# ENHANCED BREAST CANCER PREDICTION USING ADASYN AND OPTIMIZED LIGHTGBM: A STUDY ON THE BREAST CANCER COIMBRA DATASET

Paul Olujide ADEBAYO[1], Rasheed Gbenga JIMOH[2],
Babatunde Waheed YAHYA[3], Joseph Bamidele AWOTUNDE[4],
Oluwafisayo Babatope AYOADE[5]
[1,2,4]Department of Computer Science, University of Ilorin, Nigeria
[3]Department of Statistics, University of Ilorin, Nigeria
[5]School of Computer, Data and Mathematical Sciences, Western Sydney University, Australia
[1*]adebayo.po@unilorin.edu.ng, [2]jimoh_rasheed@unilorin.edu.ng, [3]wbyahya@unilorin.edu.ng,
[4]awotunde.jb@unilorin.edu.ng, [5]22053430@student.westernsydney.edu.au

Keywords: Breast Cancer Prediction, ADASYN, LightGBM, Class Imbalance, Hyperparameter Optimization, Grid Search Cross-Validation

Abstract: *Breast cancer poses a significant challenge, necessitating robust predictive models for early diagnosis and treatment planning. This study aims to enhance a machine learning model using Adaptive Synthetic Sampling (ADASYN) to address class imbalance in the Breast Cancer Coimbra Dataset (BCCD). ADASYN generates synthetic samples for the minority class, balancing the dataset and improving model sensitivity to malignant cases. We employed the Light Gradient Boosting Machine (LightGBM) model, known for its high efficiency, and Hyperparameter tuning using Grid Search with cross-validation optimized LightGBM. Our findings show a substantial improvement in performance, with the optimized model significantly outperforming existing approaches. The model achieved higher accuracy, precision, recall, and F1-score, demonstrating the effectiveness of addressing class imbalance and hyperparameter optimization. Enhanced predictive accuracy can lead to earlier detection and more precise treatment planning, ultimately improving patient outcomes. Future research will explore advancements in ensemble methods and deep learning architectures.*

## 1. INTRODUCTION

Breast cancer is among the most widespread and potentially life-threatening illnesses that impact women globally [1]. World Health Organization (WHO) emphasizes that breast cancer is typically identified through routine screenings or when observable symptoms like breast lumps or changes manifest [2]–[4]. Over time, several studies reveal that the variation in mortality rates is influenced by factors such as early detection and access to treatment [5]. There has been an improvement in survival rates, particularly for early-stage cases. In contrast, advanced-stage cases exhibit lower survival rates [6]. Attempts to discover the disease at an early stage made the Global Breast Cancer Initiative advocate that the recommendation approach involves continuous screening, early detection, and comprehensive breast cancer management [7]. The main goal is to reduce breast cancer mortality rates on a global scale. Given the potential for survival through early detection, maintaining a continuous screening routine is crucial in achieving this objective.

Early disease detection can be accomplished by creating a predictive model, ensuring improved patient treatment. Notably, prior research has successfully harnessed machine learning models to detect breast cancer, demonstrating notable performance [8]–[12]. Additionally, lightgBM, known as Light Gradient Boost Machine, stands out in terms of efficient performance. It is claimed for its speed, memory efficiency, support for distributed computing, and GPU acceleration. Widely employed in various ML tasks, LightGBM is suited for highly dimensional feature spaces. It has exhibited exceptional performance in breast cancer detection compared to traditional models [13], [14]. Moreover, earlier studies highlighted the advantages of employing hyperparameter optimization to

achieve the best possible performance from machine learning models, especially hyperparameter tuning on highly imbalanced big data [15]. Although [13] presented a Tree-Structured Parzen Estimator for hyperparameter optimization of a predictive model, the authors also addressed the imbalance dataset with a synthetic minority oversampling technique (SMOTE), thereby enhancing the performance of the LightGBM model. This approach optimizes resource usage, reduces manual effort, and adapts to specific data and problem characteristics. However, despite the model's effectiveness on a widely used Wisconsin diagnostic breast cancer (WDBC) dataset regarding accuracy, the authors were silent on their model performance on BCCD.

However, no preceding research conducted hyperparameter tuning in experiments involving highly imbalanced breast cancer datasets without implementing feature selection or employing a technique like SMOTE to address extreme class imbalances in the BCC dataset. In our study's findings (refer to Fig 3), all features in the dataset prove significant and warrant consideration for cancer prediction. Therefore, this study adopts adaptive synthetic sampling (ADASYN) to handle the highly imbalanced class within the dataset of Breast cancer Coimbra.

Furthermore, considering the significance of low-dimensional space features in predicting cancer within the BCCD, we utilized a grid search algorithm for hyperparameter optimization to enhance the performance of the lightGBM classifier. Grid search proves feasible for exploring a comprehensive set of hyperparameter combinations, especially in low-dimensional data [16], [17]. This approach allows for exploring predefined hyperparameter values to detect the optimal combination that yields the best model performance.

Therefore, the research prioritized hyperparameter optimization of the ML model rather than relying on an ensemble or a classifier with default parameter settings. The key contributions of this study encompass the following:

1. Addressing class imbalance by implement Adaptive Synthetic Sampling (ADASYN) to effectively balance the Breast Cancer Coimbra Dataset (BCCD), improving model sensitivity to the minority class.
2. Utilized Light Gradient Boosting Machine (LightGBM) and conducted comprehensive hyperparameter tuning using Grid Search with cross-validation (GridSearcheCV) to enhance model performance.
3. Developed a robust predictive model that enhances early breast cancer detection and precise treatment planning, potentially leading to better patient outcomes.
4. Provided insights into the practical application of advanced machine learning techniques for breast cancer prediction, emphasizing their impact on early diagnosis and effective treatment strategies.

The learning is organized into separate sections, commencing with Section 2, which comprehensively reviews relevant existing research in the domain of discourse. Section 3 explores the practical details of the anticipated methods to actualize the study's aim. Section 4 discusses the investigational performance of the model and further compares the proposed model result with existing work. Section 5 presents the conclusion of the findings and provides insights for future work.

