Date of publication JUN-30, 2024, date of current version FEB-11, 2024.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.2.3544

An Optimized Framework Based on Data Exploration and Dynamic Ensemble-Based Models for Breast Cancer Prediction

AYMAN AISABRY[1,2], HAMZAH ALI ABDULRAHMAN QASEM[2], MALEK AIGABRI[1], AMIN MOHAMED AHSAN[2], MOGEEB A.A. MOSLEH[2], F. E. HANASH[2,3]

[1] Department of Computer Science, Sana'a University, Sana'a, Yemen
[2] Department of Computer Science, International University of Technology Twintech, Sana'a, Yemen
[3] Emirates International University, Sana'a, Yemen

Corresponding author: Ayman Alsabry (aymanalsabri@su.edu.ye).

ABSTRACT Breast cancer (BC) is a major global health concern. Detecting BC at an early stage gives more treatment options and can help avoid more aggressive treatments. The use of machine learning (ML) in BC prediction offers significant potential for improving the accuracy and speed of diagnosis, personalizing treatment, and identifying high-risk patients. However, there are significant challenges associated with the use of ML, including the need for high-quality data and more flexible models with optimal parameters to achieve high efficiency. In this paper, we propose an optimized framework based on multi-stage data exploration. This framework is designed to provide a comprehensive approach to data exploration, ensuring that the data is well-prepared for ML. In addition, the framework includes dynamic ensemble-based classifiers, which combine multiple independent classifiers to improve accuracy and mitigate the risk of overfitting in conjunction with the cross-validation techniques. These classifiers are optimized using Bayesian hyperparameter tuning, which involves selecting the optimal values for the various hyperparameters of the model. This approach can significantly improve the prediction accuracy of the resulting model. The study evaluates the framework using the publicly available Wisconsin Diagnostic Breast Cancer (WDBC) dataset and compares our results with other state-of-the-art models. The experimental results show that the best result is 100% for accuracy and recall with hyperparameters of (Ensemble Method = AdaBoost, Number of learners = 322, learning rate = 0.9350, and the Maximum number of splits = 1). The highest performance has been achieved with the proposed framework compared with the other models in terms of accuracy (mean = 99.35%, best = 100%, worst = 98.7%, and Standard Deviation = 0.325). The framework can potentially improve the accuracy and efficiency of BC prediction, ultimately leading to better outcomes for patients.

KEYWORDS Data Exploration; Ensemble Classifier; Hyperparameters Tuning; Machine Learning.

I. INTRODUCTION

BREAST cancer (BC) is a type of cancer that develops in the breast cells. It is one of the most common cancer types, affecting women and men of all ages, although it is more common in women [1, 2]. The exact causes of BC are not fully understood, but there are several risk factors that can increase the likelihood of developing the disease [3]. One of the most significant risk factors is age. The older a person is, the higher the risk of developing BC. Other risk factors include family history, personal history, certain genetic mutations, and exposure to certain hormones.

Symptoms of BC can vary, but they often include a lump or mass in the breast, a change in the breast size or shape, skin changes, such as dimpling or puckering, and nipple discharge. It is important to note that not all lumps in the breast are cancerous; many other conditions can cause similar symptoms. However, it is always important to see a healthcare provider if any changes are noticed in the breast [4].

Diagnosing BC usually involves a combination of imaging and biopsies. Imaging tests, such as mammograms, ultrasounds, or MRIs, can help detect abnormalities in the breast tissue, while a biopsy involves taking a small sample from a tissue for further examination under a microscope [5, 6].

The treatment depends on several factors, including the type and stage of the cancer, as well as the individual's overall health. Treatment options may include a surgery, radiation

therapy, chemotherapy, hormone therapy, or targeted therapy [7].

Prevention of BC involves maintaining a healthy lifestyle, such as regular exercise, maintaining a healthy weight, and limiting alcohol intake. It is also important for individuals with a higher risk to have regular BC screenings and to discuss their risk factors with their healthcare providers. BC can be a life-threatening disease, but with early detection and appropriate treatment, the outlook for many individuals is positive. It is important for individuals to be aware of their risk factors and seek medical attention if any changes are noticed in their breasts [8]. Ongoing research and advances in treatment continue to provide hope for individuals affected by BC [9].

Early-stage detection is critical in improving individuals' prognosis and survival rates with the disease. The treatment can be initiated before cancer has a chance to spread beyond the breast, which can significantly improve the outcomes [10]. One of the most significant benefits of early-stage detection is that less invasive treatment options may be available. For example, if BC is detected at an early stage, a lumpectomy, i.e. removing only the cancerous tissue and not the entire breast, may be a viable treatment option. This can have significant physical and psychological benefits for patients, as they can retain their breasts and avoid more extensive surgery [11]. In addition to improving treatment options and outcomes, early-stage detection can reduce the need for more aggressive treatments, such as chemotherapy. When BC is detected at an advanced stage, the likelihood of cancer spreading to other parts of the body is higher, which means that more aggressive treatment options may be necessary. However, if the cancer is detected early, there may be a greater opportunity to avoid such aggressive treatments [12]. Early-stage detection of BC also means that individuals can receive support and guidance earlier in their cancer journey. This can include psychological and emotional support, as well as practical support, such as access to financial assistance and resources for managing the side effects of treatment.

Machine learning (ML) has emerged as a powerful tool for BC prediction, offering the potential to improve the accuracy and speed of diagnosis. ML algorithms can analyze huge amounts of data, identify patterns, and make predictions based on those patterns. In the context of BC, this can include analyzing medical images, genetic data, and other patient data to identify the signs of BC. Diagnosis accuracy can be increased by training ML algorithms to recognize subtle patterns that may be challenging for human specialists to recognize [13, 14]. ML can also improve the speed of diagnosis. Traditional methods of diagnosis can be time-consuming and may require multiple tests, while ML algorithms can provide a rapid and accurate diagnosis based on a single image or dataset [15]. Another key benefit is the potential to personalize treatment. ML algorithms can be trained on patient data to identify patterns and predict treatment outcomes. This can help healthcare professionals tailor treatments to individual patients based on their unique characteristics, improving the outcomes and reducing the risk of adverse effects [16]. Additionally, ML can help healthcare professionals identify high-risk patients who may benefit from more frequent screening or surveillance. By analyzing patient data, ML algorithms can identify patients at a higher risk of developing BC and recommend appropriate screening and surveillance strategies [17, 18].

Ensemble classifiers have gained popularity in ML in recent years, with numerous benefits attributed to their use. An ensemble classifier is a combination of multiple individual classifiers, which are trained independently, and then their predictions are combined to make a final prediction [19]. This approach has been shown to improve the accuracy and robustness of models. Another important benefit of ensemble classifiers is their ability to alleviate the overfitting [20].

Overfitting occurs when a model is trained on a particular dataset to the extent that it begins to memorize the data rather than learn the underlying patterns and relationships. As a result, the model may perform well on the training data but poorly on new, unseen data [21]. Ensemble classifiers can help avoid overfitting by combining the predictions of multiple individual classifiers, reducing the risk of over-reliance on any particular model. Using multiple independent models, ensemble classifiers can help ensure the resulting predictions are robust and reliable [20]. Cross-validation is a widely used technique in ML commonly used in conjunction with ensemble classifiers to help avoid overfitting. It involves dividing the data into multiple subsets, with one subset used for testing and the remaining for training. By repeating this process multiple times, rotating the test, and training subsets, cross-validation can estimate the model's performance on new, unseen data. This approach can help ensure that the model is not overfitting to the training data, as it is being evaluated on multiple different subsets of the data [22].

Hyperparameter tuning is a crucial aspect of ML that involves selecting the optimal values for the various hyperparameters of a model. A key benefit of hyperparameter tuning is improving the accuracy of a model. By fine-tuning the hyperparameters, the model can capture the underlying patterns and relationships in the data more effectively, resulting in more accurate predictions. This is particularly important in applications where high accuracy is critical, such as medical diagnosis or financial forecasting. Another important benefit of hyperparameter tuning is the improved robustness of the model. Robustness refers to a model's ability to perform well in a wide range of environments and conditions. By tuning the hyperparameters, the model can be more robust to changes in the underlying data distribution or the data itself. This is particularly important in applications with highly dynamic or noisy data, such as in online advertising or fraud detection [23].

This study aims to accommodate the previous challenges by developing a multi-stage approach to data exploration and preprocessing to ensure the data is well-prepared for ML, tuning the hyperparameters of dynamic ensemble-based classifiers, which combine multiple independent classifiers to achieve better performance. Cross-validation is used in conjunction with ensemble classifiers to help avoid overfitting.

The rest of this paper is organized as follows: Section II summarizes the literature of related works, and Section III illustrates the proposed study framework; Section IV presents the data exploration methods; Section V discusses the preprocessing techniques; and Section VI describes the ML models to predict BC. The experimental findings utilizing the proposed framework are presented in Section VII. Finally, the paper is concluded in Section VIII.

II. LITERATURE REVIEW

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is widely used in ML studies, especially in medical diagnosis.

The dataset contains 569 BC tumor biopsy samples, each with 30 features. The features include the mean, standard deviation, and worst values of 10 different characteristics, such as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, fractal dimension, and diagnosis (M=malignant, B=benign). The dataset has been widely studied, and numerous ML techniques have been applied to classify whether a tumor is benign or malignant.

