent fully interpretable model for calculating the fidelity. Due to the fact that FIR provides a balanced measure of interpretability and fidelity with a cutoff of 0.5, RF, KNN, SVM, AB, and GB (FIR=0.54) achieved the most balanced model.

5. CONCLUSION

A comparison of multiple ML methods for the identification of breast cancer is presented in this paper. The WBCDc dataset was analyzed using eight different ML algorithms: LR, DT, RF, KNN, SVM, AB, GB, and GNB. The optimal ML approach is determined by calculating, comparing, and analyzing various outputs. The highest accuracy is based on the confusion matrix, accuracy, sensitivity, precision, and AUC. A detailed examination of the proposed models reveals that LR and SVM obtained a higher testing accuracy 99.12%, sensitivity of 97.73%, specificity of 100%, FDR of 0.00%, false omission rate of 2.27%, AUC of 98.86% and exceeds all other algorithms. Various evaluations using classification and explainability are conducted to determine the best-balanced model in terms of accuracy and explainability. As a result, the most explainable prediction model is one that incorporates RF, KNN, SVM, AB, and GB. It should be mentioned that all of the observed results are only related to the WBCD database. This could be considered a study's limitation. As a result, it's important to think about how to use these algorithms in future projects and to check the results obtained from this database using methods from other databases. Ongoing research aims to improve accuracy with more illness classes by applying existing and new algorithms to massive data sets.

REFERENCES

[1] "Breast cancer," *WHO*. Online [Available]: https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action (accessed Feb. 18, 2022).

[2] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 2016, pp. 1–4, doi: 10.1109/ICEDSA.2016.7818560.

[3] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 226–229, doi: 10.1109/R10-HTC.2017.8288944.

[4] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Procedia Computer Science*, vol. 191, pp. 487–492, 2021, doi: 10.1016/j.procs.2021.07.062.

[5] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast cancer detection using machine learning algorithms," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 114–118, doi: 10.1109/CTEMS.2018.8769187.

[6] P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 796–801, doi: 10.1109/ICSSIT48917.2020.9214267.

[7] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset," in *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, 2018, pp. 5–9, doi: 10.1145/3184066.3184080.

[8] F. Pedregosa *et al.*, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] "Scikit-learn user guide," *Scikit-Learn*. Online [Available]: https://scikit-learn.org/stable/user_guide.html (accessed Mar. 01, 2022).

[10] A. Pajankar and A. Joshi, "Getting started with NumPy," in *Hands-on Machine Learning with Python*, Berkeley, CA: Apress, 2022, pp. 23–30, doi: 10.1007/978-1-4842-7921-2_2.

[11] "Matplotlib overview," *Matplotlib*. Online [Available]: https://matplotlib.org/stable/contents.html (accessed Mar. 01, 2022).

[12] A. Pajankar and A. Joshi, "Introduction to pandas," in *Hands-on Machine Learning with Python*, Berkeley, CA: Apress, 2022, pp. 45–61, doi: 10.1007/978-1-4842-7921-2_4.

[13] M. Waskom, "Seaborn user guide and tutorial," *Seaborn*. Online [Available]: https://seaborn.pydata.org/tutorial.html (accessed Mar. 10, 2021).

[14] E. Bisong, "Google colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Canada: Apress, 2019, pp. 59–64.

[15] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer Wisconsin (diagnostic) data set," *Kaggle*, 2017. Online [Available]: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data (accessed Mar. 01, 2021).

[16] M. Lecocke and K. Hess, "An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data," *Cancer Informatics*, vol. 2, pp. 313–327, 2006, doi: 10.1177/117693510600200016.

[17] E. M. Senan *et al.*, "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–10, 2021, doi: 10.1155/2021/1004767.

[18] E. Celik and S. I. Omurca, "Improving parkinson's disease diagnosis with machine learning methods," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–4, doi: 10.1109/EBBT.2019.8742057.

[19] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, 2011, doi: 10.1007/s11222-009-9153-8.

