Optimized Machine Learning Performance with Feature Selection for Breast Cancer Disease Classification

Koirunnisa, Amril Mutoi Siregar, Sutan Faisal
Departement of Informatics, Faculty of Computer Science, University Buana Perjuangan, Karawang 41361, Indonesia

ARTICLE INFO

Article history:

Received November 03, 2023
Revised December 20, 2023
Published December 23, 2023

Keywords:

Breast Cancer;
Classification;
Machine Learning;
Supervised Algorithm;
PCA

ABSTRACT

The prevalence of breast cancer is relatively high among adults worldwide. Particularly in Indonesia, according to the latest data from the World Health Organization (WHO), breast cancer accounts for 1.41% of all deaths and continues to increase. In order to address this growing issue, a proactive approach becomes essential. Therefore, the objective of this study is to classify the diagnosis of breast cancer into two categories: Benign and Malignant. Moreover, this classification pattern can serve as a benchmark for early detection and is expected to reduce mortality and cancer rates in breast cancer cases. The dataset used in this study is obtained from Kaggle and consists of 569 rows with 32 attributes. Various machine learning algorithms, such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naïve Bayes (NB), are employed for the classification analysis in this disease. . This study uses Principal Component Analysis (PCA) for optimized feature selection techniques with dimension reduction are employed on the dataset prior to modeling the data. Our highest accuracy model is the Support Vector Machine (SVM) with an RBF kernel, utilizing c-value selection. Additionally, the Logistic Regression (LR) model achieves an accuracy of 97.3%. However, it is worth noting that the precision and recall of the SVM model are both 100%. Moreover, the Receiver Operating Characteristic (ROC) curve indicates that the SVM graph surpasses the LR graph, which can be attributed to the results obtained from the confusion matrix calculation, where the False Positive Rate is found to be 0. Consequently, the overall performance evaluation of the SVM model with an RBF kernel, along with the utilization of the c-value selection approach, is significantly superior. This is primarily due to the fact that the SVM model does not make any incorrect predictions by classifying something as positive when it is actually negative.

Corresponding Author:

Amril Mutoi Siregar, Departement of Informatics, Faculty of Computer Science, University Buana Perjuangan, Karawang 41361, Indonesia
Email: amrilmutoi@ubpkarawang.ac.id

1. INTRODUCTION

According to data on cancer provided by the World Health Organization (WHO), breast cancer ranks second in terms of the leading causes of mortality among women, affecting approximately 95% of countries globally. The annual incidence of breast cancer exceeds 2.3 million cases, rendering it the most prevalent form of cancer among adults [1], [2], In Indonesia, breast cancer accounts for 1.41% of all deaths. Until the present time, surgical intervention has been a prevalent approach to address breast cancer, with subsequent administration of chemotherapy or radiation as deemed necessary. Consequently, in the event that this ailment is identified at an early stage, the potential detrimental consequences can be promptly averted. The utilization of machine learning algorithms for classification holds immense potential in facilitating the medical practitioners' endeavors in this realm [3], [4].

Our methodology for addressing this particular classification problem entails the utilization of a supervised learning classifier. Specifically, we employ five renowned classifiers, namely the Neural Network (NN) with ReLU (Rectified Linear Unit) and Sigmoid layer, Support Vector Machine (SVM) utilizing both the Linear kernel and the RBF (Radial Basis Function) kernel with c-value selection, Decision Tree (DT) employing the default Gini index, Random Forest (RF) with the default n-estimator, Logistic Regression (LR), K-Nearest Neighbor (KNN), and Naïve Bayes (NB). The evaluation of their performance is conducted through the application of Principal Component Analysis (PCA). PCA is a technique utilized for feature selection, which entails dimension reduction by converting a large number of features in a dataset into a smaller number [5], the kernel calculation utilizing Eigenvalue was conducted, and the results were visualized using a Pareto plot for the purpose of facilitating the identification of an optimal PCA value. In this study, various machine learning algorithms commonly employed include the Artificial Neural Network (ANN), which is inspired by the human brain and has the capability to capture intricate relationships among the features in the breast cancer dataset [6]. Another algorithm, Support Vector Machine (SVM), is a robust binary classification method that is well-suited for breast cancer classification due to its ability to identify the optimal hyperplane [7] that maximally separates the two classes with a margin. Decision Tree (DT) is a classification algorithm that makes decisions based on a set of predefined rules in the context of breast cancer classification [8], It maps features such as tumor size, shape, and density to make informed decisions. Random Forest (RF) is known to provide more stable results and minimize overfitting compared to a single decision tree [9]. Logistic Regression (LR), despite its name, is used for classification problems. In the context of breast cancer, it models the probability that a tumor is malignant based on specific features [10]. And subsequently, NB employs statistical techniques to estimate the probabilities of each class based on the distribution of features, yielding favorable outcomes for specific datasets [11].

Numerous prior investigations have examined the application of Principal Component Analysis (PCA) in various machine learning algorithms, as highlighted [12]. This study employs PCA by dividing the data into a ratio of 70:30, and utilizes SVM, LR, KNN, DT, NB, and RF as the proposed models. The highest level of accuracy achieved is 96.5% by utilizing SVM with the UCI Wisconsin BC dataset. Another study conducted by Verghese *et al* utilized Principal Component Analysis (PCA) to classify High Dimensional Low Sample Size Data (HDLSS) [13]. In this study, the researchers employed the Cor-Ex classifier and compared various machine learning algorithms. The results demonstrated that the Support Vector Machine (SVM) with the Radial Basis Function (RBF) kernel achieved the highest accuracy of 94.56%. The UCI Wisconsin BC dataset was utilized for this analysis.

In a separate investigation by H. Chiu *et al.*, PCA was employed for breast cancer detection. The researchers utilized the UCI Wisconsin BC dataset and found that the data processed solely using the PCA technique yielded results comparable to those obtained by extracting characteristics using the same technique and subsequently combining them with the Multilayer Perceptron (MLP). After analyzing the learning process and combining it with SVM, the accuracy reached 86.97% [14]. Another research from Assegie, *et al*, developed a model to diagnose breast cancer with Adaptive Boosting. The researcher using Kaggle dataset by using Pearson correlation for feature selection and reach the accuracy is 92.5% [15].