One of the earliest studies on the WDBC dataset was conducted by Wolberg et al. in 1995 [24]. They used a backpropagation neural network to classify the tumors and achieved an accuracy of 96.5% with their model. Although the study was conducted over two decades ago, it remains an important reference for applying ML to medical diagnosis.

In 2016, Aalaei et al. [25] explored using a genetic algorithm for feature selection in diagnosing BC. Three different datasets: WDBC, Wisconsin Prognostic Breast Cancer (WPBC), and Wisconsin Breast Cancer" dataset (WBC), were used to evaluate the algorithm's performance. The results showed that it could identify relevant features and improve accuracy in all three datasets. The study concluded that genetic algorithms could be an effective tool for feature selection in BC diagnosis.

In 2017, Jeyasingh et al. [26] examined a modified version of the Bat Algorithm (MBA) for feature selection to identify and remove irrelevant features from an original dataset. The Bat Algorithm was adapted by incorporating simple random sampling to select instances from the WDBC dataset randomly. The modified bat algorithm was found to be effective in selecting features that could accurately classify the WDBC dataset. The results also showed that the modified bat algorithm reduced the number of features used in classification while still achieving a high accuracy rate.

In 2018, Yue et al. [27] examined the use of various ML algorithms, including Artificial Neural Network (ANN), Decision Tree (DT), Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN), to improve the accuracy of BC diagnosis and prognosis. The results suggest that ML can be used to improve the accuracy and efficiency of BC diagnosis and prognosis.

In 2019, Latchoumi et al. [28] examined the use of a bio-inspired weighted quantum particle swarm optimization (QPSO) and smooth support vector machine (SVM) ensembles for the identification of abnormalities in medical data. The QPSO algorithm was used to optimize the parameters of the SVM model, while the SVM model was used to classify medical data into normal and abnormal classes. Results showed that the proposed method achieved better accuracy of 98.42% than traditional methods, demonstrating its potential for use in medical data analysis. Showrovet et al. [29] selected 10 features from the WDBC dataset and trained three different models on them: ANN, Naive Bayes, and SVM. They achieved a better accuracy of 96.72% with the SVM model.

In 2020, Sheth et al. [30] proposed frameworks based on Feature Selection Technique based on Support Vector Machine (FSTBSVM), ML, and Jaya optimization techniques. The framework achieved an accuracy of 94.36%.

In 2021, Chugh et al. [31] compared the performance of deep learning (DL) and ML by surveying computer systems that use one of these techniques. The study found that, with large data, the DL showed better performance while the ML was suitable for small datasets. Ara et al. [32] practiced various ML classifiers and conducted a performance evaluation to determine the best ML classifier to predict BC more accurately. Six classifiers were used: SVM, Logistic Regression (LR), K-Nearest Neighbors (KNN), DT, Naive Bayes, and Random Forest (RF). With an accuracy of 96.5%, the RF and SVM classifiers surpassed the other four. Gopal et al. [33] developed methods to leverage the Internet of Things (IoT) and ML for early BC diagnosis. Three classifiers, Multi-Layer Perceptron (MLP), Random Forest (RF), and logistic regression, were used on the WDBC dataset. The MLP classifier surpassed the other two with an accuracy of 98% and a lower error rate than LR and RF. Assegie et al. [34] applied the adaptive boosting (Adboost) algorithm to eliminate the bias towards the benign observations that appeared when using the decision tree. The Adboost and DT algorithms achieved an accuracy of 92.53% and 88.80%, respectively. Lahoura et al. [41] employed a cloud-based extreme learning machine (ELM) approach and achieved an impressive accuracy of 98.68%.

In 2022, Hemavathi et al. [35] applied deep learning techniques to predict BC early. First, through a heat map, the proposed framework reduced the WDBC dataset's attributes to 14 out of 30 attributes. Second, the model used the optimal attributes to train ML algorithms, such as LR, KNN, SVM (Radial Basis Function), SVM (Linear), DT, RF, and Gaussian Naive Bayes (NB), with Bagged Trees, Subspace discriminant, and Random Under-Sampling (RUS) Boosted Trees. Finally, the study compared the performance metrics of the earlier ML classifiers. The RF performed better than other classifiers, with a 98.6% accuracy rate. Monirujjaman Khan et al. [36] compared the performance of four ML algorithms' predictions: RF, LR, DT, and KNN. The findings showed that the LR had the best result of 98% accuracy. Samieinasab et al. [37] developed a framework based on ensemble techniques to predict BC more accurately. The Extra Tree algorithm is used to pick the suitable features as inputs to the classification model with the stacking approach. The proposed framework was applied to WDBC's BC dataset and achieved an accuracy of 98%. Rasool et al. [38] aimed to develop four exploratory techniques: feature distribution, hyperparameter optimization, correlation, and elimination. Compared with the three ML models, polynomial SVM gained 99.3% accuracy. Bhardwaj et al. [2] implemented MLP, KNN, and RF on the WDBC dataset. The main objective of the study was to sort the tumors as benign or malignant. The RF classifier achieved the highest accuracy of 96.24% compared to other classifiers. Saleh et al. [1] proposed a deep recurrent neural network (RNN) model consisting of five hidden layers and five dropout layers. The Keras–Tuner method was used for the WDBC dataset to extract the optimal features and feed these features to the RNN model. Compared with the regular classifier models, the optimized deep RNN achieved the best performance. Christo et al. [39] proposed a framework based on feature selection methods and RF-ML techniques to classify the tumors in the WDBC dataset. The study applied this technique to various datasets. The accuracy of the proposed framework on the WDBC dataset was 97.1%. Ogundokun et al. [40] utilized hyperparameter optimization techniques to enhance the performance of ANN and CNN. They also conducted a comparative analysis of their performance against SVM and MLP. Notably, the ANN achieved an impressive accuracy of 99.2% when evaluated using the WDBC dataset.

According to previous studies, using ML in BC prediction offers significant potential for improving the accuracy and speed of diagnosis, personalizing treatment, and identifying

high-risk patients. However, there are also significant challenges associated with the use of ML, including the need for high-quality data and more flexible models with optimal parameters to achieve high efficiency and avoid the overfitting problems. Addressing these challenges will be essential in realizing the full potential of ML for BC prediction. Table 1 provides a summary of the review of the literature.

Table 1. Summary of the review of the literature

Author	Year	Dataset	Methods / Techniques	Result
Saleh et al. [1]	2022	WDBC	Optimized deep RNN with Keras–Tuner	96.74 accuracy
Bhardwaj et al. [2]	2022	WDBC	MLP, KNN, and RF	RF = 96.24% accuracy
Rasool et al. [38]	2022	WDBC	Optimized SVM with bayesian hyperparameter optimization	99.3% accuracy
Samieinasab et al. [37]	2022	WDBC	Extra Tree algorithm is used for features selection with Bagging, Boosting, and Voting	0.982 accuracy
Hemavathi et al. [35]	2022	WDBC	deep learning	RF = 98.6% accuracy rate
Monirujjaman Khan et al. [36]	2022	WDBC	RF, LR, DT, and KNN	LR = 98% accuracy
Christo et al. [39]	2022	WDBC	RF-ML with Feature selection	97.1% accuracy
Ogundokun et al. [40]	2022	WDBC	Optimized ANN and CNN with hyperparameter optimization, SVM, and MLP. PSO feature selection applied to WDBC	ANN = 99.2% accuracy
Assegie et al. [34]	2021	WDBC	Adboost and DT	92.53% accuracy
Gopal et al. [33]	2021	WDBC	MLP, RF, and logistic regression,	MLP = 98% accuracy
Ara et al. [32]	2021	WDBC	SVM, LR, KNN, DT, Naive Bayes, and RF	SVM and RF = 96.5% accuracy
Lahoura et al. [41]	2021	WDBC	Cloud-based ELM	98.68% accuracy
Sheth et al. [30]	2020	WDBC	Optimized FSTBSVM with Jaya optimization techniques	94.36%. accuracy
Showrov et al. [29]	2019	WDBC	SVM with 10 selected features	96.72% accuracy
Latchoumi et al. [28]	2019	WDBC	WQPSO with smooth SVM	98.42% accuracy
Jeyasingh et al. [26]	2017	WDBC	RF ML model with features chosen by MB algorithm	96% accuracy
Aalaei et al. [25]	2016	WDBC	ANN model with features chosen by GA algorithm	97.3 accuracy

III. METHODOLOGY

In this section, the suggested framework will be covered, along with study questions.

The study aims to answer the following questions:

- Do the data exploration techniques help researchers build accurate predictive models for detecting BC?
- Does hyperparameter tuning play a critical role in detecting BC accurately?

To answer these questions, a framework is proposed, as demonstrated in Figure 1. The framework has ten significant steps outlined as follows:

1. Acquiring the WDBC datasets from the Kaggle repository.
2. Describing the dataset and distribute features to study the uniqueness of features.
3. Studying the feature relationships by calculating the Pearson correlation coefficient (r) and drawing the correlation heat map.
4. Studying the features and determine the importance of the features using the correlation and chi-squared test.
5. Cleaning dataset (remove missing values and duplicate instances).
6. Balancing the target class using the Synthetic Minority Oversampling Technique (SMOTE).
7. Splitting the dataset into a training dataset and a testing dataset.
8. Applying the Bayesian hyperparameter tuning algorithm to the implementation of the dynamic training model with five predictive ensemble models (AdaBoost, RUSBoost, LogitBoost, GentleBoos, and Bag) on the datasets; evaluate and select the best model.
9. Training the regular ML models (fine tree, logistic regression (LR), medium gaussian SVM (MGSVM), and fine KNN).
10. Using the accuracy, precision, recall, and F1-Score matrices to evaluate the performance of the proposed dynamic ensemble model and compare it with the regular ML model; the evaluation considered two conditions: with or without feature selection.