## 2. RELATED WORK

Breast cancer is among the primary contributors to worldwide mortality. Beyond conventional cancer detection approaches, contemporary technologies empower experts with various adaptive methods to identify and diagnose BC in women. In conjunction with emerging technologies, diverse data science methods aid in collecting and analyzing cancer-related data, contributing to predicting this life-threatening disease. ML algorithms have effectively analyzed cancer-related data among various data science technologies. As an illustration, a study recently aimed to demonstrate the enhancement of classification accuracy through machine learning algorithms [18]. The study tests the model performance using the Breast Cancer Coimbra and Banknote Authentication datasets. Results show that the BPNN-ZMP approach outperforms the novel backpropagation neural network, achieving 12.12% and 11.46% improvements for the Breast Cancer Coimbra and Banknote Authentication datasets.

In recent decades, there has been a gradual rise in the utilization of machine learning applications within the medical domain. Nevertheless, the crucial elements for diagnosis involve the data gathered from patients and the assessment conducted by medical professionals. ML classifiers have played a role in reducing human faults and providing swift, in-depth analysis of medical data [19]. Various ML classifiers are available for data modelling and

forecasting. In our research, we employed Grid Search cross-validation to identify optimal hyperparameters for the machine learning model LightGBM. Additionally, incorporating Adaptive Synthetic Sampling addresses the biases notably associated with class imbalance dataset.

In the previous study, a variety of machine learning procedures, including Decision Trees, K-Nearest Neighbors, Genetic Algorithms, Support Vector Machines, Ensemble Methods, Deep Learning, Hybrid Models, Particle Swarm Optimization, and Novel Genetic Algorithm-based Deep Neural Networks, to predict breast cancer. [20] mention that the ensemble of the Decision Tree and K-Nearest Neighbors (KNN) model demonstrates exceptional performance, achieving 100% accuracy with a train-test data ratio of 90:10 per cent. However, KNN model classification accuracy is 87.5% when the training and testing datasets are 80:20. Both [20] and [21] highlight the effectiveness of the K-Nearest Neighbors algorithm. However, Hasdyna et al. [21] show improved performance with the Gain Ratio method applied to KNN. Mishra et al. [22] combined the Genetic Algorithm (GA) with the Gradient Boosting Classifier. This merger was reported to outperform other classifiers regarding prediction accuracy, the area under the curve (AUC), and F-measure values. Sinan Basarslan and Kayaalp [23] apply SVM along with other algorithms for breast cancer classification and find that SVM performs well, while [24] compare six different models and indicate that SVM shows better performance metrics. Mishra et al. [22] use a combination of the Genetic Algorithm and the Gradient Boosting Classifier, signifying the potential practical pertinency of this ensemble model for breast cancer prediction. The experimental findings have helped researchers focus on some risk factors (Glucose, Leptin, Resistin, BMI, and Age) highlighted as critical attributes for automatic BC prediction, specifically on the BC Coimbra Dataset. Sinan Basarslan and Kayaalp [23] employ deep learning algorithms, including Recurrent Neural Network (RNN), and report that RNN performs best, achieving 92% accuracy on breast cancer datasets.

Similarly, Barwal et al. [25] propose a hybrid approach of K-nearest neighbors with singular value decomposition (SVD) and Grey Wolf Optimization (GWO) named SVOF-KNN for breast cancer detection. The model uses risk metrics from regular blood analysis in the BCCD, including insulin, glucose, HOMA, Leptin and resistin, as earlier discovered by [22]. The hybrid approach SVOF-KNN achieved an accuracy of 87.8% while its error rate is 0.1769 based on the selected features achieved by Grey Wolf Optimization. Despite the BCC dataset's relatively low-dimensional features, [26] propose a binary particle swarm optimization (BPSO) method for feature selection. The proposed model efficiently improves the efficacy of Clinical Decision Support Systems (CDSS) in predicting breast cancer automatically. In the experiment, out of all the models evaluated, only the Gradient Boosting Classifier (GBC) attains an average classification accuracy of 76% with all nine features in the BCCD. However, after feature selection, the classifier performance on BCCD achieved 80.3% accuracy with the GBC classifier, which is insignificant.

Previous literature reviewed indicates the challenge of low accuracy in existing methods. Rhmann [27] addresses the critical need for early cancer detection. It highlights the challenge of low accuracy in existing methods attributed to imbalanced classes in BCC datasets. The proposed solution is a new genetic algorithm-based deep neural network designed to detect prostate and breast cancers. The efficient performance of GA-DNN is equated with other techniques, including support vector machine (SVM), random forest, and deep neural network (DNN). The outcomes indicate that GA-DNN outperforms the other techniques for prostate cancer and the Breast Cancer Coimbra dataset. Notably, the optimized DNN within GA-DNN yields the best results when dealing with large datasets. The study also established the identified challenge of low accuracy.

The literature examined earlier highlights the difficulty of achieving high accuracy in current methodologies. In order to resolve the challenge of imbalanced datasets in BC prediction models, Alsabry et al. [28] utilize the SMOTE to balance the target class in BCCD. Two approaches are compared: one uses the original dataset, and the other incorporates SMOTE to address class imbalance. The comparison of thirteen models, consisting of various machine learning classifiers, indicates a significant improvement in BC performance when implemented by SMOTE. However, the classifier with the highest accuracy after the implementation of SMOTE is the Optimized LogitBoost model, achieving an accuracy rate of 88%. Although there is improvement in the accuracy of this approach, much is expected in terms of an efficient model for better accuracy.

There are notable shortcomings in the current body of research, including poor classification pinpoint accuracy for BCCD. Additional advancements in

the classification model are needed to improve the predictive accuracy, facilitating the early detection of BC, specifically its evaluation of BCCD. This study gains significance by evaluating various: (1) hyperparameter tuning techniques with a primary focus on addressing the classification problem of cancer analysis and (2) the field of imbalanced class distribution problems, particularly in machine learning and classification tasks. If adequately addressed, the latter results in achieving high prediction in accuracy, recall, F1-score, and other system of measurement [29].

Numerous research models have addressed some of BC's classification challenges using ML techniques, as evidenced by studies conducted using these techniques [22], [28]. To our knowledge, enhancing breast cancer prediction with ADASYN and optimizing hyperparameters using the Grid Search algorithm to improve the LightGBM classifier has not been explored, particularly on the BCC dataset. Therefore, this study seeks to improve predictive accuracy with focus on earlier detection and more precise treatment planning on ultimately enhance patient outcomes.