Based on the problems previously describe and with the support of the result of research that has been done previously related to the implementation on feature selection for optimizing machine learning algorithms in breast cancer classification we made improve on this research using seven different famous supervised learning classifiers. Our contribution is PCA technique is applied with different techniques to each classifier. Each classifier is compared against each various based on performance metrics, especially with ROC (Receiver Operating Characteristic), and Confusion matrix. with the result of the classification by supervised algorithm, patients with existing parameters can be classified between Benign and Malignant cancer. So that, this pattern can be used for benchmark diagnosis so that can be detected early and is expected to be able to reduce mortality and cancer rates in breast cancer, and this is our main contribution for research. This model is anticipated to aid pathologists in conducting examinations with greater consistency and efficiency in order to detect breast cancer diagnoses.

2. METHODS

The proceedings conducted in this investigation commence with a scholarly examination. Scholarly examinations are conducted with the objective of discovering the theoretical underpinnings employed, as well as seeking relevant scientific literature to bolster the investigation. In the entirety of this investigation, the subsequent procedures or stages of research will be executed Fig. 1.

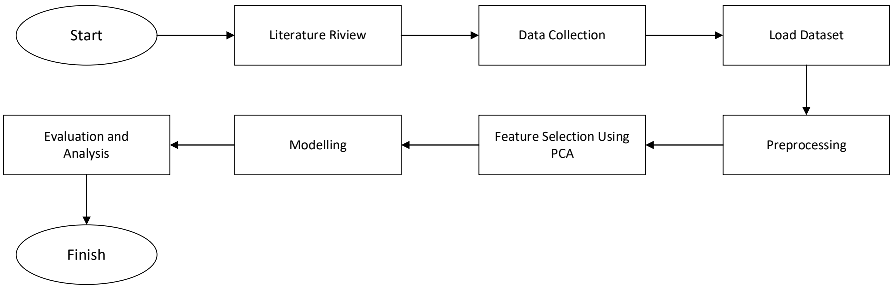


Fig. 1. Flowchart of research stage

2.1. Dataset and Attribute Description

The dataset on breast cancer was obtained from the Kaggle website by M Yasser H, an AI & ML Engineer at Media Agility Bengaluru, Karnataka, India. The dataset is accessible at the following URL: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset. The decision to use this dataset was influenced by its inclusion in the UCI Winsconsin Dataset, which is presented in a binary diagnostic format and consists of a total of 569 samples. This selection represents a unique approach to research, as most commonly employed datasets are derived from the UCI Winsconsin Breast Cancer Dataset. The description of the BC Kaggle dataset is depicted in the following section Table 1.

Table 1. Kaggle BC Dataset Description

No	Features of BC Kaggle	Description
1	ID	Unique ID
2	Diagnosis	Target: M-Malignant is cancerous B- Benign is non-cancerous
3	Radius	Average of distances from center to circumference points.
4	Texture	Standard deviation (SD) of gray-scale value.
5	Perimeter	Gross distance between the snake points
6	Area	Total number of pixels on the inside of the snake along with one half of the pixels in circumference.
7	Smoothness	Local variance in length difference.
8	Compactness	Perimeter^2/Area.
9	Concavity	Intensity of the contours concave parts.
10	Concave Points	The number of contour concavities.
11	Symmetry	The difference in length between lines perpendicular to the major axis in both directions to the cell boundary.
12	Fractal Dimension	Coastline estimation. A higher value leads to a less normal contour representing a higher risk of malignancy.

Next, we do an EDA (Exploratory Data Analysis). We can see there are two final classes in dataset which is diagnosis attribute. The class is Malignant and Benign, and this class will be the target of the classification as shown in Fig. 2.

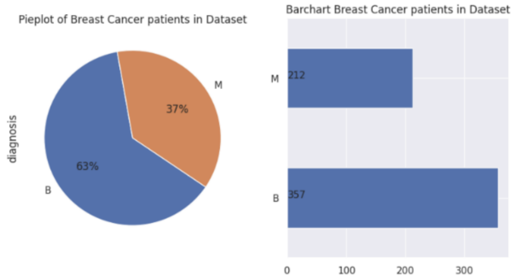


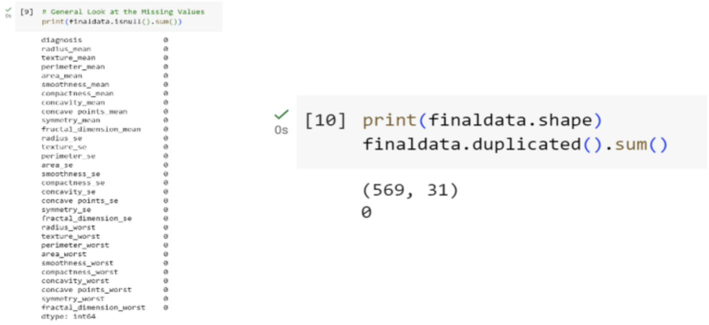
Fig. 2. EDA distribution class in dataset

2.2. Preprocessing

In the process of generating high-quality data, a number of techniques were employed:

1) The initial step involves Data Cleaning, whereby the dataset undergoes a thorough cleansing process to ensure optimal quality. In this process, the observation values are meticulously presented by avoiding the repetition of prefixes for each feature [16]. This technique aims to identify and eliminate any id columns, incomplete data (i.e., missing values), and duplicate data while preserving the essence and substance of the dataset. It is worth noting that the dataset under

consideration contains no missing values or duplicate entries, as depicted in the provided information Fig. 3.



```
[9] # General Look at the Missing Values
    print(finaldata.isnull().sum())

    diagnosis                  0
    radius_mean                0
    texture_mean               0
    perimeter_mean             0
    area_mean                  0
    smoothness_mean            0
    compactness_mean           0
    concavity_mean             0
    concave points_mean        0
    symmetry_mean              0
    fractal_dimension_mean     0
    radius_se                  0
    texture_se                 0
    perimeter_se               0
    area_se                    0
    smoothness_se              0
    compactness_se             0
    concavity_se               0
    concave points_se          0
    symmetry_se                0
    fractal_dimension_se       0
    radius_worst               0
    texture_worst              0
    perimeter_worst            0
    area_worst                 0
    smoothness_worst           0
    compactness_worst          0
    concavity_worst            0
    concave points_worst       0
    symmetry_worst             0
    fractal_dimension_worst    0
    dtype: int64
```

```
[10] print(finaldata.shape)
     finaldata.duplicated().sum()

     (569, 31)
     0
```

Fig. 3. General look for missing values and duplicate data

2) In this study, a process of data transformation is conducted in order to enhance the accuracy and efficiency of the algorithm. This process involves encoding the categories into numeric values and subsequently segregating them [17]. For instance, we assign the value of 1 to the diagnosis label "Benign" and the value of 0 to the diagnosis label "Malignant". It is important to note that the other feature, apart from the diagnosis, consists of numerical data.
3) Furthermore, data scaling and normalization are also carried out with the objective of normalizing the data by utilizing Z-Score scaling to achieve standardization. The formula employed in this normalization process is as follow (1).