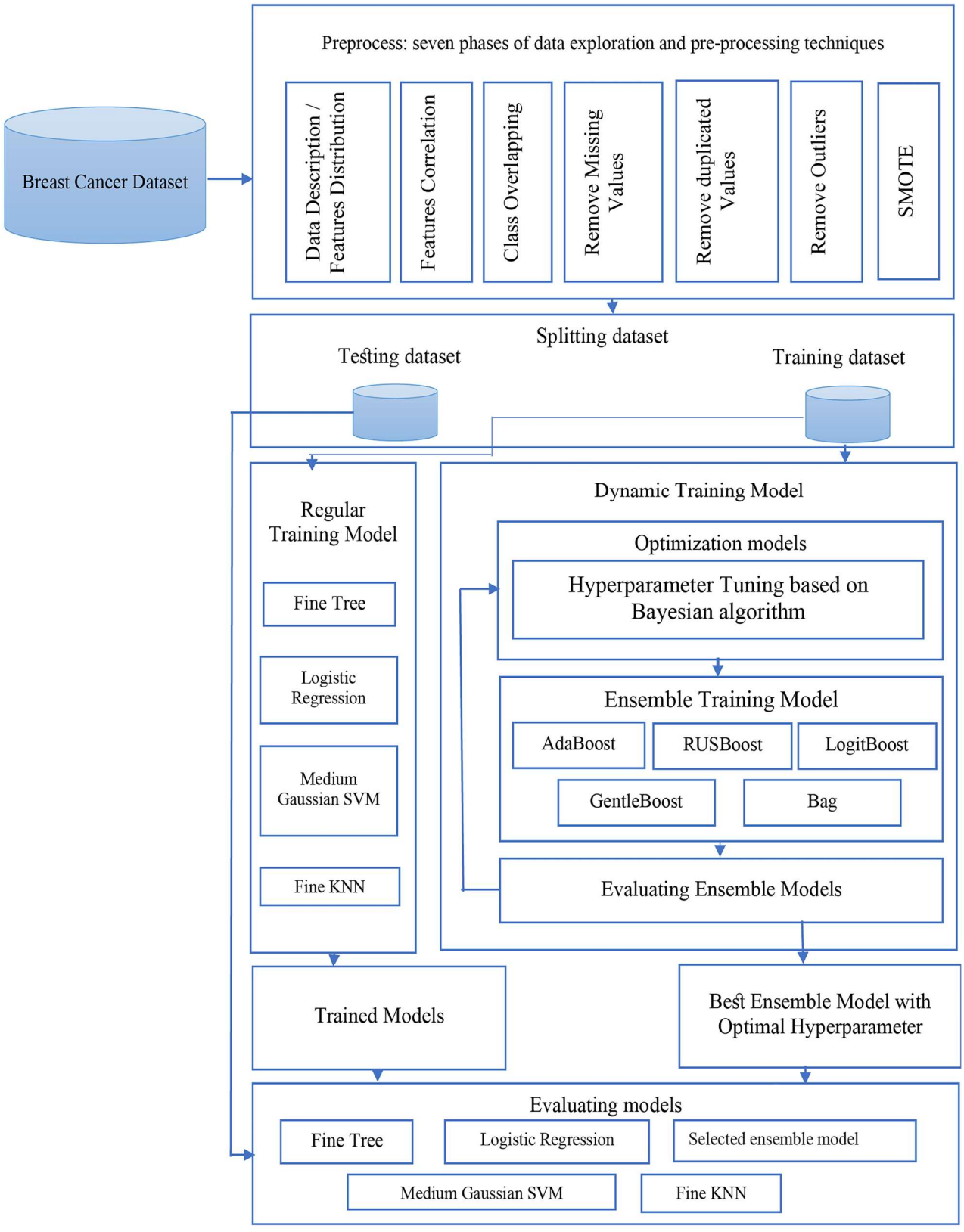

Figure 1. A proposed framework for breast cancer prediction.

IV. DATA EXPLORATION

A. DATASET DESCRIPTION:

The experiments in this study used the WDBC Dataset. The dataset can be found in the Kaggle Repository [42]. WDBC contains 569 instances, distributed as 357 benign and 212 malignant, with no missing values or duplicated instances. The dataset consists of 31 features, comprising one categorical feature (output class) and 30 numerical features. The statistical information for the numerical features (F2-F31) is presented in Table 2.

Figure 2 shows a colormap visualization technique to represent the relationship between the features and the distribution of data points across a range of values. It is evident

that there is an unequal distribution of the target class and data overlap in certain features, such as F6, F10, F11, F13, F16, F17, F18, F21, F23, F26, F30, and F31. This overview led us to conduct a more thorough analysis in the following sections.

Table 2. Wisconsin Diagnostic Breast Cancer Dataset Features

No	Feature	Missing value	Min	Median	Max
F2	Radius mean	None	6.981	13.37	28.11
F3	Texture mean	None	9.710	18.84	39.28
F4	Perimeter mean	None	43.79	86.24	188.5
F5	Area mean	None	143.50	551.1	2501
F6	Smoothness mean	None	0.05263	0.09587	0.1634
F7	Compactness mean	None	0.01938	0.09263	0.3454
F8	Concavity mean	None	0	0.06154	0.4268
F9	Concave points mean	None	0	0.0335	0.2012
F10	Symmetry mean	None	0.106	0.1792	0.304
F11	Fractal dim. Mean	None	0.04996	0.06154	0.09744
F12	Radius SE	None	0.1115	0.3242	2.873
F13	Texture SE	None	0.3602	1.108	4.885
F14	Perimeter SE	None	0.757	2.287	21.98
F15	Area SE	None	6.802	24.53	542.2
F16	Smoothness SE	None	0.001713	0.00638	0.03113
F17	Compactness SE	None	0.002252	0.02045	0.1354
F18	Concavity SE	None	0	0.02589	0.396
F19	Concave points SE	None	0	0.01093	0.05279
F20	Symmetry SE	None	0.007882	0.01873	0.07895
F21	Fractal dim. SE	None	0.0008948	0.003187	0.02984
F22	Radius worst	None	7.93	14.97	36.04
F23	Texture worst	None	12.02	25.41	49.54
F24	Perimeter worst	None	50.41	97.66	251.2
F25	Area worst	None	185.2	686.5	4254
F26	Smoothness worst	None	0.07117	0.1313	0.2226
F27	Compactness worst	None	0.02729	0.2119	1.058
F28	Concavity worst	None	0	0.2267	1.252
F29	Concave points worst	None	0	0.09993	0.291
F30	Symmetry worst	None	0.1565	0.2822	0.6638
F31	Fractal dim. Worst	None	0.05504	0.08004	0.2075

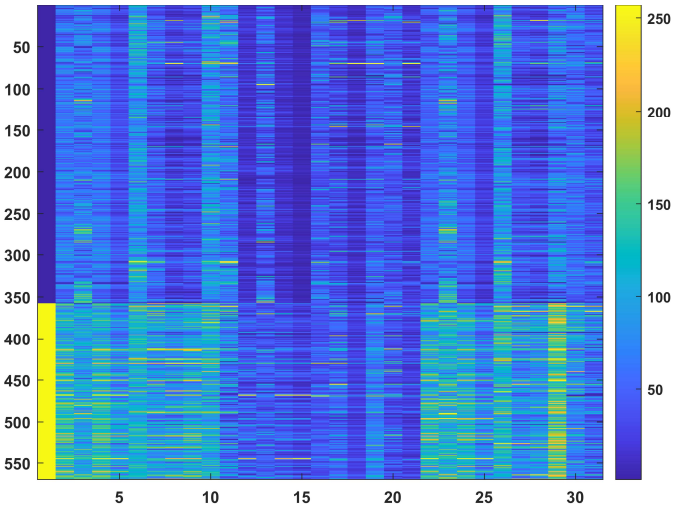

Figure 2. Data Visualization using Colormap.

B. CHECKING THE CORRELATION BETWEEN FEATURES:

Investigations into feature correlation are important because they aid in identifying relationships between different dataset features. By understanding the correlation between features, data scientists can better understand the underlying structure of the data and make more informed decisions about which features to include in their models. Correlation investigations can also help identify potential sources of bias or outliers in a dataset, which can be useful for improving the accuracy and reliability of predictive models. Figure 3 shows the correlation between the features, whereas Table 3 shows the features selected to compare the performance of the proposed framework with and without feature selection.

C. CHECKING THE OVERLAP BETWEEN CLASSES:

The effects of overlapping target classes on ML training can be quite severe. If an algorithm cannot accurately distinguish between two classes due to overlapping features, it will not be able to learn how to correctly classify new data points into either class. This can lead to inaccurate predictions and poor performance on unseen data points.

The target class overlap is depicted in Figures 4 and 5. In particular, the smoothness mean, fractal dimension mean, texture se, smoothness se, concave points se, and fractal dimension se features overlap.

D. CHECKING THE IMBALANCE PROBLEM:

Data distribution is crucial for disease prediction. Figure 6 shows the target class (malignant, benign) distribution, where 37.25% of instances are malignant and 62.75% are benign. In this case, the imbalance problem needs to be handled; otherwise, the variance in the number of target classes may cause bias when selecting the training set randomly.