## 3. MATERIALS AND METHOD

The study introduces an efficient predictive model for classifying BCCD, showcasing the proposed structural design framework for breast cancer diagnosis in Fig 1. This framework employs GridsearchCV for hyperparameter optimization of the LightGBM classifier. The data undergoes preprocessing before implementing adaptive synthetic sampling and is subsequently divided into train and test datasets. After this, the training data is subjected to ADASYN to address bias in the predictive outcome. The hyperparameter search is further enhanced through grid search and cross-validation, contributing to the overall accuracy of classifiers. Afterwards, 20% of the overall BCCD, representing test data, is used to estimate the performance of the LightGBM classifier.

### 3.1 BREAST CANCER COIMBRA

The dataset, obtained from the University of California, Irvine (UCI) [30], consists of 64 women diagnosed with breast cancer and 52 healthy subjects. Nine (9) attributes represent anthropometric data – collected during routine blood analysis and a label attribute that signifies the presence of cancer-positive clinical results. The nine attributes that serve as potential risk factors for BC and their respective units are Body Mass Index (BMI) - Kg/m$^2$, Level of glucose - mg/dL, Age – year, Homeostasis model assessment (HOMA) – none, MCP-1 - Pg/dL, Adiponectin - Ug/ml, Leptin - Ng/ml, and Resistin - ng/mL. The purpose of this study is to forecast the future development of BC in subjects. To achieve this, we enhanced the LightGBM model using a grid search for hyperparameter optimization to forecast the likelihood of future breast cancer in individuals.

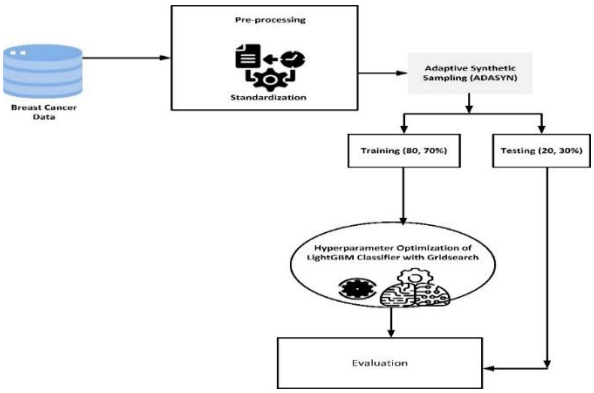


*Fig. 1 The Proposed Prediction Model for Breast Cancer*

### 3.2. THE PROPOSED PREDICTION MODEL

Constructing a model for predicting breast cancer involves the use of machine learning methods to analyze data and make predictions about the likelihood of a person having breast cancer. Early detection is pivotal with this tool and intervention. Several mathematical and statistical concepts are utilized in the model development to unveil hidden patterns in the domain of interest.

#### 3.2.1 Z-SCORE STANDARDIZATION

Z-score standardization, also known as standardization or zero-mean normalization, is a technique for preprocessing data used in statistics and ML to transform a dataset so that it has a mean (average) of 0 and a unitary standard deviation. This process is applied to individual features (columns) in the Coimbra breast cancer dataset, making them have a similar scale, which can benefit various machine learning algorithms. The procedure involves each data point in the feature deducting the mean (μ) of that feature from the data point and subsequently dividing the result by the standard deviation (σ) of that feature. The formula for standardization is as follows:

$$Standardized\ (z_i) = \frac{(x_i - \mu)}{\delta} \qquad (1)$$

where $X_i$ represents the original data point, while $\mu$ denotes the mean of the feature, and $\delta$ is the standard deviation of the feature. Hence, Fig 2 presents the normal distribution of the Z-score standardization of the BCCD.

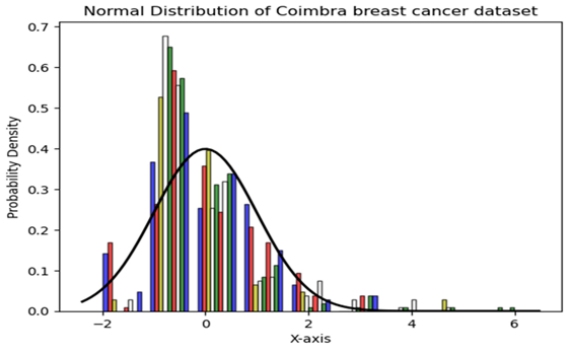


*Fig. 2. Normal distribution of BCCD*

However, a valuable step has been taken in exploring the correlation features in the breast cancer dataset. Correlation analysis aids in identifying feature pairs exhibiting high correlations with each other, indicating that the two features provide similar evidence to the model. Under such circumstances, consider removing one of the redundant features to reduce dimensionality and potentially improve model performance. Fig 3 shows the correlation plot of the BCCD. The correlation result, 0.93, from Fig 3 indicates that HOMA and Insulin are highly correlated.


*Fig. 3 Correlation map for BCCD*

### 3.2.2. ADAPTIVE SYNTHETIC SAMPLING

Adaptive Synthetic Sampling, often called ADASYN, is a data resampling technique used in imbalanced classification problems. In this scenario, the imbalance dataset exhibits a significant underrepresentation of one class (the minority class) compared to the other class (the majority)[31], [32]. ADASYN addresses the issue by generating synthetic samples for the minority class adaptively, focusing more on difficult-to-classify instances. In this study, the ADASYN technique was applied to the trainset of the breast cancer Coimbra dataset. In the process, the minority class was oversampled. Fig 4 presents statistical information on the trainset before and after applying ADASYN.

### 3.2.1. HYPERPARAMETER OPTIMIZATION (HPO)

Hyperparameter optimization, known as hyperparameter tuning, is used to find a machine learning model's optimal set of hyperparameters. These hyperparameters are predetermined and not derived from the data; they are established prior to the commencement of training. The reason for optimizing hyperparameter is finding the globally optimal value, denoted as $x^*$, for the objective function $f(X)$ to be minimized and evaluated for any arbitrary $x \in X$. Mathematically, the term can is expressed as:

$$x^* = \arg\min_{x \epsilon X} f(x) \quad (2)$$

where $X$ denotes a hyperparameter space encompassing discrete, categorical, and continuous variables [33]. Utilizing efficient hyperparameter optimization methods can make discovering the optimal hyperparameters for these models more straightforward when developing diverse machine learning models.