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

X is The standardization of a feature is a crucial aspect to consider in data analysis. The symbol, μ is represents the mean value of the entire dataset, while σ symbolizes the standard deviation value. In order to provide a clear visual representation. The Fig. 4 presents a distribution histogram showcasing the numerical data both before and after the normalization process.

The histogram demonstrates that the distribution of numerical data prior to normalization on the left side is indicative of a normal and satisfactory distribution. The selection of the Z-score is justified due to its effectiveness in scaling data with a normal distribution and its insensitivity to outliers [18]. Upon standardization on the right side, the Z-score effectively preserves the shape of the distribution [19], as illustrated in Fig. 4.

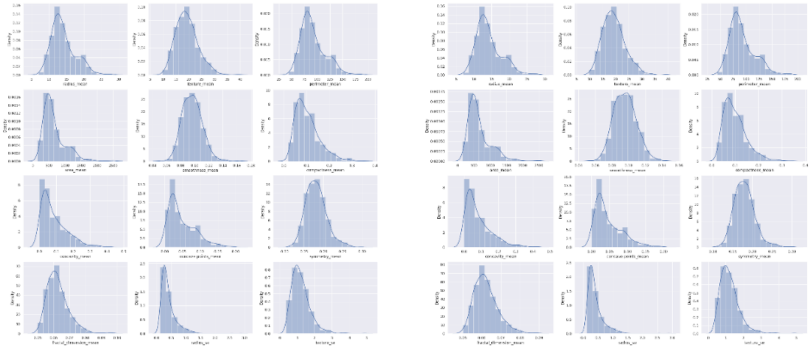


Fig. 4. Histogram of numerical data before and after normalization

The dataset consisting of 569 data points is divided into training and test data. The ratio of the two datasets is set at 80:20, meaning that the training data accounts for 80% while the test data comprises 20%. Given the limited presence of outliers in the dataset, the 20% set aside for testing is deemed sufficient for the objective evaluation of the final models. Consequently, the training data consists of 455 points, whereas the testing data consists of 114 points. After the data has been partitioned, it undergoes Principal Component Analysis (PCA) as a preliminary step. Subsequently, the data is fed into the model and subjected to testing procedures.

2.3. Feature Selection Using PCA

Principal Component Analysis (PCA) aims to project the component vectors onto a lower dimension in order to maximize the variances of the displayed data [20]. If the number of features expands for a given constant sample size, and the out-of-bounds eigen values of the population covariance matrix are sufficiently large compared to the main components, they will merge with the population covariance matrix [21]. The dataset used in this study has a high dimensionality, with 32 features. This large number of features impedes the achievement of the best result and leads to overfitting. Consequently, PCA is employed on this dataset to transform the 32 features into 4 features, thereby enhancing the performance of the result. PCA is utilized to reduce the dimensionality by converting a large number of features in a dataset into a small number of features known as the principal component [22]. The primary advantages of PCA include reducing overfitting, removing correlated features, and improving the performance of machine learning algorithms [23].

In this study, we employed a pareto plot, as shown in Fig. 5, to examine the Eigenvalue and determine the optimal number of principal components (PCs) for the subsequent modeling process. In a pareto chart of PCA Pareto curve, the x-axis typically represents the component numbers, while the y-axis represents the singular values or variance explained by each component. This curve provides a visual representation of the contribution of each component to the total variance in the dataset [24]. The pareto plot indicates that the optimal number of PCs is 4, with a cumulative explained variance of 0.79. The concept of cumulative explained variance is a fundamental aspect of PCA, which is a dimension reduction technique employed in multivariate data analysis. Subsequently, we utilized the dimension of the training and testing data for further use in modeling.

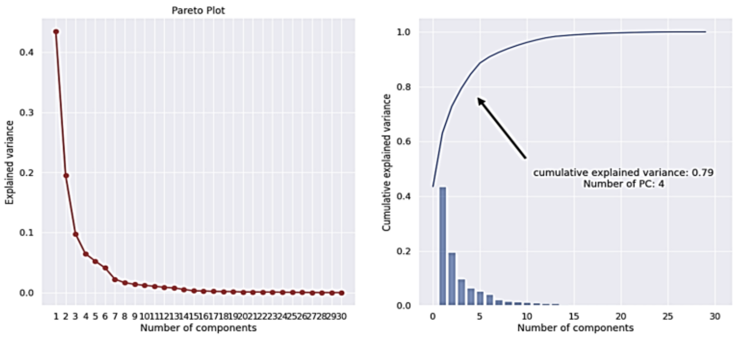


Fig. 5. Pareto plot of eigenvalue in PCA

2.4. Models Overview

1) Artificial Neural Network (ANN)

Over the past few decades, ANN has been widely utilized in various studies, establishing it as a related research field. Notably, these networks have achieved tremendous success, particularly in breast cancer classification and early-stage prognosis. ANN models typically consist of three layers: Input, Hidden, and Output [25], [26]. The layers are comprised of interconnected neurons with non-linear activation functions to enhance the network's nonlinear capacity. Firstly, the Input layer receives the data, which is then transmitted to the Hidden layer for analysis. The result is then sent back to the Output layer, where the displayed result is shown However, training an ANN is likely to involve a lengthy series of computational processes due to these limitations. The activation process of Hidden nodes in an ANN is explained in (2)

$$\begin{gathered} J_1 = \sum_k V_{lk} X_k + b_1 \\ K_1 = g_1(J_1) \end{gathered} \tag{2}$$

The activation function specified in (2) is represented by g_1, while V_{lk} is defined as the weight linked to both the input layer and the hidden layer. The term b_1 denotes the bias between the input and hidden layers at each connection. Additionally, X_k signifies the input at the input layer, J_1 is the summation of the weighted input with bias, and K_1 represents the output at the hidden layer of the activation function.

$$\begin{gathered} J_m = \sum_l V_{ml} X_l + b_m \\ K_m = g_m(J_m) \end{gathered} \tag{3}$$

In the provided (3), g_m represents the activation function, V_{ml} denotes the weights associated between the output layer m and hidden layer l, and b_m is the bias between the hidden and output layers at each connection. The term X_l signifies the output of the hidden layer at each node, J_m is the summation of weights at the output layer, and K_m represents the final output of the output layer. Referring to (2) and (3), m signifies the output layer, l is the hidden layer, and k represents the input layer.