Table 3. Feature selection using correlation, and chi-squared test

No	Feature	Correlation >= 0.6 **(See Figure 3)**	Importance using chi2		Selected Features
			Value	Ranking	
F2	Radius mean	*F3, F4, F8, F9, F12, F14, F15, F22, F,24, F25, F29*	174.1708	7	Selected
F3	Texture mean	*F23*	68.1022	18	Non-Selected
F4	Perimeter mean	*F2, F5, F7, F8, F9, F12, F14, F15, F22, F24, F25, F28, F29*	170.6867	8	Selected
F5	Area mean	*F2, F4, F8, F9, F12, F14, F15, F22, F24, F25, F29*	176.6110	6	Selected
F6	Smoothness mean	*F7, F9, F10, F11, F26*	41.2315	23	Non-Selected
F7	Compactness mean	*F4, F6, F8, F9, F10, F11, F17, F18, F19, F24, F27, F28, F29, F31*	96.4315	15	Selected
F8	Concavity mean	*F2, F4, F5, F7, F9, F12, F14, F15, F17, F18, F19, F22, F24, F25, F27, F28, F29*	159.1932	9	Selected
F9	Concave points mean	*F2, F4, F5, F6, F7, F8, F12, F14, F15, F19, F22, F24, F25, F27, F28, F29*	192.8852	5	Selected
F10	Symmetry mean	*F6, F7, F30*	31.3718	25	Non-Selected
F11	Fractal dim. Mean	*F6, F7, F17, F21, F31*	8.3221	28	Non-Selected
F12	Radius SE	*F2, F4, F5, F8, F9, F14, F15, F22, F24, F25*	177.0891	12	Selected
F13	Texture SE	-	9.8869	26	Non-Selected
F14	Perimeter SE	*F2, F4, F5, F8, F9, F12, F15, F19, F22, F24, F25, F29*	107.6787	13	Selected
F15	Area SE	*F2, F4, F5, F8, F9, F12, F14, F22, F24, F25*	155.2731	11	Selected
F16	Smoothness SE	-	1.5256	30	Non-Selected
F17	Compactness SE	*F7, F8, F11, F18, F19, F21, F27, F28, F31*	43.9593	22	Non-Selected
F18	Concavity SE	*F7, F8, F17, F19, F21, F28*	73.1931	17	Non-Selected
F19	Concave points SE	*F7, F8, F9, F14, F17, F18, F21, F29*	63.9390	19	Non-Selected
F20	Symmetry SE	-	1.6977	29	Non-Selected
F21	Fractal dim. SE	*F11, F17, F18, F19, F31*	8.7185	27	Non-Selected
F22	Radius worst	*F2, F4, F5, F8, F9, F12, F14, F15, F24, F25, F28, F29*	199.7237	2	Selected
F23	Texture worst	*F3*	75.3487	16	Non-Selected
F24	Perimeter worst	*F2, F4, F5, F7, F8, F9, F12, F14, F15, F22, F25, F28, F29*	206.9704	1	Selected
F25	Area worst	*F2, F4, F5, F8, F9, F12, F14, F15, F22, F24, F29*	197.7166	3	Selected
F26	Smoothness worst	*F6, F7, F27, F31*	54.9273	20	Non-Selected
F27	Compactness worst	*F7, F8, F9, F17, F26, F27, F29, F30, F31*	99.0969	14	Selected
F28	Concavity worst	*F4, F7, F8, F9, F17, F18, F27, F29, F31*	155.4776	19	Selected
F29	Concave points worst	*F2, F4, F5, F7, F8, F9, F14, F19, F22, F24, F25, F27, F28*	195.7161	4	Selected
F30	Symmetry worst	*F10, F27*	45.5458	21	Non-Selected
F31	Fractal dim. Worst	*F7, F11, F21, F26, F27, F28*	32.9465	24	Non-Selected

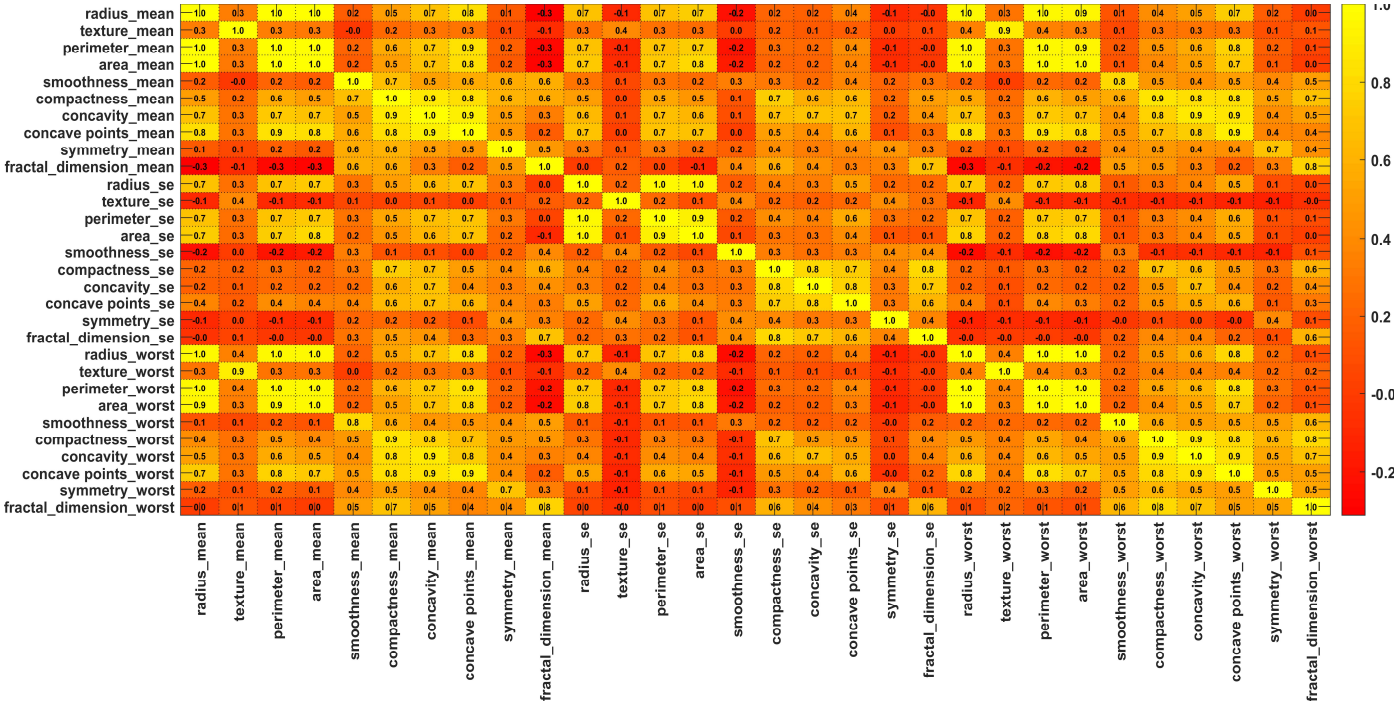

Figure 3. Features Correlation Heatmap.

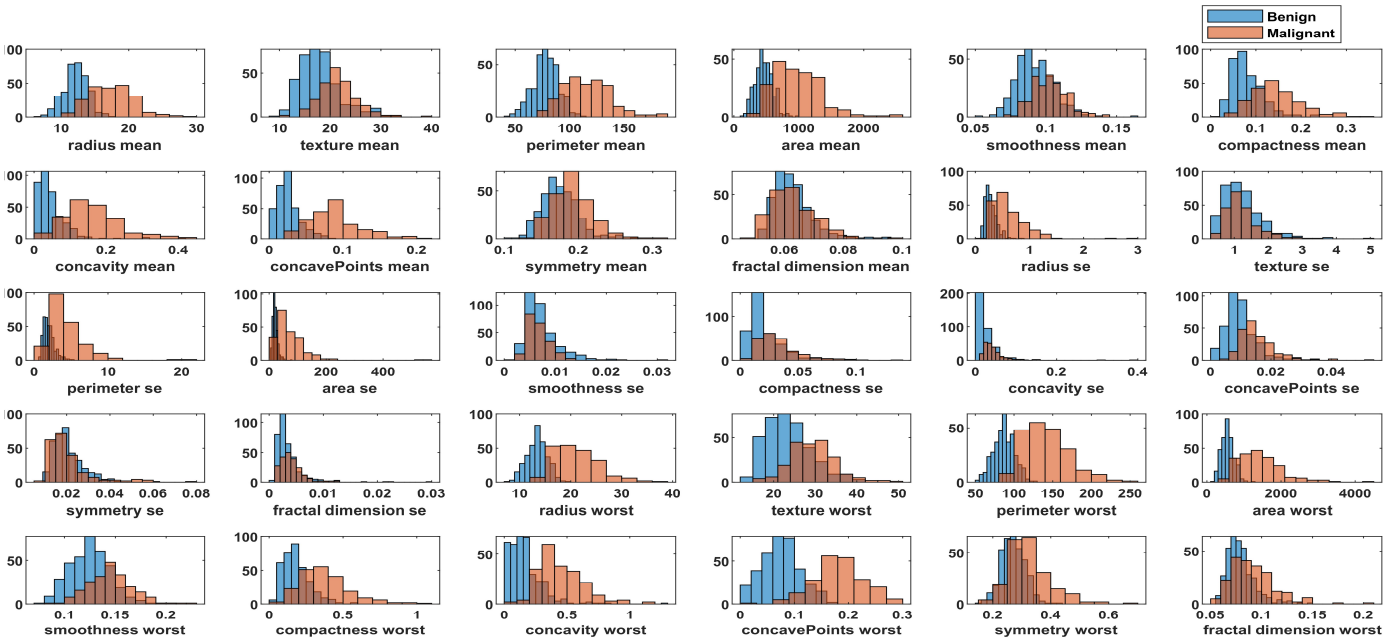

Figure 4. Features Distribution.

