Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Breast Cancer Diagnosis Using Wisconsin Dataset: A Non-Feature Selection Approach

Abdullah Al Mamun[1], Touhid Bhuiyan[2], Aminur Sarker[1], Shahin[1], Md Maruf Hassan[3]

[1]Department of CSE, Daffodil International University, Bangladesh
[2]School of IT, Washington University of Science and Technology, VA, USA
[3]Department of CSE, Southeast University, Bangladesh

Breast cancer ranks among the foremost causes of mortality in women, with a diagnosis rate of one in eight. Early detection improves therapy outcomes. Breast cancer can be predicted using different machine learning (ML) methods. We tested SVM, KNN, DT, RF, and LR models using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. After labelling and standardizing the dataset, we trained and evaluated the models using K-fold cross-valuation. Performance was assessed using F1 score, accuracy, precision, recall, and confusion matrices. The RF model had the highest accuracy 96.49%, followed by SVM 99.12% after hyperparameter adjustment. Logistic Regression, Decision Tree, and KNN achieved high accuracy 99.12%, 94.74% and 96.49%, respectively. These result show that ML algorithms can help breast cancer diagnosis earlier, increasing therapy and prognosis.

Keywords: Breast Cancer Prediction, Machine Learning (ML), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF); Logistic Regression (LR), Cross-Validation; Hyperparameter Tuning; WDBC Dataset

1. Introduction

Among the most common cancer worldwide, breast cancer kills women most of the time. Approximately 508,000 women died from breast cancer, in 2011, according to the WHO. Current figures show that one in eight women will develop breast cancer [**Error! Reference source not found.**]. Early and precise detection improves patient survival rates through timely intervention and therapy. Mammograms and biopsies are effective but invasive and error-prone. Breast cancer, a primary cause of death in women, starts in breast tissue and spreads. In 2018, the disease caused 9.6 million deaths globally, with a 50\% increase in cases by 2040.Data mining and big data technology help forecast and treat breast cancer, improve patient care, and lower healthcare costs[**Error! Reference source not found.**].Data-driven medical diagnoses are now possible because to machine learning (ML) advances.

These algorithms can analyze massive medical data and find patterns that diagnosticians may

miss.

This compare of multiple ML models for breast cancer prediction using the WDBC dataset. [10]SVM, KNN, DT, RF, and LR are examined. To optimize prediction accuracy [5], precision, recall, and other performance measures without feature selection, we focus on hyperparameter tuning. Hyperparameter tuning optimizes model performance by systematically altering learning process parameters [7] to identify the optimum algorithm configurations. Structure of the rest of the paper: Section 2 reviews machine learning breast cancer prediction literature. Section 3 describes the dataset and preprocessing, whereas Section 4 describes the training and evaluation machine learning models and methods. Section 5 shows results and discusses model performance. Section 6 finishes the analysis and suggests breast cancer prediction research directions.

2. Related Work

As you may know, classification is a very important part of machine learning, and a lot of study has been done on different medical datasets of breast cancer in this area. By using different classifier models for the classification problem, these different research projects got different results in terms of how well they worked. Below is a list of:

S. A. Abdulkareem, et al. [1] [2021] Wisconsin Breast Cancer Dataset (WBCD) and the Recursive Feature Elimination (RFE) algorithm are used in this study to show how well the Random Forest and XGBoost classifiers work for finding breast cancer. The high level of accuracy reached by XGBoost 99.02% shows that ensemble models are useful for medical tasks. When it comes to classification tasks, ensemble methods often work better than single classifiers. This is often seen in finding breast cancer, and machine learning classifiers like SVM and Random Forest have been used a lot.

Naji Mohammed Amine et al. [2] [2021] This paper's author says that improving the WDBC prediction for high accuracy is important to keep treatment and survival rates up to date. Once they had the results, they used five machine learning algorithms on the Breast Cancer Wisconsin Diagnostic dataset: SVM, RF, LR, DT (C4.5), and KNN. The goal of this study is the use machine-learning model to identify and diagnose breast cancer and find the best ones in terms of confusion matrix accuracy and precision. Support vector Machine did better than all the other classifiers and got the best accuracy (97.2%).