In this study, we employ three dense hidden layers in a ReLU and Sigmoid activation, as illustrated in Fig. 6. ReLU is utilized to replace negative values with zero and leave positive values unchanged. Subsequently, two dropout levels are introduced, along with Sigmoid activation with batch size=32 and epoch=200. The implementation of a dropout layer is considered to address the issue of overfitting, where the validation loss is high while the training loss is low. Therefore, to overcome this problem, we employ Early Stopping with verbose=2 and patience=40, as depicted in Fig. 7.

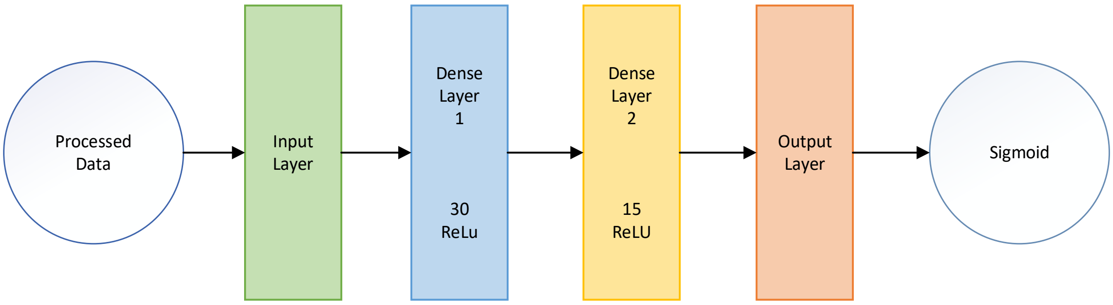


Fig. 6. Proposed ANN flowchart method

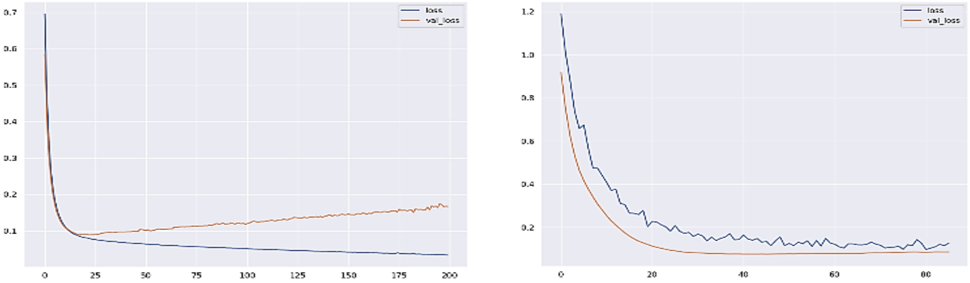


Fig. 7. ANNs visualization before and after using Early Stopping

Early Stopping is a technique employed in model training that terminates the training process prematurely if indications of overfitting are detected, measured through a metric on the validation dataset. This technique helps prevent the model from excessively memorizing the training data and enhances its ability to generalize [17].

2) Support Vector Machine (SVM)

This method is a supervised machine learning approach used in pattern classification to enhance security and service quality. SVM proves to be effective in many classification cases by constructing an optimal hyperplane with maximum geometric margins [28]. The SVM formula is:

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x_j) + b \tag{4}$$

In the context of (4) is N is represent the number of training samples, a_i denotes the weights calculated during the training process, y_i corresponds to the class label of the i training sample, $K(x_i, x_j)$ is stands for kernel, and b is represent the bias in term. In this study, we compare the linear and RBF kernel to see which kernel has the best C value. The formula of Linear and RBF kernel is:

$$K(x_i, x_j) = x_i . x_j \tag{5}$$

$$K(x_i, x_j) = exp\left(-\frac{||x_i, x_j||^2}{2\sigma^2}\right) \tag{6}$$

In the context of (5) that x_i is the feature vector of a sample and x_j is the feature vector of the training sample. Then, in (6) $||x_i, x_j||^2$ is the square of the Euclidean distance between x_i and x_j, exp is the exponential value of the number x that is an Euler's number (approximately 2.71828), and σ is the parameter for the width of the RBF kernel. The Linear and RBF kernel functions are widely used, as they are derived directly from the inner product of the original features. They are particularly beneficial in scenarios where they offer advantages such as fewer parameters and fast processing [28]. In such cases, opting for alternative parameter functions becomes essential for better suitability, so we choose the C parameters**.**

The RBF kernel with a value of c=2 showcases the most optimal performance in terms of training and test accuracy, as indicated in Fig. 8.

Fig. 8 The parameter C in SVM functions as a penalty parameter for classification errors present in the training data. It regulates the extent of penalty assigned to data points that fall on the incorrect side of the separating hyperplane. A higher value of C results in a greater penalty, thereby making the model more rigorous in addressing classification errors during training. By appropriately selecting the value of C, we can strike an optimal balance between fitting the model to the training data and preventing overfitting.

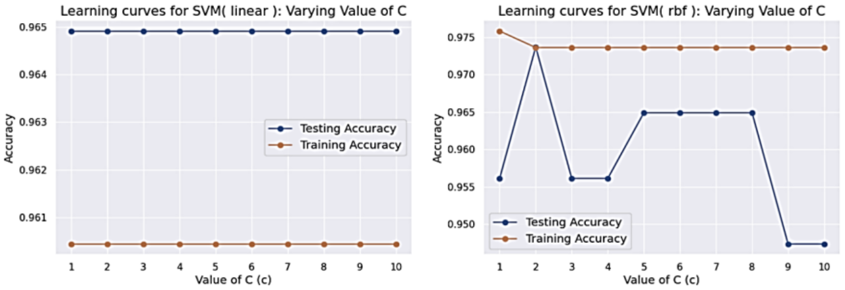


Fig. 8. SVM linear and RBF kernel for best c value

3) Random Forest (RF)

This approach involves the creation of a decision tree based on a random subset of designated data. Subsequently, it leverages each tree's forecast to determine the most favorable solution through voting [29]. This process entails constructing a decision tree for each subset and deriving a forecast output from each tree, followed by voting for each prediction outcome. The final prediction is determined based on the outcome with the highest number of votes. Random forest utilizes the Gini coefficient to construct decision trees [30]. Assuming the training set encompasses n features, the Gini index coefficient, derived from the CART learning system, is employed for constructing decision trees. The Gini coefficient quantifies the dissimilarity between values within a frequency distribution. It can be defined as follows:

$$Gini(T) = 1 - \sum_{j=1}^{n} (Pj)^2 \tag{7}$$

In the context of (7) , a Gini coefficient of zero states perfect likeness and a coefficient value of 1 expresses maximal inequality between values. If a dataset T contains examples from n classes and Pj is the relative frequency.