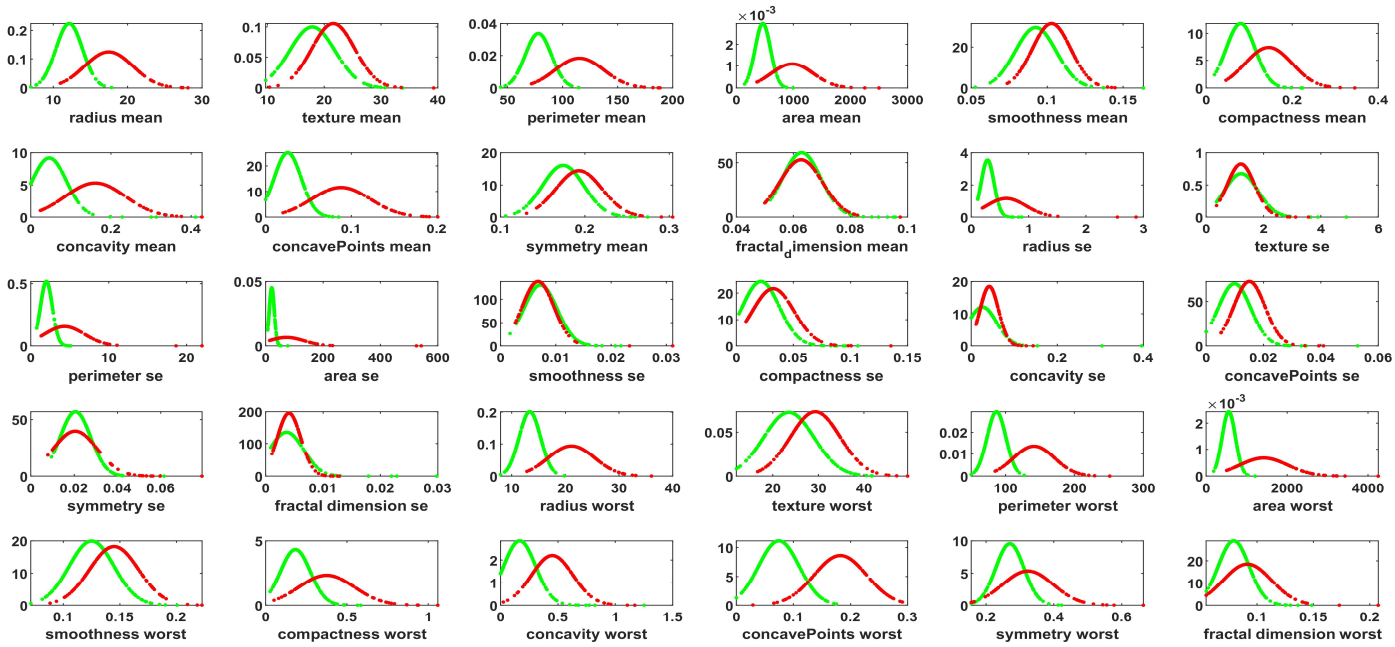

Figure 5. Classes Normal Distribution of Features.

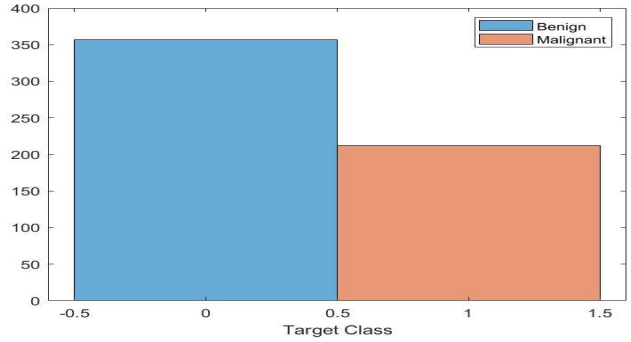

Figure 6. Target Class Distribution.

E. CHECKING THE OUTLIERS VALUE:

Outliers are data points far away from the majority of the other data points in a dataset. Outliers can be caused by errors in data collection, or they may be legitimate observations that are simply rare. In ML, outliers can have a significant impact on the accuracy of models. Outliers can significantly bias the distribution mean and standard deviation. When we add a large outlier, the mean is more than doubled, and the standard deviation is over ten times larger. However, the median and interquartile range have not changed much.

Figure 7 shows three methods to identify outliers: three standard deviations (3-SD) above and below the mean; 1.5 times the interquartile range (IQR) above or below the third and first quartiles; and third, three scaled median absolute

deviations above and below the median. The three methods are represented by different colors: red, blue, and black, respectively, in Figure 7.

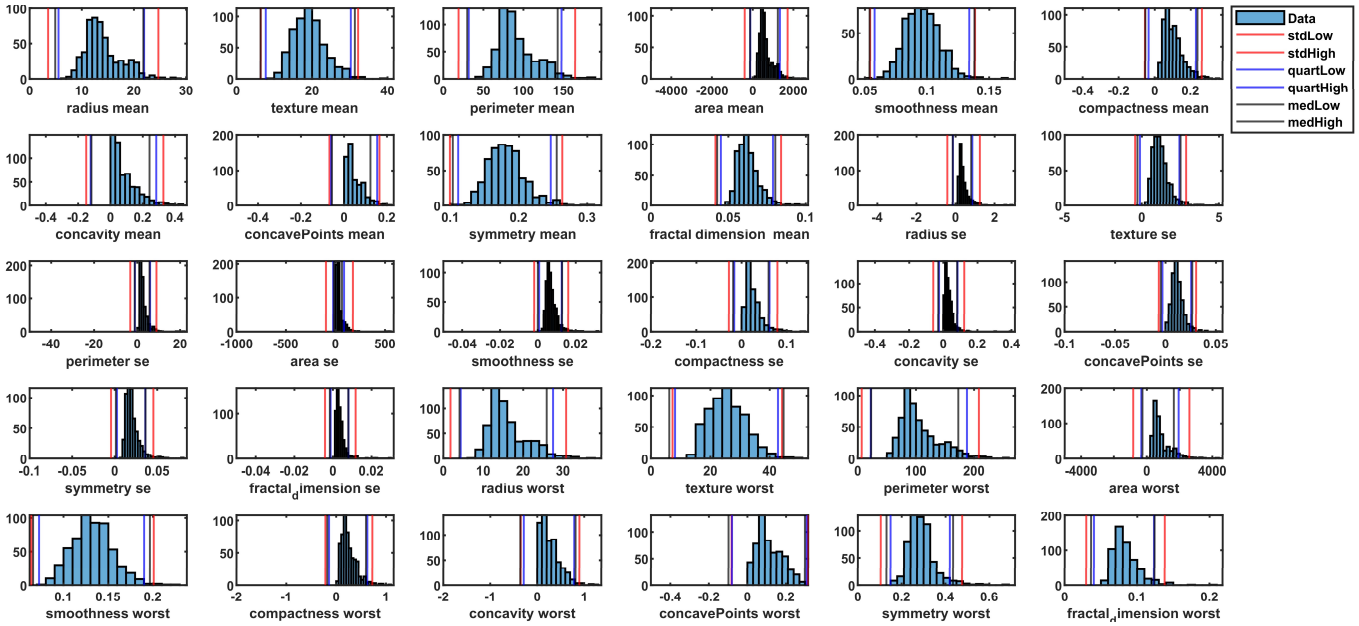

Figure 7. Identify Outliers.

V. PREPROCESSING

Raw data is often messy and should be preprocessed before being ready to be used. After the previous data analysis, a sense of what the data looks like is formed. The following are the major preprocessing steps performed by the WDBC.

A. HANDLING OUTLIERS' VALUES:

In healthcare, outliers' data is not necessarily bad; outliers are often expected on large enough sets of data due to natural variations. 3-SD (μ+3σ) and below (μ-3σ) means are useful for normally distributed data. However, due to the sensitivity of both metrics, the bounds may shift significantly with skewed data or large outliers. In contrast, both the interquartile range and median absolute deviation methods are more useful for general distributions and unaffected by extreme outliers. This study applied the 3-SD method for removing extremely large outliers from the WDBC dataset, as illustrated in Figure 7.

B. HANDLING IMBALANCE PROBLEM:

The study used SMOTE techniques to balance the target class by creating new instances for the minority class. SMOTE determines the new points by finding their collinearity with the neighboring points of the minority class.

VI. PREDICTIVE MODELS

The study used an ensemble classifier, which is a powerful tool in ML, as it combines multiple individual classifiers to create a dynamic model for accurate and reliable predictions. These classifiers include AdaBoost, LogitBoosting, GentleBoost, RUSBoost, and Bagged decision trees. It is important to tune the parameters of the individual classifiers in order to maximize their performance within the ensemble. This is done through 5-fold cross-validation and the Bayesian optimization techniques, which allow efficient exploration of parameter space. The performance of the dynamic model was compared with other state-of-the-art models, including fine trees, LR, MGSVM, and fine KNN. The comparison was conducted using accuracy, precision, recall, and the F1 score.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+F}, \quad (1)$$

$$Precision = \frac{TP}{TP+FP}, \quad (2)$$

$$Rcall = \frac{TP}{TP+FN}, \quad (3)$$

$$F1-Score = \frac{2\times Precision\times Recall}{Precision+Recall}. \quad (4)$$

where:

TP = True Positive,
FP = False Positive,
TN = True Negative,
FN = False Negative.