HPO comprises four key elements: initially, an estimator, which can manifest as either a regressor of a classifier featuring one or multiple objective functions; secondly, a defined search space; thirdly, an optimization technique for identifying the optimal combinations; and fourthly, a function to assess and compare the efficacy of different hyperparameter configurations [34]. Several optimization techniques exist [34]–[36], but the traditional techniques are unsuitable for finding optimal hyperparameter solutions. In this study, Grid Search optimization techniques and cross-validation are employed in the experiment, and their effects influence the performance of the LightGBM algorithm classification task despite its weakness. However, the faintness of the Grid search technique lies in its inability to perform optimally in ample search space. Nonetheless, the grid search technique proves suitable for hyperparameter tuning while searching for the parameters that could improve LightGBM performance.

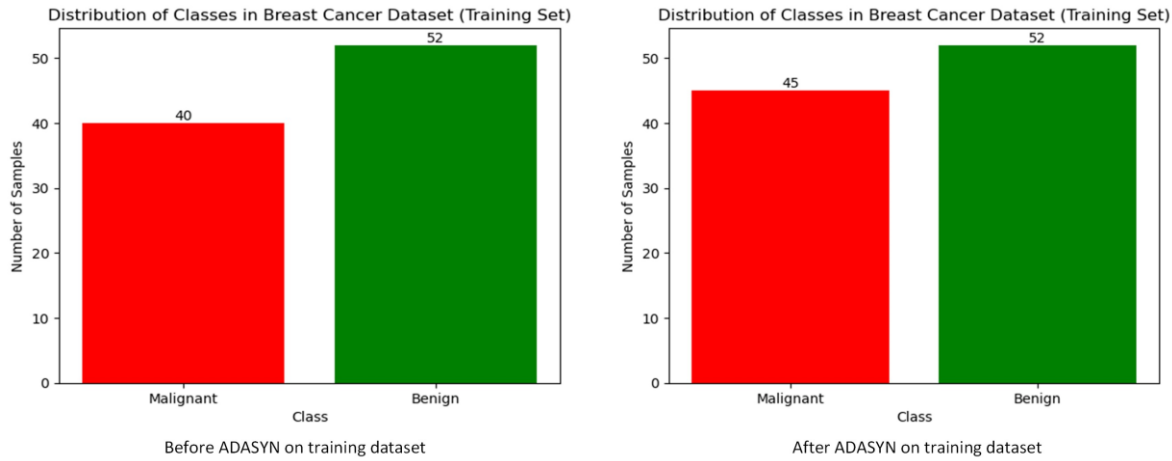


*Fig. 4 Malignant Instance Distribution in the Trainset and Post-ADASYN Growth*

**Algorithm** 1: Hyperparameter Algorithm

**Input**: *Objective function, performance metrics*

**Input**: *Hyper-parameters to tune, their types, an optimization technique*

**Initialize**: *Best configuration, Best performance score*
*Train ML model with default hyper-parameter configuration*
*Initialize: Search space with a sizeable hyper-parameter domain*

**Loop***:*
*a. Evaluate the model with the current hyper-parameter configuration*
*b. Update best configuration and performance if necessary*
*c. Narrow search space based on well-performing regions or explore new spaces*
*d. If convergence criteria are met, the* **exit loop**

**Output***: Best-performing hyper-parameter configuration*

### 3.2.2. OPTIMIZATION TECHNIQUE

Conventional optimization techniques [35] are unsuitable for hyperparameter optimization since HPOs differ. In this study, we employed the grid search to explore hyperparameter configuration space [37] and cross-validation for a reliable, optimal result. t stands for "Grid Search Cross-Validation." GridSearchCV is integrated with hyperparameter tuning, systematically exploring a range of values for the target hyperparameters through extensive cross-validation [38]. It assesses the model's performance by systematically testing and validating it for each unique combination of values within the dictionary [39]. It is a systematic and straightforward approach, ensuring no combination of hyperparameters is left unexplored. This set of hyperparameters minimizes a predefined loss function. It maximizes the model's effectiveness, resulting in improved results with reduced errors. This process serves to reduce errors within the models during this experiment.

### 3.2.3. MACHINE LEARNING MODEL

An ML model is a computational system that utilizes algorithms and statistical patterns to autonomously learn, predict, or make decisions without explicit programming for a particular task. [40], [41]. It processes and analyzes data to identify underlying patterns and relationships, enabling it to make informed predictions or classifications based on new, unseen data. A LightGBM model was used to predict breast cancer, as adopted in this experiment.

### 3.2.4. LIGHT GRADIENT BOOSTING MACHINE

LightGBM is a gradient-boosting framework designed for fast and efficient machine learning tasks. Developed by Microsoft in April 2017, it is an open-source algorithm for Gradient-Boosting Decision Trees (GBDT). Unlike traditional gradient-boosting algorithms,

**Algorithm** 1: Hyperparameter Algorithm

**Input**: *Objective function, performance metrics*

**Input**: *Hyper-parameters to tune, their types, an optimization technique*

**Initialize**: *Best configuration, Best performance score*
*Train ML model with default hyper-parameter configuration*
*Initialize: Search space with a sizeable hyper-parameter domain*

**Loop***:*
*a. Evaluate the model with the current hyper-parameter configuration*
*b. Update best configuration and performance if necessary*
*c. Narrow search space based on well-performing regions or explore new spaces*
*d. If convergence criteria are met, the* **exit loop**

**Output***: Best-performing hyper-parameter configuration*

LightGBM uses a histogram-based approach for constructing decision trees, enabling faster training and improved accuracy. LightGBM grows decision trees leaf-wise, focusing on splitting the leaf with the highest gain in variance, which enhances performance by effectively managing weak and strong learners (small and big gradients) [42].

### 3.2.2. OPTIMIZATION TECHNIQUE

Conventional optimization techniques [35] are unsuitable for hyperparameter optimization since HPOs differ. In this study, we employed the grid search to explore hyperparameter configuration space [37] and cross-validation for a reliable, optimal result. It stands for "Grid Search Cross-Validation." GridSearchCV is integrated with hyperparameter tuning, systematically exploring a range of values for the target hyperparameters through extensive cross-validation [38].