4) Decision Tree (DT)

This technique comprises nodes and branches that resemble a hierarchical tree structure. Each node corresponds to a specific feature, while each leaf provides information about the outcome, which can be either discrete or continuous. The branches convey the rules, and the distinct parameters are established [31], [32]. Our decision tree (DT) utilizes 'entropy' as the criterion, with a maximum depth of ten and a random state number of 100. The uncertainty within the dataset is measured using entropy as a metric. The entropy function, which characterizes the degree of uncertainty, is defined as:

$$H(K) = \sum_{c \in C} -p(c)log_2 p(c) \tag{8}$$

In the context of (8), C belongs to the class of dataset either it is benign or malignant. K is dataset and $p(c)$ belongs to proportion of number of elements related to dataset K and class c. The visualization shown in the Fig. 9.

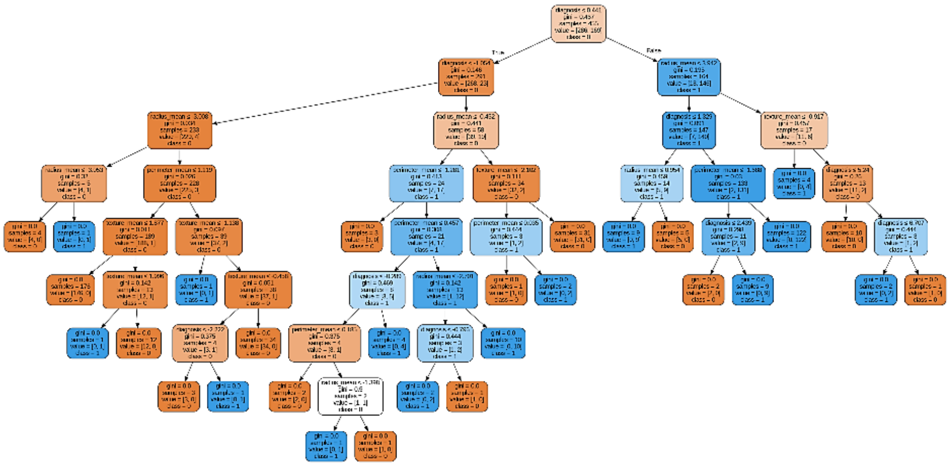


Fig. 9. Decision tree visualization

5) K-Nearest Neighbor (KNN)

This method is a supervised algorithm that relies on the majority vote of its k value. It is considered a non-parametric algorithm, as the classification of the test data points is based on the nearest training data points, without making any assumptions about their underlying distribution. The accuracy of this model can be enhanced as the number of nearest neighbors, determined by the value of k, increases [33]. The k nearest neighbors collectively contribute their votes in support of a new instance that is in close proximity to them. Various methods can be employed to measure distances in this algorithm, with the 'Euclidean' distance being a popular choice. The formula for calculating the Euclidean distance is :

$$d(x, y) = \sqrt{(x - y)^2} \tag{9}$$

In this study we use for k value is 5 which in the context of (9) is the Euclidean distance (d) between two points of x and y, the distances are calculated using the above formula for each neighbor, and the majority class among these neighbors determines the classification of the new instance.

6) Logistic Regression (LR)

This approach is a supervised algorithm utilized for predicting probabilities associated with a target variable [34]. The target variable is built upon linear regression, which evaluates the output and minimizes the error. It employs a complex approximation function, such as the sigmoid or logistic function, to generate predictions. The logistic regression formula is employed to model the probability of a binary outcome (0 or 1) [35] based on the dataset., models predict the probability using the logistic function. The logistic function is defined as:

$$P(Y = 1) = \frac{1}{1 + e - (b_0 - b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)} \tag{10}$$

where in (10), $P(Y = 1)$ is the probability of the positive class (1), e is the base of the natural logarithm, b_0 is a intercept term, and $b_1, b_2, \dots, b_k$ is the coefficients of the predictor variables $x_1, x_2, \dots, x_k$. The parameter C in logistic regression represents the inverse of the regularization strength.

7) Naïve Bayes (Gaussian NB)

The algorithm discussed in this study is a fundamental outcome in the fields of probability and statistics. It can be defined as a conceptual framework used for decision-making. In the context of Naive Bayes (NB), the variables are conditionally independent [36]. NB can be employed to analyze data that have direct influence on each other, in order to establish a model. Moreover, NB is also well-suited for ranking multiple databases [37], [38]. In this particular research, the default formulation of the NB equation is presented as follows:

$$P(y \mid x_1, x_2, \dots, x_n) = \frac{P(y)P(x_1|y)P(x_2|y) \dots P(x_n|y)}{P(\,x_1), P(x_2), \dots, P(x_n)} \tag{11}$$

where in (11), $P(y \mid x_1, x_2, \dots, x_n)$ is the posterior probability of class y given features of x, $P(y)$ is the prior probability of class $y, P(x_i|y)$ is the likelihood of feature x_i given class y, and $x_1, x_2, \dots, x_n$ are the features. The Naïve bayes classifier then predict the class with the highest posterior probability.

3. RESULTS AND DISCUSSION

After preprocessing the data, the performance of the classifier is visually represented using various performance metrics. The preprocessing of the dataset involves replacing missing values and extracting the minimum and maximum values. Additionally, data scaling and dimensional reduction using Principal Component Analysis (PCA) are applied to all machine learning algorithms utilized in this study. Confusion matrix and Receiver Operating Characteristic (ROC) curve are also employed for evaluating the performance

3.1. Performance Measurement with confussion matrix and ROC curve

We applied seven different methods to the dataset and measured the performance of each model using metrics such as Accuracy, Precision, and Recall. The formulation of these performance metrics is as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{12}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

The meanings of the terms (12), (13), (14) are provided:
TP = True Positive (Correctly classified as having breast cancer) - TN = True Negative (Correctly classified as not having breast cancer) - FP = False Positive (Classified as having breast cancer, but actually they do not have it: Error of type I) - FN = False Negative (Classified as not having breast cancer, but actually they have it: Error of type II) [39], [40]:

Accuracy refers to the number of data points that the machine learning model accurately predicts out of the total data points, and it can be calculated using (12). Precision is the percentage of relevant elements that the model correctly predicts and can be calculated using (13). Meanwhile, Recall is the percentage of relevant elements that are correctly classified by the model out of all relevant elements, and it can be calculated using (14). The results of the performance of each algorithm are shown in Table 2, and the confusion matrix with the x-label representing the test data and the y-label representing the model's predictions is shown in Fig. 10**.**