This is Kadhim, R. R. et al. [3] [2022]. The main point of the study is to compare different ways to classify breast cancer using machine learning algorithms. With a score of 96.77, extreme randomise trees had the best F1-score out of the eleven models tested using the Wisconsin dataset. Specificity, sensitivity, precision, accuracy, and F1 score were used to rate how well each model worked. The goal of this study is to help find breast cancer early by finding the best Machine Learning models for classification.

Hossin, M. M., et al. [4][2023] This article looks at eight machine learning methods for finding breast cancer. These are LR,RF,KNN,DT,AB,SVM,GB, and GNB. The Wisconsin Diagnostic Dataset is used to test these models and make sure they work. Sensitivity, specificity, Accuracy, and area under the curve (AUC) were used to measure how well the model worked. Logistic Regression: Out of all the methods, it works 99.12% of the time. Researchers said

that the study shows how important it is to find and treat breast cancer early so that people can live.

Rasool, Abdur, et al. [5] (2022) This paper is mostly about the WDBC approach. The author used a four-layer data exploratory method (DET) to make the model work better. This technique included feature selection, correlation analysis, and hyperparameter optimization. The polynomial SVM model was the most accurate 99.3%. It was followed by the LR model 98.6%, the KNN model 97.35%, and the EC model 97.61%. The study used K-fold cross-validation and confusion matrices to show that the models worked even better. These results are in line with other study that has shown that SVM models are better at diagnosing breast cancer.

Aboudr MAA et al. [6] (2023) this study suggests the FLN algorithm as a way to make Breast Cancer diagnoses more accurate. A) The FLN method can get rid of overfitting; b) it can handle binary and multiclass classification problems; and c) it can work like a kernel-based support vector machine with the structure of a neural network. They used the WBCD, which is a breast cancer database. The experiment showed that the suggested FLN method worked very well, with an average of 98.37% accuracy, 95.44% precision, 99.40% memory, 97.644% F-measure, 97.654% G-mean, 96.444% MCC, and 97.854% specificity using the WBCD. This shows that the FLN method is a good way to diagnose BC, and it might also help with other problems in the healthcare field that have to do with applications.

Sara Ibrahim et al. [7] (2021) The WBCD was used to test the author's suggested approach in this paper. For reducing the number of dimensions, analysis of correlation. Well-known machine learning models were tested to see how well they worked, and the seven best ones were picked for the next step. Tuning the hyperparameters was done to make the algorithms work better. Two different vote methods mixed with the classification algorithms that worked the best. Hard voting picks class that pick the most votes, while soft voting picks the class that has the best chance of winning. With an accuracy of 98.24%, a high precision of 99.29%, and a recall value of 95.89%, the suggested method did better than the best work that had been done before.

Arpit Bhardwaj et al. [8] (2022) by this work compares four algorithms that are used for the WBCD dataset. These are MLP, KNN, GP, and RF, which are all classification algorithms. which was made by taking samples of the breast with a fine needle. We used genetic programming (GP), random forest (RF), multilayer perceptron nearest neighbor (MLP), and K-nearest neighbor (KNN) on the WBCD dataset to sort the patients into those who are benign and those who are cancerous. RF has a classification rate of 96.24%, which is better than all the other classifiers. Based on the data of the suggested method, probable breast cancer is labelled.

Adel S. Assiri et al. [9] [2020] The WBCD was used to compare how well different cutting-edge machine learning classification methods worked. Based on their F3 score, the three best models were then chosen. The F3 score is used to stress how important false positives are in classifying breast cancer. Simple logistic regression learning, support vector machine learning with stochastic gradient descent optimization, and multilayer perceptron network are the three classifiers that are used for ensemble classification with a vote system. With a success rate of 99.42%, the hard voting (majority-based voting) method works better than the most recent

WBCD algorithm.

