


See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/379663149>

Comparative analysis of machine learning models for breast cancer prediction and diagnosis: A dual-dataset approach

Article in Indonesian Journal of Electrical Engineering and Computer Science · June 2024
 DOI: 10.11591/ijeecs.v34.i3.pp2032-2044

CITATIONS
 0

4 authors:




Muhammad Zeerak Awan

Namal University Mianwali

4 PUBLICATIONS 58 CITATIONS

SEE PROFILE




Mirza Zain

Quaid-i-Azam University

1 PUBLICATION 0 CITATIONS

SEE PROFILE

READS
 204




Muhammad Shoaib Arif

Air University of Islamabad

100 PUBLICATIONS 1,017 CITATIONS

SEE PROFILE



Kamaleldin Abodayeh

Prince Sultan University

168 PUBLICATIONS 1,907 CITATIONS

SEE PROFILE

Comparative analysis of machine learning models for breast cancer prediction and diagnosis: A dual-dataset approach

Muhammad Zeerak Awan¹, Muhammad Shoaib Arif^{2,3}, Mirza Zain Ul Abideen⁴,
Kamaleldin Abodayeh²

¹Centre for AI and Big Data, Namal University Mianwali, Mianwali, Pakistan

²Department of Mathematics and Sciences, College of Humanities and Sciences, Prince Sultan University, Riyadh, Saudi Arabia

³Department of Mathematics, Air University, PAF Complex E-9, Islamabad, Pakistan

⁴Department of Biotechnology, Quaid-i-Azam University, Islamabad, Pakistan

Article Info

Article history:

Received Jan 22, 2024

Revised Feb 16, 2024

Accepted Mar 10, 2024

Keywords:

Breast cancer

Data mining

Datasets

Machine learning

Model evaluation

ABSTRACT

Breast cancer is ranked as a significant cause of mortality among females globally. Its complex nature poses principal challenges for physicians and researchers for rapid diagnosis and prognosis. Hence, machine learning algorithms are employed to forecast and identify diseases. This study discusses the comparative analysis of seven machine learning models, e.g., logistic regression (LR), support vector machine (SVM), k-nearest neighbor classifier (KNN), decision tree classifier (DT), random forest classifier (RF), Naïve Bayes (NB), and artificial neural network (ANN) to predict breast cancer using Wisconsin breast cancer and breast cancer datasets. In the Wisconsin breast cancer dataset, KNN depicted 99% accuracy, followed by RF (98%), SVM (96%), NB (96%), LR (96%), ANN (93%), and DT (92%). On the contrary, in the breast cancer (BC) dataset, the highest accuracy was achieved by LR at 83%, and the lowest was achieved by DT (65%), which depicted that the numeric dataset WBC has better accuracy than the breast cancer dataset.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Muhammad Shoaib Arif

Department of Mathematics and Sciences, College of Humanities and Sciences, Prince Sultan University

Riyadh, 11586, Saudi Arabia

Email: marif@psu.edu.sa

1. INTRODUCTION

Breast cancer continues to be a significant global health issue and a leading cause of death among women globally. The cause of occurrence involves genetic and environmental factors [1]. 25% of hereditary cases are due to mutations affecting high penetrant genes, e.g., HER2, BRCA1, BRCA2, TP53, PTEN, CDH1, and STK11, and moderate penetrant genes, e.g., CHEK2, BRIP1, ATM, and PALB2 [2]. GLOBOCAN 2020 data depicts the estimated number of new cases in women as 2.3 million, with 6.9% mortality over the five years respectively [3]. According to the LLR (log-linear Regression) model, women above the age of 75 are more likely to get Breast Cancer, followed by women between the ages of 55 and 64 [4]-[6].

The complex phenomenon of a varied nature requires accurate and timely diagnostic and prognostic approaches. Thus, using machine learning (ML) approaches in the medical sector is essential to help forecast by analyzing and configuring data. Because of their strong classification results, many researchers utilize these algorithms to address complex problems [7]. Data mining using machine learning is being utilized in the clinical domains to arrange and comprehend extensive data more readily using a computer-assisted detection (CAD) system that employs machine learning techniques to give reliable Breast Cancer diagnosis [8].