Table 2. Performance each algorithm based on confussion matrix

Model and	Benign (0)			Malignant (1)		
Accuracy Score	Accuracy	Precision	Recall	Accuracy	Precision	Recall
ANN (96.49%)	96%	96%	99%	96%	98%	93%
SVM:RBF (97.3%)	97%	96%	100%	97%	100%	93%
DT (91.8%)	91%	94%	94%	91%	94%	94%
RF (95.6%)	95%	97%	96%	95%	93%	95%
KNN (93.8)	93%	94%	96%	93%	93%	91%
LR (97.3%)	97%	97%	99%	97%	98%	95%
NB (94.7%)	94%	93%	99%	94%	97%	88%

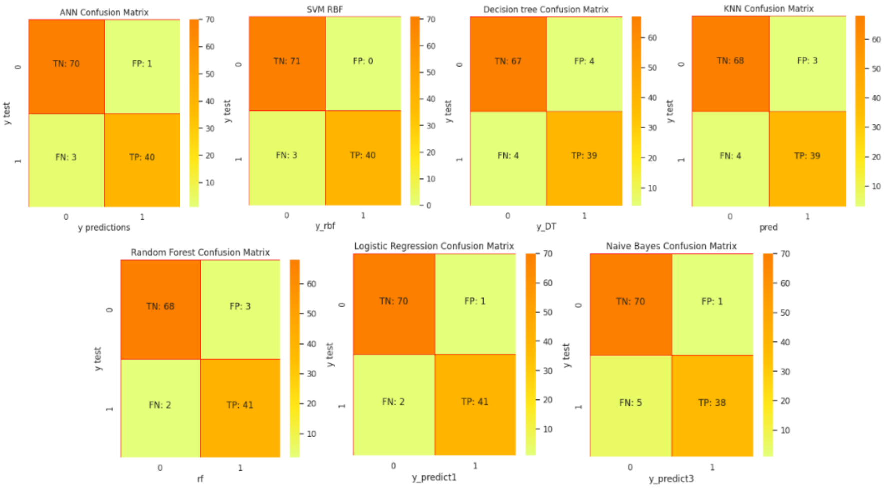


Fig. 10. Confussion matrix of all models

3.2. Performance Proposed Method Comparison with ROC Curve

We have implemented a Receiver Operating Characteristic (ROC), which serves as the conventional tool for model selection and assessment in problems involving the classification of two classes [41]. The ROC curve can be calculated by utilizing the True Positive Rate (TPR) and False Positive Rate (FPR) results obtained from the calculation of the confusion matrix shown in (15) and (16). The TPR and FPR values for each model are presented in Table 3, and the visualization is depicted in Fig. 11, with the x-axis representing the TPR (True Positive Rate) and the y-axis representing the FPR (False Positive Rate) measurement for each algorithm. The formula used is as follows:

$$TPR = \frac{TP}{\text{Actual Positive}} = \frac{TP}{\text{TP} + \text{FN}} \tag{15}$$

$$FPR = \frac{FP}{\text{Actual Negative}} = \frac{FP}{\text{TN} + \text{FP}} \tag{16}$$

It is evident that there is convergence among all the machine learning classifier models. The highest accuracy is achieved by the Support Vector Machine (SVM) and Logistic Regression (LR), both with an accuracy of 97.3%, while the lowest accuracy is observed in the Decision Tree model, which achieves an accuracy of 91.8%. The SVM model exhibits a higher ROC curve and a better FPR value compared to Logistic Regression, despite both models having the same accuracy. This indicates that SVM outperforms Logistic Regression in terms of classification performance. The results of the FPR and TPR for all the methods used are presented.

Table 3. Result of FPR and TPR each algorithm

Method	FPR	TPR
ANN	0.0,0.01408541,1.0	0.0,0.93023256,1.0
SVM RBF	0.0,0.0,1.0	0.0,0.94023256,1.0
DT	0.0,0.05633803,1.0	0.0,0.88697674,1.0
RF	0.0,0.04225352,1.0	0.0,0.94348837,1.0
KNN	0.0,0.04225352,1.0	0.0,0.90697674,1.0
LR	0.0,0.01408451,1.0	0.0,0.95348837,1.0
NB	0.0,0.01419451,1.0	0.0,0.89372093,1.0

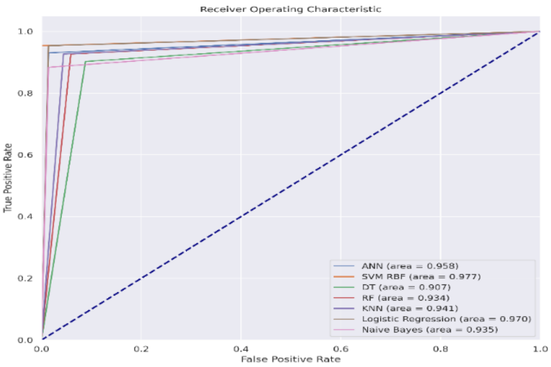


Fig. 11. ROC of all models

3.3. Performance Comparison Previous Study

Comparison of our work with the most related works show in Table 4.

Table 4. Performance comparison with previous study

Author	Dataset	Method	Accuracy
S. Ara *et al.*, 2021 [12]	UCI WBCD, 569 instances, 32 features	SVM	96.5%
Verghese *et al.*, 2021 [13]	UCI WBCD, 569 instances, 32 features	SVM:RBF	94.5%
H. Chiu *et al.*, 2020 [14]	UCI WBCD, 569 instances, 32 features	MLP + SVM	86.9%
Assegie *et al.*, 2020 [15]	Kaggle, 569 instances, 32 features	DT	92.5%
Proposed	**Kaggle, 569 instances, 32 features**	**SVM:RBF & LR**	**97.3%**

4. CONCLUSION

Based on research conducted using datasets obtained from the Kaggle site, we have explored breast cancer classification using feature selection with Principal Component Analysis (PCA) implemented into several supervised machine learning algorithms. The results obtained indicate that SVM and LR achieve the highest accuracy, reaching 97.3%. However, the ROC curve shows that the SVM graph is higher than the LR graph, which can be attributed to the results of the confusion matrix calculation, where the False Positive (FP) value is 0 and the False Positive Rate (FPR) is also 0. A FP and FPR value of 0 is considered favorable, as it signifies that the classification model accurately predicts instances as negative when they do not belong to the class in question. In cases such as breast cancer disease, minimizing FP is crucial. When the FP value is 0, it indicates that the model does not mistakenly classify something as positive when it is actually negative. Consequently, this is considered a positive outcome. Thus, the overall performance of SVM with RBF (Radial Basis Function) kernel and utilizing the c-value selection approach surpasses that of all the machine learning algorithms tested in this study. For future research, several avenues can be explored to further improve accuracy and enhance the classification diagnosis for breast cancer patients. These include applying alternative feature selection methods and optimizers, such as forward selection, to obtain the optimal set of attributes and selecting different features to increase the accuracy value.