### 3.2.3. MACHINE LEARNING MODEL

An ML model is a computational system that utilizes algorithms and statistical patterns to autonomously learn, predict, or make decisions without explicit programming for a particular task. [40], [41]. It processes and analyzes data to identify underlying patterns and relationships, enabling it to make informed predictions or classifications based on new, unseen data. A LightGBM model was used to predict breast cancer, as adopted in this experiment.

### 3.2.4. LIGHT GRADIENT BOOSTING MACHINE

LightGBM is a gradient-boosting framework crafted for swift and efficient execution in machine learning assignments. Developed by the Microsoft team in April 2017, it is an open-source algorithm for Gradient-boosting decision tree (GBDT). It uses a histogram-based approach for decision tree construction, differentiating it from traditional Gradient-boosting algorithms. LightGBM builds decision trees leaf-wise [42], allowing for faster training and improved accuracy. The leaf-wise strategy is used for tree growth to identify a leaf with the highest gain in variance to perform a split. It takes records of weak and strong learners (small and big gradients, $g_i$).

When utilizing LightGBM, defining hyperparameters, including the number of iterations, is necessary. Table 1 presents the specifics of the hyperparameters employed in this study.

*Table 1 LightGBM hyperparameters*

| *LightGBM (Parameter)* | *Value* |
|---|---|
| *num_leaves* | *[31, 50, 100]* |
| *max_depth* | *[-1, 10, 20, 30]* |
| *learning_rate* | *[0.01, 0.1, 0.2]* |
| *n_estimators* | *[300, 400, 500]* |

**Algorithm 2: Adaptive Synthetic Sampling**

Input: Training set - $X$ with its element $\beta$ having zero mean and unitary standard deviation
Output: Synthetic minority samples $S \sqsubseteq X_{ADASYN}$

Split to compute majority and minority samples into: $n_{major}, n_{minor} \subset X$ training dataset

Hence, $G = (n_{major} - n_{minor}) \times \beta$They compute the quality of samples to synthesize within the class.

For each sample $n_i \in n_{minor}$ Class:
Compute $\Delta_i$ //represents majority of samples within the k-nearest neighbours of x.
Compute $r_i = {}^{\Delta_i}/_{k}$
Compute $g_i \leftarrow r_i \times G$
Loop from 1 to $g_i$
$s_i \leftarrow x_i + (x_{minor(i)} - x_i) \times \gamma$ synthesize data sample, where $\gamma \in [0,1]$
End loop
End for
Return oversampled minority Train set

### 3.3 EXPERIMENTAL SETTING

*Table 2 Establishing the environment for the proposed system*

| *Resource* | *Description* |
|---|---|
| *Central Processing Unit* | *Intel(R) Core™ i5-5200U @2.20GHz* |
| *RAM* | *12GB* |
| *Software* | *Jupyter Notebook* |
| *Language* | *Python* |

### 3.4. MODEL EVALUATION

Model evaluation is crucial to assessing a machine learning model's performance. The most common methods include training/testing splits, cross-validation, and diverse evaluation metrics, including accuracy, precision, recall, F1-score, and ROC/AUC. Moreover, by linking these concepts, evaluating the model's effectiveness and reliability is essential. However, the choice of method and metrics depends on the problem type and goals [43]. Consequently, this study not only underscores the importance of model evaluation but also presents classifier model performance metrics and their respective formulas for evaluation in Table 3.

*Table 3 Metric for evaluating performance*

| *Metrics* | *Formula* |
|---|---|
| *Accuracy (%)* | $\frac{TP+TN}{TP+TN+FP+FN} \times 100$ |
| *Recall (%)* | $\frac{TP}{TP+FN} \times 100$ |
| *Precision (%)* | $\frac{TP}{TP+FP} \times 100$ |
| *Specificity (%)* | $\frac{TN}{TN+FP} \times 100$ |

Furthermore, the AUC-ROC curve visually represents the classification model's performance. It is a widely used and vital statistic for evaluating its effectiveness. The ROC curve visually demonstrates the model's performance at various threshold values. The ROC is drawn between the parameters False positive and True positive rates. Hence, the effectiveness of a classifier increases with the AUC value during the model classification.

## 4. RESULTS AND DISCUSSION

### 4.1 MODEL PERFORMANCE ON BREAST CANCER

The LightGBM classifier, considering all features essential in the breast cancer Coimbra dataset, was employed to train and test the ML model. This significance becomes evident when computing the participation feature correlation, as illustrated in Fig 3. Notably, among the various features in the dataset, only Insulin and HOMA demonstrate the most substantial correlation value of 0.93.
To further enhance the LightGBM performance, gridsearchCV was employed, automating the quest for optimal hyperparameters in classification. The improvement is particularly notable after averaging 3-fold cross-validation, a result obtained using a grid search algorithm to determine optimal hyperparameters. Compared to its performance without grid search, LightGBM demonstrates optimum accuracy, precision, recall, F1 score, and AUC rates by 91.67%, 85.71%, 100%, 92.31%, and 97%, respectively. The Receiver Operating Characteristic curve further validates these advancements in classifier performance.

*Table 4 Breast cancer prediction metrics*

| *Model* | *Accuracy (%)* | *Recall (%)* | *Precision (%)* | *F1 score (%)* | *AUC* | *CV* | *Train/Test* |
|---|---|---|---|---|---|---|---|
| *LightGBM* | *74.29* | *66.67* | *80* | *72.72* | *73.53* | *3* | *70/30* |
| ***GridsearchCV+ LighGBM*** | ***91.67*** | *100* | *85.71* | *92.31* | *97.0* | *3* | *80/20* |
| *GridsearchCV+ LighGBM* | *87.5* | *83.33* | *90.91* | *86.96* | *90* | *5* | *80/20* |
| ***GridsearchCV+ LighGBM*** | ***91.67*** | *91.67* | *91.67* | *91.67* | *97.0* | *10* | *80/20* |
| *Gridsearch + LightGBM* | *87.5* | *83.33* | *90..91* | *86.96* | *90* | *-* | *80/20* |
| *Gridsearch+ LightGBM (without ADASYN)* | *74.29* | *77.78* | *73.68* | *75.68* | *74.0* | *3* | *70/20* |