Neha Panwar et al. [10] [2020] in this study, we use different Machine Learning (ML) techniques to figure out if a patient has BC or not. SVM,k-NN,NB,DT, and LR will be used to sort the WDBC dataset in this work. Before classification, there is a preprocessing step where five different classifiers are used with the fivefold cross-validation method. Performance factors like sensitivity, accuracy, and specificity are used to measure how well classification works. Confusion metrics are also used to measure performance. It was found that SVM worked best, with a precision of 99.12% after the normalization process in

3. Preliminary Section

A. Data Description: The Wisconsin Breast Cancer (WBC) dataset is derived from the UCI repository of machine learning datasets [17]. This collection includes 569 instances, categorized as either benign or malignant, with 357 instances (62.74 percent) identified as benign and 212 instances (37.25 percent) as malignant. The dataset is segmented into two categories, B for benign and M for malignant. Breast cancer stands as the most frequently diagnosed condition in healthcare, and its incidence is on the rise annually. Beyond the sample code numbers and class labels, the dataset features 32 characteristics related to breast cancer, such as the mean radius, texture, area, smoothness, compactness, and concavity [18, 19]. Cases labeled as benign are considered less harmful to the body, whereas those labeled as malignant are deemed harmful due to their cancerous nature in our research. The dataset contains 16 instances with missing values for features, which are typically filled using the mean method. To guarantee the integrity of the data, the dataset is randomized at the end.

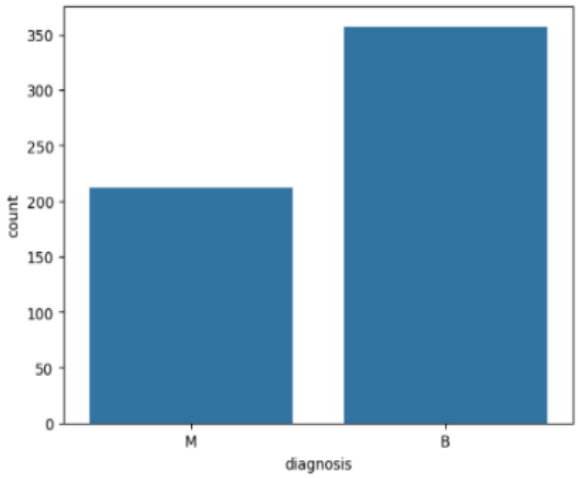

Figure 1. WISCONSIN BREAST CANCER DIAGNOSTIC DATASETS.

B. Preprocessing:

Table 1. Features categorization of WDBC dataset.

NO.	Feature	NO.	Feature	NO.	Feature

1	Radius mean	11	Radius SE	21	Radius worst
2	Texture mean	12	Texture SE	22	Texture worst
3	Perimeter mean	13	Perimeter SE	23	Perimeter worst
4	Area mean	14	Area SE	24	Area worst
5	Smoothness mean	15	Smoothness SE	25	Smoothness worst
6	Compactness mean	16	Compactness SE	26	Compactness worst
7	Concavity mean	17	Concavity SE	27	Concavity worst
8	Concavepts. mean	18	Concavepts. SE	28	Concavepts. worst
9	Symmetry mean	19	Symmetry SE	29	Symmetry worst
10	Fractaldim. mean	20	Fractal dim. SE	30	Fractaldim. worst

C. Performance Evaluation Metrics: Four distinct cross-valuation metrics precision, recall, F1 score, and accuracy were examined in this work. The values of the confusion matrix allow one to ascertain these measures. True positives (TP) when it predict yes and the actual data is also yes; true negatives (TN) when the prediction no and the actual data also no; false positives (FP) when the prediction yes but the actual data is no; and false negatives (FN) when the prediction is no but the actual data is yes. The formulae below allow one to calculate accuracy, F1-score, precision, recall:

$$\text{Precision(P)} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall(R)} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{F1-score} = \frac{2 \times P \times R}{P+R} \quad (3)$$

$$\text{Accuracy(A)} = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

D. Methodology:

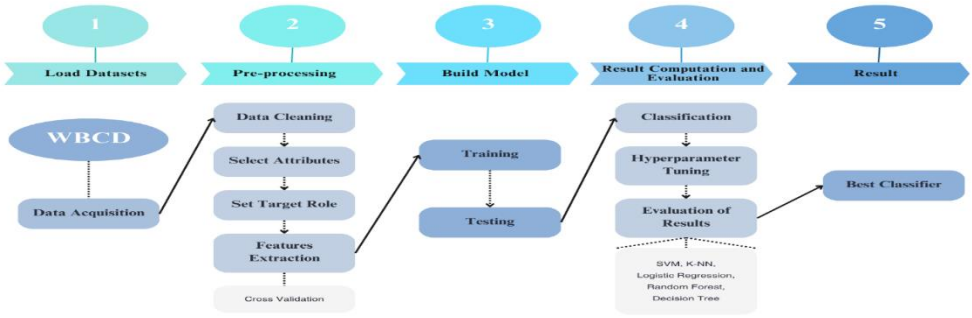

Figure 2 Process Flow Diagram

Our main aim in this study is to identify the most dependable and effective approach for breast cancer detection. SVM and Hyper parameter Tuning, RF, LR, DT (C4.5), and KNN on the WDBC have been used among other machine learning methods to do this. We next evaluated the results to choose the model with best accuracy. Figure 2 shows the proposed design.