REFERENCES

[1] X. Lin, L. Liu, and Z. Yu, “A Generic-Driven Wrapper Embedded With Feature-Type-Aware Hybrid Bayesian Classifier for Breast Cancer Classification,” *IEEE Access*, vol. 7, pp. 119931–119942, 2019, https://doi.org/10.1109/ACCESS.2019.2932505.

[2] R. Roslidar and A. Rahman, “A Review on Recent Progress in Thermal Imaging and Deep Learning Approaches for Breast Cancer Detection,” *IEEE Access*, vol. 8, 2020, https://doi.org/10.1109/ACCESS.2020.3004056.

[3] I. Hirra *et al.*, “Breast Cancer Classification From Histopathological Images Using Patch-Based Deep Learning Modeling,” *IEEE Access*, pp. 24273–24287, 2021, https://doi.org/10.1109/ACCESS.2021.3056516.

[4] H. N. Khan, A. R. Shahid, A. H. Dar, and H. Alquhayz, “Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network,” *IEEE Access*, vol. 7, pp. 165724–165733, 2019, https://doi.org/10.1109/ACCESS.2019.2953318.

[5] A. Hassan, “Performance Analysis of Supervised Classifiers using PCA based Techniques on Breast Cancer,” *Int. Conf. Eng. Emerg. Technol.*, pp. 1–6, 2019, https://doi.org/10.1109/CEET1.2019.8711868.

[6] A. Ameh, M. Abdullahi, S. Balarabe, H. Hassan, and H. Chiroma, “Intelligent Systems with Applications Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer),” *Intelligent Systems with Applications*, vol. 14, 2022, https://doi.org/10.1016/j.iswa.2022.200066.

[7] A. U. L. Haq *et al.*, “Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques,” *IEEE Access*, vol. 9, 2021, https://doi.org/10.1109/ACCESS.2021.3055806.

[8] H. U. A. Chen, K. Mei, Y. Zhou, N. A. N. Wang, and G. Cai, “Auxiliary Diagnosis of Breast Cancer Based on Machine Learning and Hybrid Strategy,” *IEEE Access*, vol. 11, pp. 96374-96386, 2023, https://doi.org/10.1109/ACCESS.2023.3312305.

[9] T. Khater, A. Hussain, S. Member, R. Bendardaf, and S. Member, “An Explainable Artificial Intelligence Model for the Classification of Breast Cancer,” *IEEE Access*, p. 1, 2023, https://doi.org/10.1109/ACCESS.2023.3308446.

[10] A. S. Elkorany, M. Marey, K. M. Almustafa, and Z. F. Elsharkawy, “Breast Cancer Diagnosis Using Support Vector Machines Optimized by Whale Optimization and Dragonfly Algorithms,” *IEEE Access*, vol. 10, pp. 69688–69699, 2022, https://doi.org/10.1109/ACCESS.2022.3186021.

[11] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, “Computer Methods and Programs in Biomedicine Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning,” *Comput. Methods Programs Biomed.*, vol. 223, p. 106951, 2022, https://doi.org/10.1016/j.cmpb.2022.106951.

[12] S. Ara, “Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms,” *International Conference on Artificial Intelligence (ICAI)*, pp. 97–101, 2021, https://doi.org/10.1109/ICAI52203.2021.9445249.

[13] S. L. Verghese, I. Y. I. Liao, T. H. Maul, and S. Y. E. W. Chong, “An Empirical Study of Several Information Theoretic Based Feature Extraction Methods for Classifying High Dimensional Low Sample Size Data,” *IEEE Access,* vol. 9, pp. 69157–69172, 2021, https://doi.org/10.1109/ACCESS.2021.3077958.

[14] H. Chiu and T. S. Li, “Breast Cancer – Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine,” *IEEE Access*, vol. 8, pp. 204309–204324, 2020, https://doi.org/10.1109/ACCESS.2020.3036912.

[15] T. A. Assegie, “An optimized K-Nearest Neighbor based breast cancer detection,” *Journal of Robotics and Control (JRC)*, vol. 2, no. 3, pp. 115–118, 2020, https://doi.org/10.18196/jrc.2363.

[16] A. H. Osman, “An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model,” *IEEE Access*, vol. 8, pp. 39165–39174, 2020, https://doi.org/10.1109/ACCESS.2020.2976149.

[17] F. Azour and A. Boukerche, “An Efficient Transfer and Ensemble Learning Based Computer Aided Breast Abnormality Diagnosis System,” *IEEE Access*, vol. 11, pp. 21199–21209, 2023, https://doi.org/10.1109/ACCESS.2022.3192857.

[18] E. K. Jadoon, F. G. Khan, S. Shah, A. Khan, and M. Elaffendi, “Deep Learning-Based Multi-Modal Ensemble Classification Approach for Human Breast Cancer Prognosis,” *IEEE Access*, vol. 11, pp. 85760–85769, 2023, https://doi.org/10.1109/ACCESS.2023.3304242.

[19] A. R. Beeravolu, S. Azam, and M. Jonkman, “Preprocessing of Breast Cancer Images to Create Datasets for Deep-

CNN," *IEEE Access*, vol. 9, 2021, https://doi.org/10.1109/ACCESS.2021.3058773.

[20] V. Patel, V. Chaurasia, R. Mahadeva, and S. P. Patole, "GARL-Net : Graph Based Adaptive Regularized Learning Deep Network for Breast Cancer Classification," *IEEE Access*, vol. 11, pp. 9095–9112, 2023, https://doi.org/10.1109/ACCESS.2023.3239671.

[21] A. Kumar, "Model Selection for Predicting Breast Cancer using Supervised Machine Learning Algorithms," *IEEE 1st International Conference for Convergence in Engineering (ICCE)*, pp. 320–324, 2020, https://doi.org/10.1109/ICCE50343.2020.9290578.

[22] E. I. D. Alkhaldi and E. Salari, "Ensemble Optimization for Invasive Ductal Carcinoma (IDC) Classification Using Differential Cartesian Genetic Programming," *IEEE Access*, vol. 10, pp. 128790–128799, 2022, https://doi.org/10.1109/ACCESS.2022.3228176.