Further, we decided to assess the performance of the classical LightGBM classifier, as presented in Table 4. The performance achieved the highest accuracy with the best Hyperparameters: {'n_estimators': 300, 'learning_rate': 0.01, 'num_leaves': 31, 'max_depth': -1}. To comprehensively evaluate the model's performance, we analyzed the ROC curve, a precise metric for evaluating model performance on unbalanced datasets [44]. The ROC curve can differentiate between false-positive and false-negative results, with an AUC value approximately equal to 1, signifying optimal model performance [45], [46]. The last row of Table 4 presents the model performance without using ADASYN to handle the imbalanced dataset. In the experiment, the best hyperparameters that give optimum result is{'learning_rate': 0.1, 'max_depth': -1, 'n_estimators': 300, 'num_leaves': 31}.(Fig. 5)

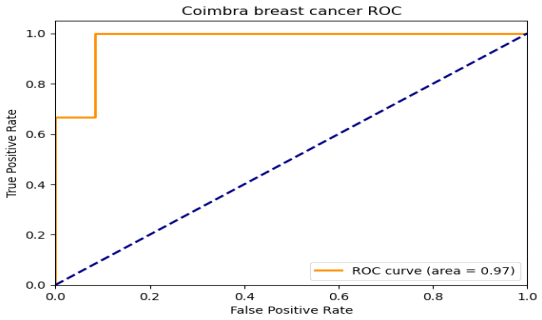


*Fig. 5 Breast cancer ROC curve LighGBM with ADASYN*

Further, we decided to evaluate the performance of the classical LightGBM model, as presented in Table 5. The performance achieved the highest accuracy with the best Hyperparameters: {'learning_rate': 0.01, 'max_depth': -1, 'n_estimators': 300, 'num_leaves': 31}. To comprehensively assess the model's performance, we analyzed the Receiver Operating Characteristic (ROC) curve, a specific metric suitable for evaluating model performance on unbalanced datasets [44]. The ROC curve provides a means to distinguish between false-positive and false-negative results, with an AUC value approximately equal to 1 indicating the best model performance [45], [46].(Fig. 6)

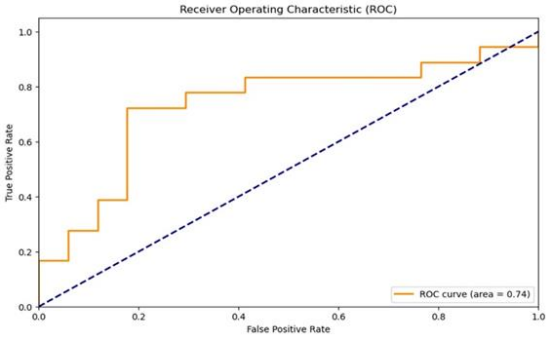


*Fig. 6 ROC curve of the LighGBM performance (without ADASYN)*

## 4.2. COMPARISON OF OUR STUDY WITH PREVIOUS WORKS

When evaluating a machine learning model's performance, the training methodology is crucial. Cross-validation is widely used in ML and statistical modeling to assess a model's effectiveness and generalizability [47]–[49]. It helps predict how well a model will perform with new, unseen data. Some existing works compare models built with and without cross-validation, finding that cross-validation is a more robust evaluation technique.

*Table 5 Comparing the proposed model with earlier research*

| *Author* | *Feature selection* | *Classifier* | *Cross-validation Method* | *Accuracy (%)* |
|---|---|---|---|---|
| [50] | *K-mean clustering (weighing feature)* | *AdaBoost* | *10-fold* | *91.27* |
| [51] | - | *EML* | | |
| [52] | *RFE* | *ANN* | *Train/test (70/30)* | *80* |
| [20] | - | *KNN* | *Train/test(90/10)* | *100* |
| [20] | - | *KNN* | *Train/test(80/20)* | *87.5* |
| [21] | *Gain Ratio* | *KNN* | - | *72.85* |
| [53] | - | *NB, RF, MLP, and SLR* | *Train/test (70/20)* | *85* |
| [46] | *Chi-Square* | *ensemble machine learning* | *20 - fold* | *78* |
| [54] | *RFE* | | *Train/test (80/20)* | *76.42* |
| *Our model* ***GridsearchCV + LightGBM*** | - | ***LightGBM*** | ***Train/test(80/20) 3-Fold CV*** | ***91.67*** |

This method involves partitioning the dataset into multiple subsets (folds) and systematically training and testing the model across various combinations of these folds. Cross-validation is preferred because it provides a more comprehensive assessment of a model's robustness and generalization capacity. This study employs 3-fold cross-validation guided by grid search to select the best hyperparameters for optimal model performance. However, our focus is on listing models built with cross-validation in the existing work. Table 5 presents a list of existing works on breast cancer Coimbra.

It is essential to recognize that comparing the direct performance of the reported results is not equitable, considering the diverse classification models, parameter configurations, and validation techniques employed. Consequently, the outcomes presented in Table 5 can serve not only to validate the efficacy of the categorization models but also to facilitate a broader comparison with prior research.

## 5. CONCLUSIONS AND FUTURE WORKS

In conclusion, our study tackles the critical challenge of breast cancer diagnosis by introducing an effective predictive model for early detection of this life-threatening disease. We investigate the use of ADASYN to address class imbalance in the training dataset and compare the model's performance with and without this adjustment. Our study employed GridSearchCV to identify the optimal hyperparameter configuration for our model by systematically exploring predefined sets of hyperparameters. We found that a 3-fold cross-validation outperformed other fold sizes in terms of reported performance metrics. Notably, a 10-fold cross-validation showed equal performance across all metrics—accuracy, recall, sensitivity, and F1 score. However, the LightGBM model's performance with a 5-fold cross-validation was unimpressive. We then compared the improved model with a standard LightGBM classifier baseline, revealing a significant performance enhancement. Our model achieved an average accuracy of 91.67%, 100% recall, 85.71% precision, 97% AUC, and an F1 score of 92.31%.