Our approach begins with data collecting and proceeds through pre-processing, which consists in main parts data cleansing, attribute selection, target role definition, and feature extraction. Then, the processed data is used to create machine learning models able to forecast breast cancer depending on fresh measurements. We expose these models to labelled new data to evaluate their performance. Usually, this is accomplished by splitting the labelled data in two subsets applying the Train_test_split technique. Referred as the training set, 80% of data set for model training; the remaining 20% set for model performance evaluation, sometime known as the test set. After the evaluation, we match the findings to find the most suitable algorithm for breast cancer detection and ascertain the model with best accuracy.

4. Machine Learning Algorithms

With our study, we effectively applied machine learning techniques to predictive analysis. The project's machine learning techniques consist in.

A. Support Vector Machine (SVM)

VM is the classifier that uses the data point to divide the data into several categories so locating the broadest maximum marginal hyperplane (MMH) [2].

B. Random Forests (RF)

Random Forests, known as random decision forests, is hybrid approaches for classification, regression, and other problems whereby huge number of decision trees generated in training and produce classes reflecting classes (clusters). Forecast (decline) for every division. Based on their fit to their training set, random decision forests are dependable.

C. K-Nearest Neighbor (KNN)

KNN classifies the query data according on the similarity measure while storing all the training data. In KNNs, k stands in for the voting process's included neighbor count. KNNs follow a similar strategy. Choosing the right value of k helps to enhance performance by means of KNN parameterization. For instance, the Euclidean distance [9] helps one find the similarity between two places.

D. Logistic Regression (LR)

An effective modelling tool with a development from linear regression is logistic regression [2]. Using a risk factor or type logistic regression evaluates the risk of disease or health condition. Simple and multiple logistic regression is link between the dependent variable (Y) - sometimes known as the outcome, or response variable - and the independent variable (Xi) - sometimes known as the exposure variable or predictor variable. Usually use estimate binary or multiclass dependent variables of it.

E. Decision Tree (DT)

Predictive modelling tool Decision Tree C4.5 finds applications in various spheres. It can be built using an algorithmic method allowing different divisions of data sets depending on various criteria.

F. Parameter Optimization

In machine learning, parameter optimization is the method used to find the ideal collection of values. The learning process is under control with reference to the values of this parameter. Grid search, random search, Bayesian optimization, gradient-based optimization, evolutionary optimization and population optimization are only a few of the several approaches for meta parameter optimization. We applied network search optimization in this work since the obtained useful outcomes would help to optimize. This approach generates candidates from the grid using the given parameter value, therefore using their ability. Maximum mutual reliance is the aim of network search. Given its disease prediction dataset, we in our situation applied scikit. GridSearchCV was the metaparameter estimation tool utilized in all prediction models. GridSearchCV conducts the analysis using a separate set of designated meta parameters and their values.

5. Experimental Results & Discussions

A. Experimental Result: This collection comprises information gleaned from microscopic analysis of breast tumors. The activity was calculated by means of a computerized scan of the needles. One of the greatest techniques to assess the existence of malignant tumors is fine needle aspiration. This data set comprises of 569 samples. Every model comprises thirty-two extracted characteristics from the main pictures. The mean, standard error, or worst-case approach of the above described functions helped one to determine the remaining ones. Figure 2 displays the variance distribution of the specified characteristics; the x-axis denotes the value of the attribute and the y-axis shows the frequency of every value for the two groups. Figure 2 displays both primary and secondary traits related to the diagnostic and illness class severe mild disease. Results given in red and blue accordingly reflect light and dark tones respectively. As is common knowledge from the literature, evaluation was based on an 80% to 20% practice test split.