Luckily, most Breast Cancer data is open source and available on data repositories for the medical research community, e.g., the Wisconsin Breast Cancer Dataset on Kaggle and the Breast Cancer Dataset on the UCI Machine Learning Repository. So, researchers are using them for analysis and prediction by applying Machine Learning algorithms [9]. So, researchers and physicians are using machine learning algorithms to develop effective predictive breast cancer detection and prognosis models. The Wisconsin (diagnostic) dataset was used by Hossin *et al.* [10] to detect breast cancer using eight machine learning algorithms, namely "logistic regression, random forest, K-nearest neighbours, decision tree, ada boost, support vector machine, gradient boosting, and Gaussian Naïve Bayes." Comparing the results of the inquiry to other similar studies in the literature, it was found that the LR approach attained a maximum accuracy of 99.12% among the eight algorithms. Looking at the latest research by Sakib *et al.* [11], the research compared various machine-learning models using the Breast Cancer Wisconsin (Diagnostic) dataset. The models checked were: support vector machine, decision tree, logistic regression, random forest, K-Nearest Neighbour, and deep learning. After some fine-tuning, the results showed that the random forest model performed the best. It exceeded all other models examined with an accuracy of 96.66% and an F1-score of 0.963. Comparing models using the Wisconsin dataset, Ak *et al.* also found an LR accuracy of 99.12% [12].

To improve, algorithms must be enhanced, and various datasets need more models for better diagnosis and prediction. There is still a lot of improvement needed even after so much progress. The effectiveness of various machine-learning classifiers, such as logistic regression, support vector machine, k-nearest neighbour, decision tree, random forest, Naïve Bayes, and artificial neural network, in the prediction and diagnosis of breast cancer is examined in this study through computational analysis. The aim is to meet the urgent requirement for accurate and timely assessments. The examined datasets, specifically the wisconsin breast cancer (WBC) and larger Breast Cancer datasets, cover different aspects of this widespread health issue. Given the inherent problems in these datasets, such as uneven class distributions and missing values, our study takes great care in pre-processing the data to guarantee that the model training is robust and unbiased.

2. METHOD

This section describes the approach to assessing ML algorithms' effectiveness through analyzing and preparing the data. The study was conducted using the following steps shown in Figure 1. We divided our research study into five sections "Data Collection, Data Pre-processing, Exploratory Data Analysis, Model Selection, and Model Evaluation". These sections will be briefly explaining and discussed one by one.

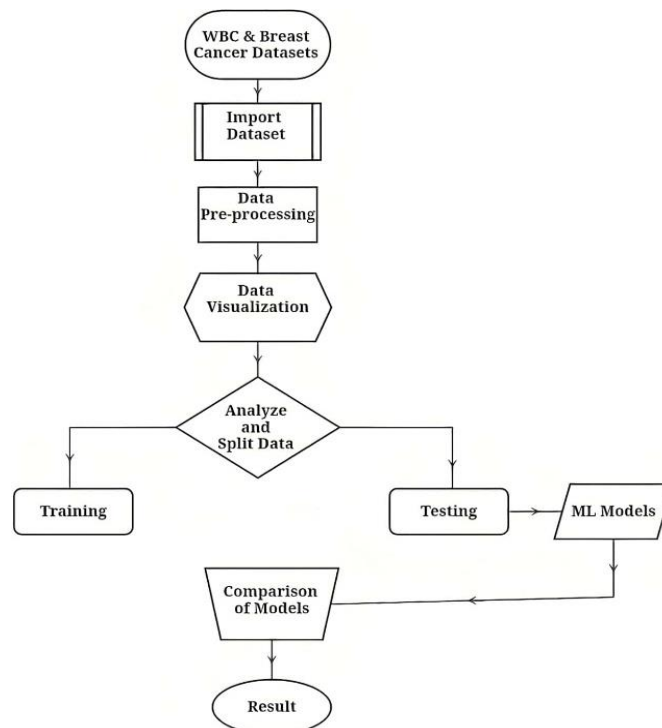


Figure 1. Proposed model by using WBC and breast cancer datasets

2.1. Data collection

Breast cancer and WBC Dataset are used in this paper. You can obtain both datasets from the UCI machine learning repository. There are 699 entities and 11 features in the WBC dataset, 458 of which are benign cases and 241 of which are malignant cases. "Clump thickness, Uniformity of cell size, Uniformity of cell shape, and other similar metrics" are examined throughout the WBC dataset. Bond at the margin, entire cell size of an epithelial layer just the cores, Genomic DNA, regular nucleoli, class is our dependent or target variable, while mitoses are independent variables [13]-[15]. Hence, data pre-processing is essential for this dataset. Our second dataset, called the Breast Cancer dataset, comes from a digital image of a breast tumor that was removed using a fine needle suction. The prognosis (malignant or benign) is recorded in the target feature. A total of 286 entities and 10 nominal and linear features makes up the dataset. The features include things like "age, menopause, tumor-size, inv-nodes, node caps, deg malign, breast, breast quad, irradiate and class(output)", along with 201 no-recurrence events and 85 recurrence events.