These findings highlight the substantial impact of hyperparameter optimization techniques and the use of ADASYN to address class imbalance in improving the efficacy of machine-learning models for breast cancer prediction, particularly on the BCCD dataset. Our research has significant implications in medicine, especially for precise and effective breast cancer prediction, contributing to improved early diagnosis and treatment planning. By showcasing our approach's efficacy, we aim to benefit the lives of those affected by this life-threatening disease. Future research could build upon this solution by utilizing remote patient monitoring wearable devices to reduce the associated risks, paving the way for further advancements in this field.

## 6. REFERENCES

[1] J. Ferlay *et al.*, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *Int. J. Cancer*, vol. 136, no. 5, pp. E359–E386, 2015, doi: 10.1002/ijc.29210.

[2] Nhs Choices, "Breast Cancer Symptoms," *WebMD*, pp. 1–5, 2010.

[3] D. Singh, N. Malila, A. Pokhrel, and A. Anttila, "Association of symptoms and breast cancer in population-based mammography screening in Finland," *Int. J. Cancer*, vol. 136, no. 6, pp. E630–E637, 2015, doi: 10.1002/ijc.29170.

[4] Centers for Disease Control and Prevention, "What Are the Symptoms of Breast Cancer?," *Breast Cancer*, pp. 1–33, 2020, [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm%0Ahttps://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm%0Ahttps://www.cdc.gov/cancer/breast/basic_info/symptoms.htm%0Ahttps://www.cdc.gov/cancer/breast/basic_info/symptoms.htm

[5] O. S. Rahmayani, R. H. Permana, and W. Witdiawati, "Early Detection of Breast Cancer According to Fertile Age Women," *Media Keperawatan Indones.*, vol. 3, no. 1, p. 32, 2020, doi: 10.26714/mki.3.1.2020.32-37.

[6] K. A. Černíková, L. Klůzová Kráčmarová, M. Pešoutová, and P. Tavel, "Patient delay in presenting symptoms of breast cancer in women in the czech republic," *Klin. Onkol.*, vol. 34, no. 1, pp. 40–48, 2021, doi: 10.48095/ccko202140.

[7] WHO, "Global Breast Cancer Initiative Iamplementation Framework: Assessing, Strengthening and Scaling up of Services for the Early Detection and Management of Breast Cancer," *World Heal. Organ.*, p. 118, 2023, [Online]. Available: https://www.who.int/publications/i/item/9789240065987

[8] T. Srinivas *et al.*, "Novel Based Ensemble Machine Learning Classifiers for Detecting Breast Cancer," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/9619102.

[9] V. D. P. Jasti *et al.*, "Computational Technique Based on Machine Learning and Image Processing for

Medical Image Analysis of Breast Cancer Diagnosis," *Secur. Commun. Networks*, vol. 2022, 2022, doi: 10.1155/2022/1918379.

[10] R. Shukla, V. Yadav, P. R. Pal, and P. Pathak, "Machine learning techniques for detecting and predicting breast cancer," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 7, pp. 2658–2662, 2019.

[11] R. H. Khan, J. Miah, M. M. Rahman, and M. Tayaba, "A Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer," *2023 IEEE 13th Annu. Comput. Commun. Work. Conf. CCWC 2023*, pp. 647–652, 2023, doi: 10.1109/CCWC57344.2023.10099106.

[12] S. O. Folorunso, J. B. Awotunde, A. A. Adigun, L. V. N. Prasad, and V. L. Lalitha, "A hybrid model for post-treatment mortality rate classification of patients with breast cancer," *Healthc. Anal.*, vol. 4, 2023, doi: 10.1016/j.health.2023.100254.

[13] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis," *Healthc. Anal.*, vol. 4, 2023, doi: 10.1016/j.health.2023.100218.

[14] S. Akbulut, I. B. Cicek, and C. Colak, "Classification of Breast Cancer on the Strength of Potential Risk Factors with Boosting Models: A Public Health Informatics Application," *Haseki Tip Bul.*, vol. 60, no. 3, pp. 196–203, 2022, doi: 10.4274/haseki.galenos.2022.8440.

[15] J. Hancock and T. M. Khoshgoftaar, "Impact of Hyperparameter Tuning in Classifying Highly Imbalanced Big Data," *Proc. - 2021 IEEE 22nd Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2021*, pp. 348–354, 2021, doi: 10.1109/IRI51335.2021.00054.

[16] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 4, p. 1502, 2016, doi: 10.12928/telkomnika.v14i4.3956.

[17] F. M. Talaat and S. A. Gamel, "RL based hyper-parameters optimization algorithm (ROA) for convolutional neural network," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 10, pp. 13349–13359, 2023, doi: 10.1007/s12652-022-03788-y.

[18] R. Jullapak and A. Thammano, "Backpropagation Neural Network with Adaptive Learning Rate for Classification," *Lect. Notes Data Eng. Commun. Technol.*, vol. 153, pp. 493–499, 2023, doi: 10.1007/978-3-031-20738-9_56.

[19] A. Bannach-Brown *et al.*, "Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error," *Syst. Rev.*, vol. 8, no. 1, 2019, doi: 10.1186/s13643-019-0942-7.

[20] Naveen, R. K. Sharma, and A. Ramachandran Nair, "Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models," *2019 4th IEEE Int. Conf. Recent Trends Electron. Information, Commun. Technol. RTEICT 2019 - Proc.*, pp. 100–104, 2019, doi: 10.1109/RTEICT46194.2019.9016968.

[21] N. Hasdyna, B. Sianipar, and E. M. Zamzami, "Improving the Performance of K-Nearest Neighbor Algorithm by Reducing the Attributes of Dataset Using Gain Ratio," *J. Phys. Conf. Ser.*, vol. 1566, no. 1, 2020, doi: 10.1088/1742-6596/1566/1/012090.

[22] A. K. Mishra, P. Roy, and S. Bandyopadhyay, "Genetic algorithm based selection of appropriate biomarkers for improved breast cancer prediction," *Adv. Intell. Syst. Comput.*, vol. 1038, pp. 724–732, 2020, doi: 10.1007/978-3-030-29513-4_54.

[23] M. Sinan Basarslan and F. Kayaalp, "Performance evaluation of classification algorithms on diagnosis of breast cancer and skin disease," *Stud. Comput. Intell.*, vol. 908, pp. 27–35, 2021, doi: 10.1007/978-981-15-6321-8_2.