A. HYPERPARAMETER OPTIMIZATION:

Hyperparameter tuning is an important part of the ML process. It involves adjusting the parameters of a model to optimize its performance. This is done by testing different combinations of hyperparameters and selecting the best one. Bayesian optimization was employed in this study. It is a powerful technique for hyperparameter tuning that uses Bayesian inference to find the optimal set of hyperparameters for a given ML model. This technique has become increasingly popular in recent years due to its ability to quickly and accurately identify the best set of hyperparameters for a given problem.

B. ADABOOST DECISION TREE:

This is an ensemble ML algorithm that uses a combination of weak learners to create a strong learner. It works by combining multiple weak learners (one-level decision trees) to create a

more powerful model. Each weak learner is trained on the same dataset but with different weights assigned to each instance, and the final model is the weighted average of all the weak learners. AdaBoost is used for both classification and regression problems, and it has been found effective in improving accuracy over single models.

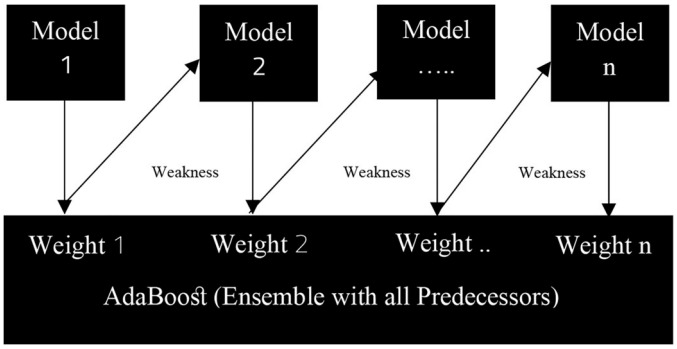

Figure 8. AdaBoost Algorithm.

C. LOGITBOOST DECISION TREE:

It is an ensemble learning technique that combines the strengths of decision trees and logistic regression to create a powerful predictive model. It uses a series of decision trees to learn from the data and then combines the results with logistic regression to produce a more accurate prediction. This technique is often used in classification problems. The convex optimization equation of LogitBoost is as follows:

$$f = \sum_t a_t h_t, \tag{5}$$

where, f represents the output of the LogitBoost ensemble model, which is the sum of the individual predictions from the decision trees (h_t) weighted by the coefficients (a_t).

The logistic loss is minimized by the equation:

$$\sum_i \log\left(1 + e^{-y_i f(x_i)}\right) \tag{6}$$

where:

- $\sum_i$: Represents the summation over all instances in the dataset.
- y_i: Denotes the true label or target value of the i-th instance.
- $f(x_i)$: Represents the predicted output or score obtained from the ensemble model for the i-th instance.
- log: Denotes the natural logarithm.
- e: Represents the base of the natural logarithm.

D. GENTLEBOOST:

An ensemble learning technique combining decision trees' powers with the boosting algorithm. It works by combining multiple weak learners (decision trees) to create a strong learner. The boosting algorithm helps reduce bias and variance, resulting in improved accuracy and better generalization performance. It is a powerful tool for classification problems, as it can handle both continuous and categorical data. The following equation may be used to compute the mean squared error:

$$\sum_i w_{t,i} \left(y_i - f_t(x_i)\right)^2, \tag{7}$$

where:

- $w_{t,i}$: The observation weights.
- $f_t(x_i)$: The prediction of f_t-regression model fitted x_i-response values.

E. DECISION TREE RUSBOOST:

This is an ensemble learning technique that combines the strengths of both decision trees and boosting algorithms. It uses a random under-sampling technique to reduce the size of the training dataset and then applies boosting to create a strong classifier from the reduced dataset. This technique has been shown to improve accuracy and reduce overfitting compared to traditional decision tree methods.

F. ENSEMBLE BAGGED DECISION TREES:

It is also an ensemble machine-learning technique that combines multiple decision trees to create a more accurate and stable prediction. It works by training multiple decision trees on different subsets of the data and then combining their predictions to form a more robust prediction. This technique can help reduce overfitting and improve accuracy.

VII. RESULTS AND DISCUSSION

To evaluate how well the proposed framework worked, three different approaches were used. The first approach used the original dataset without preprocessing to train the proposed optimized dynamic model and compare its performance with the performance of four other ML models.

The second approach involved balancing the target class distribution using the SMOTE technique, followed by training the ML models. This approach obtained encouraging performance with fine trees, logistic regression, and fine KNN. The proposed optimized ensemble model and MGSVM achieved the highest accuracy of 100%.

The third approach involved removing outliers and then balancing the target class distribution using the SMOTE technique. In comparison to earlier approaches, the findings obtained from the dynamic model of optimized ensemble classifiers, fine tree, logistic regression, fine KNN, and MGSVM models are superior. The following is the result of all the approaches with various conditions:

A. THE FIRST APPROACH (WITHOUT DOING OUTLIERS DETECTION AND SMOTE TECHNIQUES):

Table 4 shows the performance analysis of the first approach in terms of test accuracy rate (mean). The proposed optimized GentleBoost achieved a better average test accuracy of about 97.35% for the mean, 98.2% for the best, 96.5% for the worst, and 0.425 for the standard deviation.

Without feature selection, the accuracy (highest) achieved by the fine tree is 96.5%, LR is 93.8%, MGSVM is 97.3%, fine KNN is 94.7%, and the GentleBoost is 98.2%.

As shown in Table 5, feature selection reduces the performance of the ML models, where the accuracy (highest) achieved by fine tree is 91.2%, LR is 94.7%, MGSVM is 95.6%, fine KNN is 94.7%, and GentleBoost is 94.7%. The optimal GentleBoost hyperparameter is shown in Table 6.

Table 4. Performance analysis of the first approach in terms of test accuracy rate

Models	Mean	Best	Worst	SD
Fine Tree	96.5%	96.5%	96.5%	0
Logistic Regression	93.8%	93.8%	93.8%	0
MGSVM	96.45%	97.3%	95.6%	0.425
Fine KNN	92.95%	94.7%	91.2 %	0.875
Proposed	97.35**%**	98.2%	96.5%	0.425

Table 5. Performance comparison of the first approach

ML Models	Accuracy		Precision		Recall		F1-Score	
	Without	With	Without	With	Without	With	Without	With
Fine Tree	96.5%	91.2%	98%	91.5%	92.9%	90.5%	95%	90%
Logistic Regression	93.8	94.7	94.4%	95.8%	92.9%	92.9%	94%	94%
MGSVM	97.3%	95.6%	97.2	98.6%	97.6%	90.5%	97%	94%
Fine KNN	94.7%	94.7%	95.8%	94.4%	92.9%	95.2%	94%	95%
Proposed	98.2%	94.7%	100%	94.4	95.2%	95.2%	97.5%	95%

Table 6. GentleBoost optimal hyperparameter for the first approach

Bayesian hyperparameter tuning algorithm	
Hyperparameter	Value
Ensemble Method	GentleBoost
Number of learners	97
Learning rate	0.8299
Maximum number of splits	1

B. THE SECOND APPROACH (WITHOUT DOING OUTLIERS DETECTION AND DOING SMOTE TECHNIQUE):

The performance analysis of the second approach in terms of test accuracy rate is shown in Table 7.

Table 7. Performance analysis of the second approach in terms of test accuracy rate

Models	Mean	Best	Worst	SD
Fine Tree	94.75%	95.1%	94.4%	0.175
LR	96.5%	96.5%	96.5%	0
MGSVM	98.6%	100%	97.2%	0.7
Fine KNN	98.25%	98.6%	97.9%	0.175
Proposed	99.3%	100%	98.6%	0.525

The following discussion includes a thorough comparison of the best prediction model performance:

- ***Without feature selection:*** Table 8 shows the significant contribution of the SMOTE technique in improving the performance of ML models. The problem of ML models' bias toward the majority class is avoided in the training process by balancing the target class. The experiments proved that the SMOTE technique without feature selection provided significant accuracy, reaching 100% with a MGSVM and the proposed AdaBoost with Bayesian optimization algorithm. The fine tree model does not gain benefits from balancing the target class, unlike the logical regression and fine KNN models, which showed a slight improvement in performance.
- ***With feature selection***: In comparison with the first approach, the SMOTE technique provides significant results in the case of feature selection; The achieved accuracy of fine tree is 97.2%, LR is 95.8%, MGSVM is 97.9%, fine KNN is 95.1, and the optimized ensemble model is 97.9%. Table 9 displays the AdaBoost's optimal hyperparameter.