[23] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, "Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection," *IEEE Access*, vol. 7, pp. 151482–151492, 2019, https://doi.org/10.1109/ACCESS.2019.2947701.

[24] S. Alghunaim, "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context," *IEEE Access*, vol. 7, pp. 91535–91546, 2019, https://doi.org/10.1109/ACCESS.2019.2927080.

[25] A. Saber, M. Sakr, O. M. Abo-seida, A. Keshk, H. Chen, and A. Member, "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," *IEEE Access*, vol. 9, pp. 71194-71209, 2021, https://doi.org/10.1109/ACCESS.2021.3079204.

[26] R. E. Whisnant, "A Novel Data Analytics-derived Metric (Nearest Cluster Distance) Is Easily Implemented in Routine Practice and Correctly Identifies Breast Cancer Cases for Quality Review," *J. Pathol. Inform.*, vol. 13, p. 100005, 2022, https://doi.org/10.1016/j.jpi.2022.100005.

[27] K. Mohammad, M. Uddin, N. Biswas, and S. Tasmin, "Computer Methods and Programs in Biomedicine Update Machine learning-based diagnosis of breast cancer utilizing feature optimization technique," *Comput. Methods Programs Biomed. Updat.*, vol. 3, p. 100098, 2023, https://doi.org/10.1016/j.cmpbup.2023.100098.

[28] J. Ahmad, S. Akram, A. Jaffar, M. Rashid, and M. Bhatti, "Breast Cancer Detection Using Deep Learning : An Investigation Using the DDSM Dataset and a Customized AlexNet and Support Vector Machine," *IEEE Access*, vol. 11, pp. 108386-108397, 2023, https://doi.org/10.1109/ACCESS.2023.3311892.

[29] M. Minnoor and V. Baths, "Diagnosis of Breast Cancer Using Random Forests Diagnosis of Breast Cancer Using Random Forests," *Procedia Comput. Sci.*, vol. 218, pp. 429-437, 2022, pp. 429–437, 2023, https://doi.org/10.1016/j.procs.2023.01.025.

[30] T. I. Rohan, A. B. Siddik, M. Islam, and S. U. Yusuf, "A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost," *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pp. 1-4, 2019, https://doi.org/10.1109/IC4ME247184.2019.9036697.

[31] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," *IEEE Access*, vol. 10, pp. 3284–3293, 2022, https://doi.org/10.1109/ACCESS.2021.3139595.

[32] E. Strelcenia and S. Prakoonwit, "Improving Cancer Detection Classification Performance Using GANs in Breast Cancer Data," *IEEE Access*, vol. 11, pp. 71594–71615, 2023, https://doi.org/10.1109/ACCESS.2023.3291336.

[33] M. Kumari and V. Singh, "Breast Cancer Prediction system," *Procedia Comput. Sci.*, vol. 132, pp. 371–376, 2018, https://doi.org/10.1016/j.procs.2018.05.197.

[34] D. Sharma, R. Kumar, and A. Jain, "Measurement : Sensors Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning," *Meas. Sensors*, vol. 24, p. 100560, 2022, https://doi.org/10.1016/j.measen.2022.100560.

[35] A. Al Bataineh, D. Kaur, and S. M. J. Jalali, "Multi-Layer Perceptron Training Optimization Using Nature Inspired Computing," *IEEE Access*, vol. 10, pp. 36963–36977, 2022, https://doi.org/10.1109/ACCESS.2022.3164669.

[36] A. Bustamam, A. Bachtiar, and D. Sarwinda, "Selecting Features Subsets Based on Support Vector Machine-Recursive Features Elimination and One Dimensional-Naïve Bayes Classifier using Support Vector Machines for Classification of Prostate and Breast Cancer," *Procedia Comput. Sci.*, vol. 157, pp. 450–458, 2019, https://doi.org/10.1016/j.procs.2019.08.238.

[37] A. Algarni, "Convolutional Neural Networks for Breast Tumor Classification using Structured Features," I*nternational Conference of Women in Data Science at Taif University (WiDSTaif),* pp. 11–15, 2021, https://doi.org/10.1109/WiDSTaif52235.2021.9430225.

[38] M. Lopez-Perez, P. Morales-Alvarez, L. A. D. Cooper, R. Molina, and A. K. Katsaggelos, "Deep Gaussian Processes for Classification With Multiple Noisy Annotators. Application to Breast Cancer Tissue Classification," *IEEE Access*, vol. 11, pp. 6922–6934, 2023, https://doi.org/10.1109/ACCESS.2023.3237990.

[39] U. Naseem *et al.*, "An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers," *IEEE Access*, vol. 10, pp. 78242–78252, 2022, https://doi.org/10.1109/ACCESS.2022.3174599.

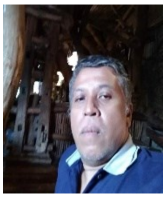
[40] Y. Yari and T. V Nguyen, "Deep Learning Applied for Histological Diagnosis of Breast Cancer," *IEEE Access*, vol. 8, pp. 162432-162448, 2020, https://doi.org/10.1109/ACCESS.2020.3021557.

[41] M. Carrington *et al.*, "Deep ROC Analysis and AUC as Balanced Average Accuracy , for Improved Classifier Selection, Audit and Explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 329–341, 2023, https://doi.org/10.1109/TPAMI.2022.3145392.

BIOGRAPHY OF AUTHORS

Koirunnisa, is an undergraduated student in the Departement of Informatics Engineering, Faculty of Computer Science at Buana Perjuangan University. Her research is centered around data mining, machine learning, and data science. Email: if20.koirunnisa@mhs.ubpkarawang.ac.id.

Amril Mutoi Siregar, received a B.Eng. degree in information technology from STMIK MIC Cikarang in 2008, and an M. Sc degree in information technology from president university in 2016. and Ph, D degrees from the IPB University in 2023. Currently, he is a lecturer of computer science at Buana perjuangan university. His research interests include Artificial intelligence, machine learning, deep learning, data mining, and data science. He can be contacted at email: amril.mutoi99@gmail.com, amrilmutoi@ubpkarawang.ac.id, orcid: https://orcid.org/0000-0001-8746-3283.



Sutan Faisal, is a lecturer in Department of Informatics Engineering, Faculty of Computer Science at Buana Perjuangan University. His research interest include Artificial intelligence, machine learning, and IoT. Email: sutan.faisal@ubpkarawang.ac.id.