[24] A. Rashid, S. S. Binta Farhad, A. Bhuyian, N. Yeasmin, M. A. Azim, and Z. Alom, "A Comparative Analysis of Machine Learning techniques on Breast Cancer diagnosis using WEKA," *Proc. 2022 25th Int. Conf. Comput. Inf. Technol. ICCIT 2022*, pp. 663–668, 2022, doi: 10.1109/ICCIT57492.2022.10055421.

[25] R. K. Barwal, N. Raheja, M. Bhiyana, and D. Rani, "Machine Learning-Based Hybrid Recommendation (SVOF-KNN) Model For Breast Cancer Coimbra Dataset Diagnosis," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 1, pp. 23–42, 2023, doi: 10.17762/ijritcc.v11i1.5983.

[26] A. K. Mishra, P. Roy, and S. Bandyopadhyay, "Binary Particle Swarm Optimization Based Feature Selection (BPSO-FS) for Improving Breast Cancer Prediction," *Adv. Intell. Syst. Comput.*, vol. 1164, pp. 373–384, 2021, doi: 10.1007/978-981-15-4992-2_35.

[27] W. Rhmann, "Optimised deep neural network for cancer disease prediction using a genetic algorithm," *Int. J. Bioinform. Res. Appl.*, vol. 18, no. 6, pp. 578–595, 2023, doi: 10.1504/IJBRA.2022.129262.

[28] A. Alsabry, M. Algabri, A. M. Ahsan, M. A. A. Mosleh, A. A. Ahmed, and H. A. Qasem, "Enhancing Prediction Models' Performance for Breast Cancer using SMOTE Technique," pp. 1–8, 2023, doi: 10.1109/esmarta59349.2023.10293726.

[29] S. Sarkar, A. Pramanik, J. Maiti, and G. Reniers, "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," *Saf. Sci.*, vol. 125, 2020, doi: 10.1016/j.ssci.2020.104616.

[30] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, 2018, doi: 10.1186/s12885-017-3877-1.

[31] Y. Feng, M. Zhou, and X. Tong, "Imbalanced classification: A paradigm-based review," *Stat. Anal. Data Min.*, vol. 14, no. 5, pp. 383–406, 2021, doi: 10.1002/sam.11538.

[32] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding

the Best Classification Threshold in Imbalanced Classification," *Big Data Res.*, vol. 5, pp. 2–8, 2016, doi: 10.1016/j.bdr.2015.12.001.
[33] H. Cho, Y. Kim, E. Lee, D. Choi, Y. Lee, and W. Rhee, "Basic Enhancement Strategies When Using Bayesian Optimization for Hyperparameter Tuning of Deep Neural Networks," *IEEE Access*, vol. 8, pp. 52588–52608, 2020, doi: 10.1109/ACCESS.2020.2981072.
[34] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020, doi: 10.1016/j.neucom.2020.07.061.
[35] C. Gambella, B. Ghaddar, and J. Naoum-Sawaya, "Optimization Models for Machine Learning: A Survey," *arXiv Prepr. arXiv1901.05331*, pp. 1–40, 2019, [Online]. Available: http://arxiv.org/abs/1901.05331
[36] P. Hijma, S. Heldens, A. Sclocco, B. Van Werkhoven, and H. E. Bal, "Optimization Techniques for GPU Programming," *ACM Comput. Surv.*, vol. 55, no. 11, 2023, doi: 10.1145/3570638.
[37] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Syst.*, vol. 200, 2020, doi: 10.1016/j.knosys.2020.105992.
[38] N. Ahmed *et al.*, "Machine learning based diabetes prediction and development of smart web application," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 229–241, 2021, doi: 10.1016/j.ijcce.2021.12.001.
[39] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif, and S. Islam, "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2019-Octob, pp. 2023–2028, 2019, doi: 10.1109/TENCON.2019.8929527.
[40] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 587–601, 2017, doi: 10.1145/3133956.3134077.
[41] M. Batta, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, p. 7, 2020.
[42] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 3147–3155, 2017.
[43] Y. Li, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction," *Appl. Comput. Math.*, vol. 7, no. 4, p. 212, 2018, doi: 10.11648/j.acm.20180704.15.
[44] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit*, vol. 30, no. 7, pp. 1145–1159, 2019.
[45] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005, doi: 10.1109/TKDE.2005.50.
[46] S. Nanglia, M. Ahmad, F. Ali Khan, and N. Z. Jhanjhi, "An enhanced Predictive heterogeneous ensemble model for breast cancer prediction," *Biomed. Signal Process. Control*, vol. 72, 2022, doi: 10.1016/j.bspc.2021.103279.
[47] A. Seraj *et al.*, "Cross-validation," *Handb. HydroInformatics Vol. I Class. Soft-Computing Tech.*, pp. 89–105, 2022, doi: 10.1016/B978-0-12-821285-1.00021-X.
[48] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *J. Am. Stat. Assoc.*, 2023, doi: 10.1080/01621459.2023.2197686.
[49] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.
[50] K. Polat and U. Senturk, "A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier," *ISMSIT 2018 - 2nd Int. Symp. Multidiscip. Stud. Innov. Technol. Proc.*, 2018, doi: 10.1109/ISMSIT.2018.8567245.
[51] Y. Celik, K. Sabanci, A. Durdu, and M. F. Aslan, "Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data," *Int. J. Intell. Syst. Appl. Eng.*, vol. 6, no. 4, pp. 289–293, 2018, doi: 10.18201/ijisae.2018648455.
[52] M. U. Ghani, T. M. Alam, and F. H. Jaskani, "Comparison of Classification Models for Early Prediction of Breast Cancer," *3rd Int. Conf. Innov. Comput. ICIC 2019*, 2019, doi: 10.1109/ICIC48496.2019.8966691.
[53] T. Khatun *et al.*, "Performance Analysis of Breast Cancer: A Machine Learning Approach," *Proc. 3rd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2021*, pp. 1426–1434, 2021, doi: 10.1109/ICIRCA51532.2021.9544879.
[54] A. Rasool, C. Bunterngchit, L. Tiejian, M. R. Islam, Q. Qu, and Q. Jiang, "Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis," *Int. J. Environ. Res. Public Health*, vol. 19, no. 6, 2022, doi: 10.3390/ijerph19063211.