Table 8. Performance comparison of the second approach

ML Models	Accuracy		Precision		Recall		F-Score	
	Without	With	Without	With	Without	With	Without	With
Fine Tree	95.1%	97.2%	93%	97.2%	97.2%	97.2%	95%	97.2%
Logistic Regression	96.5%	95.8%	93%	97.2%	100%	94.4%	96%	96%
MGSVM	100%	97.9%	100%	97.2%	100%	98.6%	100%	98%
Fine KNN	98.6%	95.1%	96.5%	94.4%	93.7%	95.8%	95%	95%
Proposed	100%	97.9%	100%	98.6%	100%	97.2%	100%	98%

Table 9. AdaBoost optimal hyperparameter for the second approach

Bayesian hyperparameter tuning algorithm	
Hyperparameter	Value
Ensemble Method	AdaBoost
Number of learners	72
Learning rate	0.9505
Maximum number of splits	3

C. THIRD APPROACH (DOING OUTLIERS DETECTION AND DOING SMOTE TECHNIQUE):

Table 10. Performance analysis of the third approach in terms of test accuracy rate (average)

Models	Mean	Best	Worst	SD
Fine Tree	96.65%	97.3%	96%	0.324
LR	95.3%	95.3%	95.3%	0
MGSVM	98.65%	100%	97.3%	0.675
Fine KNN	96.65	99.3%	94%	1.325
Proposed	99.35%	100%	98.7%	0.325

Table 10 shows the performance analysis of the proposed model and the other four ML models in terms of test accuracy rate. The proposed optimized AdaBoost achieved a better average of test accuracy of about 99.35% for the mean, 100% for the best, 98.7% for the worst, and 0.325 for the standard deviation.

A comprehensive comparison of the best performance of predictive models is discussed below:

- ***Without feature selection:*** Among the five ML models, the ensemble model (AdaBoost) with the Bayesian hyperparameter tuning algorithm and the MGSVM received the most significant performance, with 100% accuracy. The second-best performance was by the fine KNN algorithm with 99.3% accuracy, 98.7% precision, 100% recall, and 99 % F1-score. Table 11 shows the comparison of the performance of the five models for all conditions in the third approach. Table 12 displays the AdaBoost's optimal hyperparameter.
- ***With feature selection:*** The LR model outperformed the other predictive models with an accuracy of 98.7%, a precision of 97.2%, a recall of 100%, and an F1-score of 98.5%.

Table 13 shows the evaluation of performance in comparison with similar works of literature

Table 11. Performance comparison of the third approach

ML Models	Accuracy		Precision		Recall		F1-Score	
	Without	With	Without	With	Without	With	Without	With
Fine Tree	97.3%	94.7%	97.2%	94.4%	97.5%	94.9%	97%	95%
LR	95.3%	98.7%	91.5%	97.2%	98.7%	100%	95%	98.5%
MGSVM	100%	96%	100%	94.4%	100%	97.5%	100%	96%
Fine KNN	99.3%	98%	98.6%	95.8%	100%	100%	99%	98%
Proposed	100%	97.3%	100%	97.2%	100%	97.5%	100%	97%

Table 12. AdaBoost optimal hyperparameter for the third approach

Bayesian hyperparameter tuning algorithm	
Hyperparameter	Value
Ensemble Method	AdaBoost
Number of learners	322
Learning rate	0.9350
Maximum number of splits	1

Table 13. Evaluation of performance in comparison to similar works of literature

Author	Year	Classifier	Accuracy (%)
[25]	2016	ANN with GA algorithm	97.3%
[28]	2019	Smooth SVM with WQPSO	98.42%
[30]	2020	FSTBSVM with Jaya optimization technique	94.36%
[33]	2021	MLP	98%
[34]	2021	Adboost algorithm	92.53%
[41]	2021	Cloud-based ELM	98.68%
[36]	2022	LR	98%
[35]	2022	RF	98.6%
[2]	2022	RF	96.24%
[38]	2022	Polynomial SVM with grid search optimization	99.3%
[40]	2022	Optimized ANN and CNN with hyperparameter optimization	99.2%
Proposed	2023	Dynamic learning model of ensemble classifiers, including AdaBoost, RUSBoost, LogitBoost, GentleBoost, and Bag. The model used with Bayesian hyperparameter tuning algorithm	Test accuracy rate = 99.35%

VIII. CONCLUSION AND FUTURE SCOPE

In conclusion, the optimized framework based on multi-stage data exploration and dynamic ensemble-based classifiers with Bayesian hyperparameter tuning represents a significant advancement in BC prediction. The framework provides a comprehensive approach to data exploration and preprocessing, ensuring the data is well-prepared for ML. Additionally, the framework includes dynamic ensemble-based classifiers, which combine multiple independent classifiers to improve accuracy and mitigate the risk of overfitting. These classifiers are optimized using hyperparameter tuning, which selects the optimal values for the various hyperparameters of the model, leading to more accurate BC prediction. The experimental results show that the proposed framework

outperforms other models in terms of test accuracy rate (mean), precision, recall, and F1 score. The framework also achieved significant performance in different approaches, making it a practical and effective tool for BC prediction. Using the publicly available WDBC dataset, the framework is readily accessible to other researchers and practitioners, which can further improve the accuracy and efficiency of BC prediction.

According to the information provided in Table 4, the optimized dynamic ensemble-based classifiers achieved remarkable accuracy rates when trained on the original WDBC data. The best accuracy rate obtained was 98.2%, with a mean accuracy rate of 97.35%. Furthermore, balancing the target class using SMOTE demonstrated a notable impact on improving test accuracy. As indicated in Table 7 and 10, utilizing SMOTE resulted in a significant improvement of 1.0183%, ultimately leading to a best accuracy value of 100%. This improvement is in comparison to the accuracy achieved without employing SMOTE. In addition, the optimized framework has the potential to improve the accuracy of BC prediction, ultimately leading to better outcomes for patients. The framework's comprehensive approach to data exploration, combined with dynamic ensemble-based classifiers and hyperparameter tuning, represents a significant advancement in BC prediction. The framework can be applied to other domains, providing a reliable and robust tool for ML.

This study depicted how the proposed framework performed on a relatively small numerical dataset. Trials using bigger datasets, like big data, can be used in the future to expand the scope of this study. Future work can mainly focus on improving this framework to make it applicable to all types of datasets, whether numerical or images.

References

[1] H. Saleh, H. Alyami, and W. Alosaimi, "Predicting breast cancer based on optimized deep learning approach," *Computational Intelligence and Neuroscience,* vol. 2022, article ID 1820777, 2022. https://doi.org/10.1155/2022/1820777.

[2] A. Bhardwaj, H. Bhardwaj, A. Sakalle, Z. Uddin, M. Sakalle, and W. Ibrahim, "Tree-based and machine learning algorithm analysis for breast cancer classification," *Computational Intelligence and Neuroscience,* vol. 2022, article ID 6715406, 2022. https://doi.org/10.1155/2022/6715406.

[3] A. N. Hurson, T. U. Ahearn, R. Keeman, M. Abubakar, A. Y. Jung, P. M. Kapoor*, et al.*, "Systematic literature review of risk factor associations with breast cancer subtypes in women of African, Asian, Hispanic, and European descents," *Cancer Research*, vol. 82, pp. 3670-3670, 2022. https://doi.org/10.1158/1538-7445.AM2022-3670.

[4] W. H. W. Mamat, N. Jarrett, and S. Lund, "Diagnostic interval: experiences among women with breast cancer in Malaysia," *Open Access Macedonian Journal of Medical Sciences*, vol. 9, pp. 54-59, 2022. https://doi.org/10.3889/oamjms.2021.7833.

[5] H.-J. Wu and P.-Y. Chu, "Current and developing liquid biopsy techniques for breast cancer," *Cancers*, vol. 14, p. 2052, 2022. https://doi.org/10.3390/cancers14092052.

[6] Y. S. Younis, A. H. Ali, O. K. Alhafidhb, W. B. Yahia, M. B. Alazzam, A. A. Hamad*, et al.*, "Early diagnosis of breast cancer using image processing techniques," *Journal* of *Nanomaterials*, vol. 2022, article ID 2641239, 2022. https://doi.org/10.1155/2022/2641239.

[7] K. H. Lau, A. M. Tan, and Y. Shi, "New and emerging targeted therapies for advanced breast cancer," *International Journal of Molecular Sciences*, vol. 23, p. 2288, 2022. https://doi.org/10.3390/ijms23042288.

[8] E. Cava, P. Marzullo, D. Farinelli, A. Gennari, C. Saggia, S. Riso*, et al.*, "Breast cancer diet "BCD": A review of healthy dietary patterns to prevent breast cancer recurrence and reduce mortality," *Nutrients*, vol. 14, p. 476, 2022. https://doi.org/10.3390/nu14030476.

[9] D. A. Zebari, D. A. Ibrahim, D. Q. Zeebaree, H. Haron, M. S. Salih, R. Damaševičius*, et al.*, "Systematic review of computing approaches for breast cancer detection based computer aided diagnosis using mammogram images," *Applied Artificial Intelligence*, vol. 35, pp. 2157-2203, 2021. https://doi.org/10.1080/08839514.2021.2001177.

[10] M. Madani, M. M. Behzadi, and S. Nabavi, "The role of deep learning in advancing breast cancer detection using different imaging modalities: A systematic review," *Cancers*, vol. 14, p. 5334, 2022. https://doi.org/10.3390/cancers14215334.

[11] E. Fina, "Signatures of breast cancer progression in the blood: What could be learned from circulating tumor cell transcriptomes," *Cancers*, vol. 14, p. 5668, 2022. https://doi.org/10.3390/cancers14225668.

[12] F. Cardoso, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans, I. Rubio*, et al.*, "Early breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Annals of Oncology*, vol. 30, pp. 1194-1220, 2019. https://doi.org/10.1093/annonc/mdz173.