2.2. Data pre-processing

First, the dataset was thoroughly examined to make sure there were no instances of missing data, and the results showed that there were no missing values. This demonstrated the dataset's dependability and integrity. WBC (Table 1) and breast cancer (Table 2) datasets were pre-processed to remove missing values and irrelevant information. The feature Bare Nuclei status value was missing in the WBC dataset for 16 records. In the breast cancer dataset, eight absent values were recorded. In both datasets, missing values are indicated by "?". We handle missing data of WBC data by replacing the missing values "?" with 'nan', and then we input the missing values with the median. In pre-processing, we select the median rather than the mean or mode for missing values to lessen the effect of outliers.

On the other hand, we choose data transformation for the breast cancer dataset over data cleaning since much of our data is labeled in the form of words. Therefore, we use the sklearn-preprocessing library label encoder to turn our input into numeric form so that it may be machine-readable. After the pre-processing phase, we analyze both datasets. Table 1 summarizes the results of the features data explanatory analysis for the WBC dataset. Table 2 summarizes the results for the breast cancer dataset. They present a statistical overview of both datasets.

Table 1. Statistical summary of the WBC dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
Clump thickness	699.0	4.417740	2.815741	1.0	2.0	4.0	6.0	10.0
Uniformity of cell size	699.0	3.134478	3.051459	1.0	1.0	1.0	5.0	10.0
Uniformity of cell shape	699.0	3.207439	2.971913	1.0	1.0	1.0	5.0	10.0
Marginal adhesion	699.0	2.806867	2.855379	1.0	1.0	1.0	4.0	10.0
Single epithelial cell size	699.0	3.216023	2.214300	1.0	2.0	2.0	4.0	10.0
Bare nuclei	699.0	3.486409	3.621929	1.0	1.0	1.0	5.0	10.0
Bland chromatin	699.0	3.437768	2.438364	1.0	2.0	3.0	5.0	10.0
Normal nucleoli	699.0	2.866953	3.053634	1.0	1.0	1.0	4.0	10.0
Mitoses	699.0	1.589413	1.715078	1.0	1.0	1.0	1.0	10.0
Class	699.0	2.689557	0.951273	2.0	2.0	2.0	4.0	4.0

Table 2. Statistical summary of breast cancer dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
Age	286.0	2.664336	1.011818	0.0	2.0	3.0	3.0	5.0
Menopause	286.0	1.073427	0.986680	0.0	0.0	2.0	2.0	2.0
Tumor-size	286.0	4.062937	2.151187	0.0	3.0	4.0	5.0	10.0
Inv-nodes	286.0	1.073427	1.935321	0.0	0.0	0.0	1.0	6.0
Node-caps	286.0	0.251748	0.495149	0.0	0.0	0.0	0.0	2.0
Deg-malign	286.0	1.048951	0.738217	0.0	1.0	1.0	2.0	2.0
Breast	286.0	0.468531	0.499883	0.0	0.0	0.0	1.0	1.0
Breast-quad	286.0	1.793706	1.103151	0.0	1.0	1.0	2.0	5.0
Irradiate	286.0	0.237762	0.426459	0.0	0.0	0.0	0.0	1.0
Class	286.0	0.297203	0.457828	0.0	0.0	0.0	1.0	1.0

2.3. Exploratory data analysis

Exploratory data analysis helps to investigate the critical decision for further processing to build data modeling. We use Python libraries Plotly, seaborn, and Matplotlib to plot a scatter plot and heatmap of both datasets. Multi-variable scatter plots help display interactions between more than two variables in a single plot, whereas heatmaps, which synthesize data and present it graphically, give a practical visual overview of

information. In the scatter plot and heatmap (Figure 2 and Figure 3) of WBC dataset, we can see that the variables uniformity of cell size, uniformity of cell shape, and bare nuclei are highly correlated with the target variable "class."