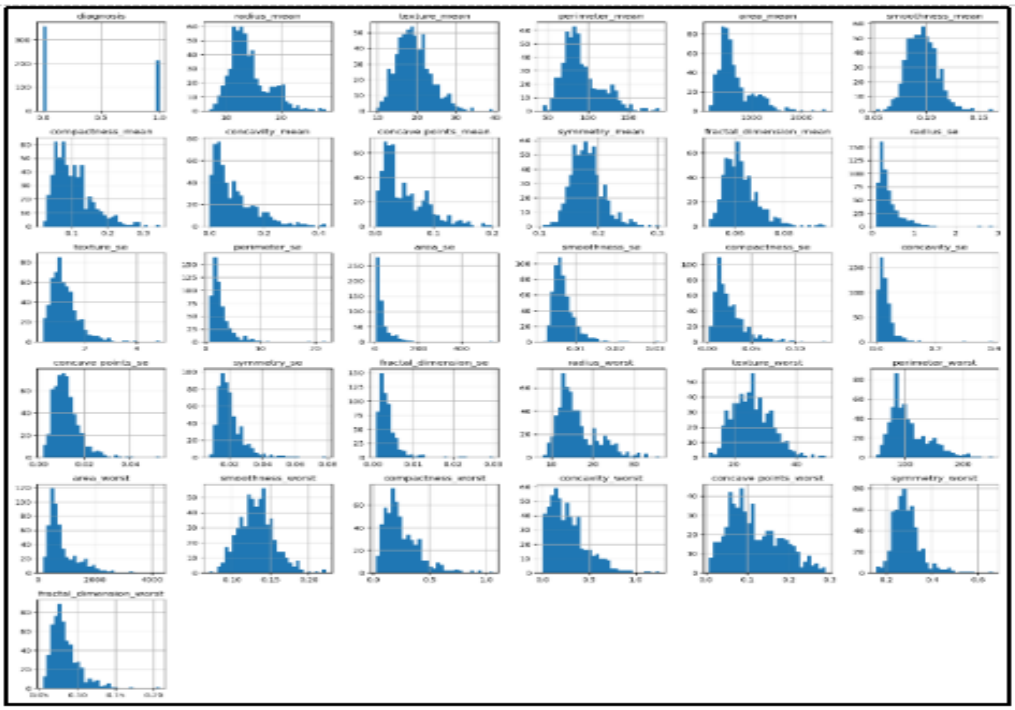

Figure 3. Feature Visualization result for WBCD.

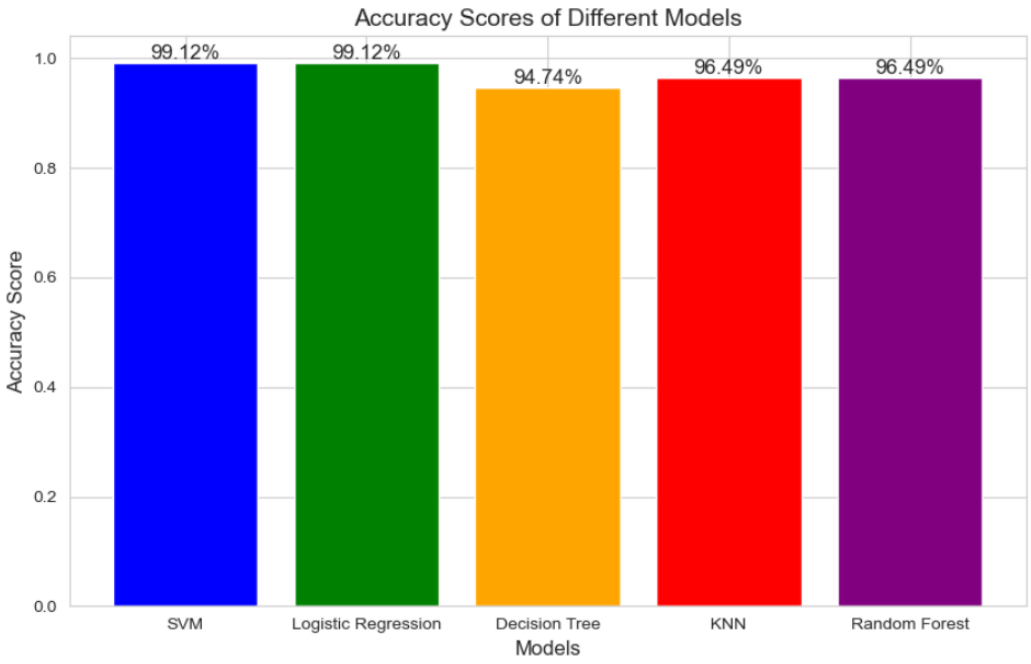

Figure 4. Performance of the machine learning algorithm on WBCD.