[13] M. A. Naji, S. El 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. https://doi.org/10.1016/j.procs.2021.07.062.

[14] R. Krithiga and P. Geetha, "Breast cancer detection, segmentation and classification on histopathology images analysis: a systematic review," *Archives of Computational Methods in Engineering*, vol. 28, pp. 2607-2619, 2021. https://doi.org/10.1007/s11831-020-09470-w.

[15] S. Bacha and O. Taouali, "A novel machine learning approach for breast cancer diagnosis," *Measurement*, vol. 187, p. 110233, 2022. https://doi.org/10.1016/j.measurement.2021.110233.

[16] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access,* vol. 5, pp. 8869-8879, 2017. https://doi.org/10.1109/ACCESS.2017.2694446.

[17] M. M. Beno, I. R. Valarmathi, S. M. Swamy, and B. Rajakumar, "Threshold prediction for segmenting tumour from brain MRI scans," International Journal of Imaging Systems and Technology, vol. 24, pp. 129-137, 2014. https://doi.org/10.1002/ima.22087.

[18] S. Sengupta and A. K. Das, "Particle swarm optimization based incremental classifier design for rice disease prediction," *Computers and Electronics in Agriculture*, vol. 140, pp. 443-451, 2017. https://doi.org/10.1016/j.compag.2017.06.024.

[19] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," *Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI'2018)*, 2019, vol. 26, pp. 758-763. https://doi.org/10.1007/978-3-030-03146-6_86.

[20] N. Liu and H. Wang, "Ensemble based extreme learning machine," *IEEE Signal Processing Letters*, vol. 17, pp. 754-757, 2010. https://doi.org/10.1109/LSP.2010.2053356.

[21] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, p. 022022, 2019. https://doi.org/10.1088/1742-6596/1168/2/022022.

[22] B. Ghojogh and M. Crowley, "The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial," *arXiv preprint arXiv:1905.12787,* 2019.

[23] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295-316, 2020. https://doi.org/10.1016/j.neucom.2020.07.061.

[24] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, pp. 570-577, 1995. https://doi.org/10.1287/opre.43.4.570.

[25] S. Aalaei, H. Shahraki, A. Rowhanimanesh, and S. Eslami, "Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets," *Iranian Journal of Basic Medical Sciences*, vol. 19, p. 476, 2016.

[26] S. Jeyasingh and M. Veluchamy, "Modified bat algorithm for feature selection with the wisconsin diagnosis breast cancer (WDBC) dataset," *Asian Pacific Journal of Cancer Prevention: APJCP*, vol. 18, p. 1257, 2017.

[27] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, p. 13, 2018. https://doi.org/10.3390/designs2020013.

[28] T. Latchoumi, T. Ezhilarasi, and K. Balamurugan, "Bio-inspired weighed quantum particle swarm optimization and smooth support vector machine ensembles for identification of abnormalities in medical data," *SN Applied Sciences*, vol. 1, pp. 1-10, 2019. https://doi.org/10.1007/s42452-019-1179-8.

[29] M. I. H. Showrov, M. T. Islam, M. D. Hossain, and M. S. Ahmed, "Performance comparison of three classifiers for the classification of breast cancer dataset," *Proceedings of the 2019 4th International Conference on Electrical Information and Communication Technology (EICT'2019)*, 2019, pp. 1-5. https://doi.org/10.1109/EICT48899.2019.9068816.

[30] P. D. Sheth, S. T. Patil, and M. L. Dhore, "Evolutionary computing for clinical dataset classification using a novel feature selection algorithm," *Journal of King Saud University – Computer and Information Sciences,* vol. 8, pp. 5075-5082, 2020. https://doi.org/10.1016/j.jksuci.2020.12.012.

[31] G. Chugh, S. Kumar, and N. Singh, "Survey on machine learning and deep learning applications in breast cancer diagnosis," *Cognitive Computation*, vol. 13, pp. 1451-1470, 2021. https://doi.org/10.1007/s12559-020-09813-6.

[32] S. Ara, A. Das, and A. Dey, "Malignant and benign breast cancer classification using machine learning algorithms," *Proceedings of the 2021 International Conference on Artificial Intelligence (ICAI)*, 2021, pp. 97-101, 2021. https://doi.org/10.1109/ICAI52203.2021.9445249.

[33] V. N. Gopal, F. Al-Turjman, R. Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using IoT and machine learning," *Measurement*, vol. 178, p. 109442, 2021. https://doi.org/10.1016/j.measurement.2021.109442.

[34] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence*, vol. 10, p. 184, 2021. https://doi.org/10.11591/ijai.v10.i1.pp184-190.

[35] N. Hemavathi, R. Sriranjani, P. Arulmozhi, M. Meenalochani, and R. Deepak, "Deep learning based early prediction scheme for breast cancer," *Wireless Personal Communications*, vol. 122, pp. 931-946, 2022. https://doi.org/10.1007/s11277-021-08933-y.

[36] M. Monirujjaman Khan, S. Islam, S. Sarkar, F. I. Ayaz, M. K. Ananda, T. Tazin*, et al.*, "Machine learning based comparative analysis for breast cancer prediction," *Journal of Healthcare Engineering*, vol. 2022, article ID 4365855, 2022. https://doi.org/10.1155/2022/4365855.

[37] M. Samieinasab, S. A. Torabzadeh, A. Behnam, A. Aghsami, and F. Jolai, "Meta-health stack: A new approach for breast cancer prediction," *Healthcare Analytics*, vol. 2, p. 100010, 2022. https://doi.org/10.1016/j.health.2021.100010.

[38] A. Rasool, C. Bunterngchit, L. Tiejian, M. R. Islam, Q. Qu, and Q. Jiang, "Improved machine learning-based predictive models for breast cancer diagnosis," *International Journal of Environmental Research and Public Health*, vol. 19, p. 3211, 2022. https://doi.org/10.3390/ijerph19063211.

[39] V. E. Christo, H. K. Nehemiah, J. Brighty, and A. Kannan, "Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest," *IETE Journal of Research*, vol. 68, pp. 2508-2521, 2022. https://doi.org/10.1080/03772063.2020.1713917.

[40] R. O. Ogundokun, S. Misra, M. Douglas, R. Damaševičius, and R. Maskeliūnas, "Medical internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks," *Future Internet*, vol. 14, p. 153, 2022. https://doi.org/10.3390/fi14050153.

[41] V. Lahoura, H. Singh, A. Aggarwal, B. Sharma, M. A. Mohammed, R. Damaševičius, *et al.*, "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, p. 241, 2021. https://doi.org/10.3390/diagnostics11020241.

[42] K. Papel, "Breast Cancer Wisconsin (Diagnostic) Dataset," [Online]. Available at: https://www.kaggle.com/code/karan1210/breast-cancer/data, 1995.

***AYMAN ALSABRY** received his BSc degree in Computer Science from the faculty of science at Taiz University, Yemen. He completed his master's degree in information technology from Taiz University in 2020. He is currently a PhD student at Sana'a University. He is currently working as Academic Affairs Director at the International University of Technology Twintech, Sana'a, Yemen. His research interests include computer science, artificial intelligence, and machine learning.*

***HAMZAH ALI ABDULRAHMAN QASEM** received his Bachelor of Technology in Information Technology from Uttar Pradesh Technical University, India in 2006 and his Master of Technology in Information Science and Engineering from Visvesvaraya Technological University, India in 2010. He has obtained his Ph.D. from Department of Computer Engineering, Aligarh Muslim University, India. He is now an assistant professor and the faculty dean of Computer Science and Information Technology at the International University of Technology Twintech, Sana'a, Yemen. His current research interests include Data Science, artificial intelligence, machine learning, deep learning, and localization in wireless sensor networks*

***MALEK ALGABRI** received his BSc and MSc degrees in Computer Science and Technology from Wuhan University of Technology, Wuhan, China. He completed his PhD at the same university in Wuhan, China, in December 2013. He is currently working as an Associate Professor of Computer Science and Technology in the Faculty of Computing and Information Technology at Sana'a University, Sana'a, Yemen. His research interests include computer science, MANETs, MPLS, artificial intelligence, and machine learning.*

***AMIN MOHAMED AHSAN** received his B.S. degree in Computer Science and Information Systems from the University of Technology, Baghdad, Iraq, in 1999 and his M.S. and Ph.D. degrees in Computer Science and Information Systems from Universiti Teknologi Malaysia, Malaysia, in 2010 and 2015, respectively. He is now an assistant professor and Head of Department of Computer Science in the faculty of Computer Science and Information Technology at the International University of Technology Twintech, Sana'a, Yemen. His research interests include computer vision and image processing, pattern recognition, artificial intelligence, machine learning, and deep learning.*

***Assoc. Prof. Dr. MOGEEB A.A. Mosleh** is the Vice President of IUTT for academic affairs and senior lecturer at Software Engineering Dep. Faculty of Engineering and Information Technology- Taiz University. He obtained his Ph.D. and MSc in Artificial Intelligent area at FCSIT – University of Malaya.*

***FOAD HANASH** is currently the CEO of International University of Technology Twintech, Secretary General of the Emirates International University, and senior lecturer of Experimental Physics, Faculty of Science, Saada University. He received his B.Sc. degree in physics from Amran University in 2008. received an M.Sc. degree in experimental physics from the Faculty of Science at Mansoura University, Egypt, in 2012 and a PhD in experimental physics from the same university. His research interests are experimental physics, optics, interference, lasers, nanotechnology, and machine learning.*