[20] D. G. Kleinbaum and M. Klein, *Logistic regression*. New York: Springer, 2010, doi: 10.1007/978-1-4419-1742-3.

[21] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, 2021.

[22] F. M. J. M. Shamrat, Z. Tasnim, P. Ghosh, A. Majumder, and M. Z. Hasan, "Personalization of job circular announcement to applicants using decision tree classification algorithm," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 2020, pp. 1–5, doi: 10.1109/INOCON50539.2020.9298253.

[23] S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 2010, pp. 91–94, doi: 10.1109/FSKD.2010.5569740.

[24] S. Suthaharan, "Support vector machine," in *Machine Learning Models and Algorithms for Big Data Classification*, Boston, MA: Springer, 2016, pp. 207–235, doi: 10.1007/978-1-4899-7641-3_9.

[25] S. Afrin *et al.*, "Supervised machine learning based liver disease prediction approach with LASSO feature selection," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3369–3376, 2021, doi: 10.11591/eei.v10i6.3242.

[26] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Systems*, vol. 192, p. 105361, 2020, doi: 10.1016/j.knosys.2019.105361.

[27] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[28] R. E. Schapire, "Explaining adaboost," in *Empirical Inference*, Berlin, Heidelberg: Springer, 2013, pp. 37–52, doi: 10.1007/978-3-642-41136-6_5.

[29] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[30] C. Bentéjac, A. Csörgő, and G. M. -Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.

BIOGRAPHIES OF AUTHORS

Md. Murad Hossin is a final-year student in the Department of Computer Science and Engineering at Daffodil International University. He is preparing to become an expert in artificial intelligence and already has expertise in machine learning. His long-term objective is to advance medical diagnostics, particularly in developing nations. Some of his research interests include machine learning, deep learning, image processing, and artificial intelligence. He may be reached through email: murad15-2547@diu.edu.bd.

F. M. Javed Mehedi Shamrat graduated from Daffodil International University with a B.Sc. in Software Engineering in 2018. He was formerly employed with Daffodil International University. He is presently employed as a lecturer at the European University of Bangladesh in the Department of Computer Science and Engineering. He has been actively engaged in collaborative research with researchers from Bangladesh, the United States of America, Canada, China, Korea, and Australia. He has several research publications published in prestigious journals (Scopus) and conferences (Scopus). His primary areas of interest in the study include the internet of things, deep learning, data science, image processing, neural networks, artificial intelligence, bioinformatics, and machine learning. He can be contacted at email: javedmehedicom@gmail.com.

Md. Rifat Bhuiyan is currently pursuing a B.Sc. degree Department of Computer Science and Engineering at Daffodil International University. He has been heavily involved in collaborative research activities with researchers in Bangladesh. His research interests include machine learning, data mining, deep learning, information security, neural network, big data, and IoT. He can be contacted at email: rifat15-2375@diu.edu.bd.

Rabea Akter Hira is a final-year student in the Department of Computer Science and Engineering at International University of Business Agriculture and Technology. She is now preparing to become an expert in the machine learning field. Some of her research interests include machine learning and deep learning. She can be contacted at email: rabea.hira@gmail.com.

Tamim Khan has been enrolled in the B.Sc. degree at Daffodil International University in the Department of Computer Science and Engineering since 2019. He has collaborated on research projects with Bangladeshi researchers, particularly in the fields of machine learning, deep learning, and image processing. His main research interests include deep learning, machine learning, image processing, IOS development, and artificial intelligence. He enjoys reading history books and contrasting them with the present. He can be contacted at email: tamim15-2515@diu.edu.bd.

Shourav Molla studying at Daffodil International University in the Department of Computer Science and Engineering, under the B.Sc. program, since 2019. He has been involved in cooperative research activities with researchers from Bangladesh and researchers from Australia, especially in machine learning, deep learning, and image processing. His primary areas of interest in the study include deep learning, machine learning, image processing, web development, and artificial intelligence. He is interested in reading history books and comparing them with the current world. He can be contacted at email: shourav15-2438@diu.edu.bd.