B. Discussions: Using the Wisconsin Diagnostic Breast Cancer (WDBC) data, we presented different machine learning (ML) algorithms for breast cancer identification. In pre-processing, we followed the usual technique using all the dataset's features without any segmentation. All models' performance was much enhanced by metaparameter optimization applied using GridSearchCV and RandomizedSearchCV.

After calibration, logistic regression and support vector machine (SVM) attained an accuracy of 99.12% demonstrating its performance in high-dimensional classification problems. Showing the need of sound in lowering redundancy, the lower performing Decision Tree model improved to 94.74% following optimization. Emphasizing the part of meta-parameter optimization to improve model performance, Random Forest models with Nearest-neighbors (KNN) attained 96.49% accuracy.

These findings demonstrate how well ML techniques particularly following meta-parameter modification can enhance early breast cancer identification. To guarantee more general clinical relevance, future research must thus validate these conclusions in bigger and more varied datasets.

6. Conclusion

WDBC dataset utilizing LR, SVM, KNN, decision tree (DT), and random forest (RF) among several ML approaches. The accuracy of every model was much raised via hyperparameter optimization. With a 99.12% accuracy, logistic regression and SVM did rather well among the models. With 96.49% accuracy, Random Forest and KNN also shown good performance; the Decision Tree model rose to 94.74%. These results imply that machine learning techniques, especially when combined with metaparameter modification can offer precise and efficient instruments for the diagnosis of breast cancer. One restriction of this work is that machine learning is limited to numerical data. We shall aim to work properly with photos using several image extraction techniques in the future.

Acknowledgment

This research work is supported by School of IT, Washington University of Science and Technology, VA, USA

References

1. Abdulkareem, S. A., & Abdulkareem, Z. O. (2021). An evaluation of the Wisconsin breast cancer dataset using ensemble classifiers and RFE feature selection. Int. J. Sci., Basic Appl. Res., 55(2), 67–80.
2. Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida AitAbdelouhahid, & Olivier Debauche. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, 487–492.
3. Kadhim, R. R., & Kamil, M. Y. (2022). Comparison of breast cancer classification models on Wisconsin dataset. Int. J. Reconfigurable Embed. Syst., ISSN 2089-4864.
4. Hossin, M. M., Shamrat, F. J. M., Bhuiyan, M. R., Hira, R. A., Khan, T., & Molla, S. (2023). Breast cancer detection: An effective comparison of different machine learning algorithms on

the Wisconsin dataset. Bulletin of Electrical Engineering and Informatics, 12(4), 2446–2456.
5. Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. Int. J. Environ. Res. Public Health, 19, 3211. https://doi.org/10.3390/ijerph19063211.
6. Albadr, M. A. A., Ayob, M., Tiun, S., AL-Dhief, F. T., Arram, A., & Khalaf, S. (2023). Breast cancer diagnosis using the fast learning network algorithm. Front. Oncol., 13, 1150840. https://doi.org/10.3389/fonc.2023.1150840.
7. Ibrahim, S., Nazir, S., & Velastin, S. A. (2021). Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. J. Imaging, 7, 225. https://doi.org/10.3390/jimaging7110225.
8. Bhardwaj, A., Bhardwaj, H., Sakalle, A., Uddin, Z., Sakalle, M., & Ibrahim, W. (2022). Tree-based and machine learning algorithm analysis for breast cancer classification. Computational Intelligence and Neuroscience, 2022, 6715406.
9. Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. Journal of Imaging, 6(6), 39.
10. Panwar, N., Sharma, D., & Narang, N. (2020). Breast cancer classification with machine learning classifier techniques. In Proceedings of the 4th International Conference: Innovative Advancement in Engineering & Technology (IAET).