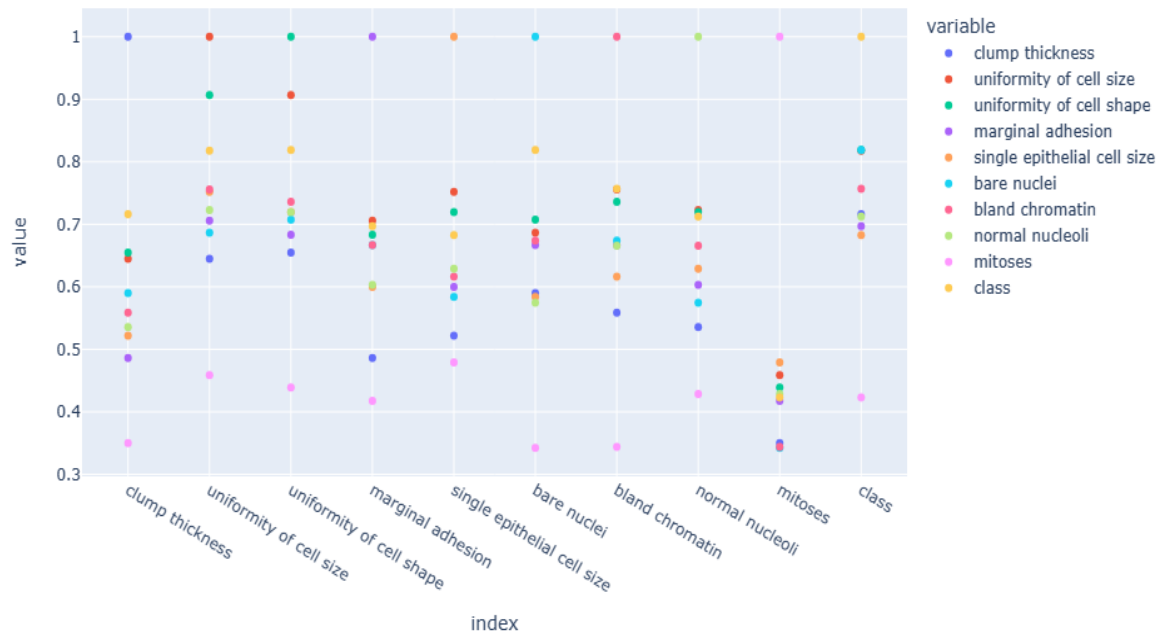


Figure 2. Scatter plot of WBC dataset

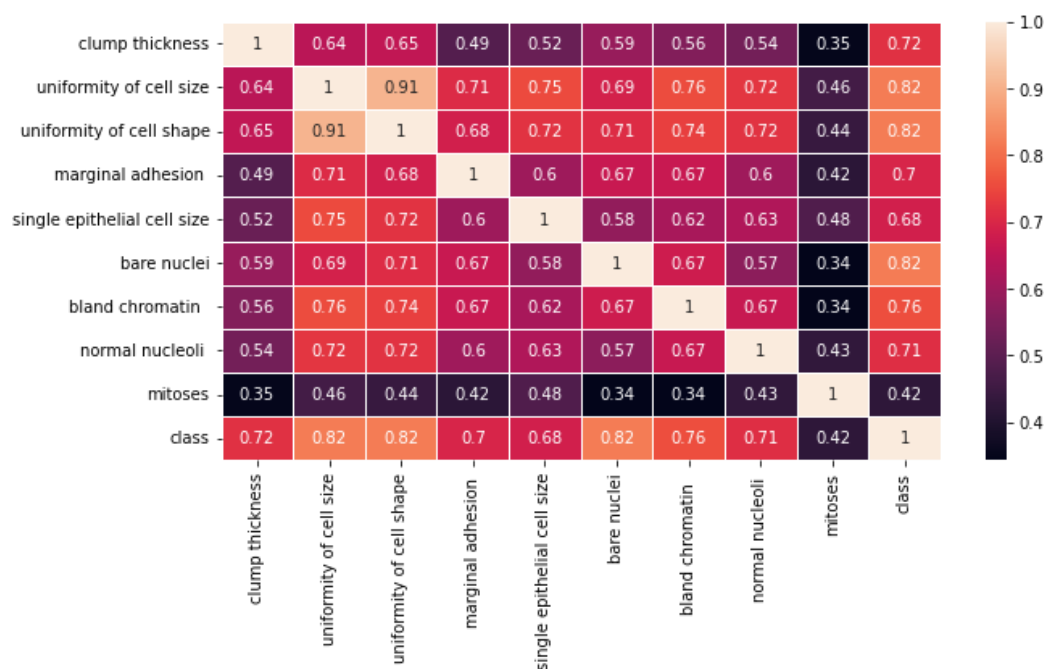


Figure 3. Heatmap of WBC dataset

We can observe that none of the independent variables in the Breast Cancer dataset have a significant relationship with the target variable, "class." The heatmap (Figure 4 and Figure 5) indicates that, among other variables, the variable "inv-nodes" has a strong positive correlation with the target variable "class."

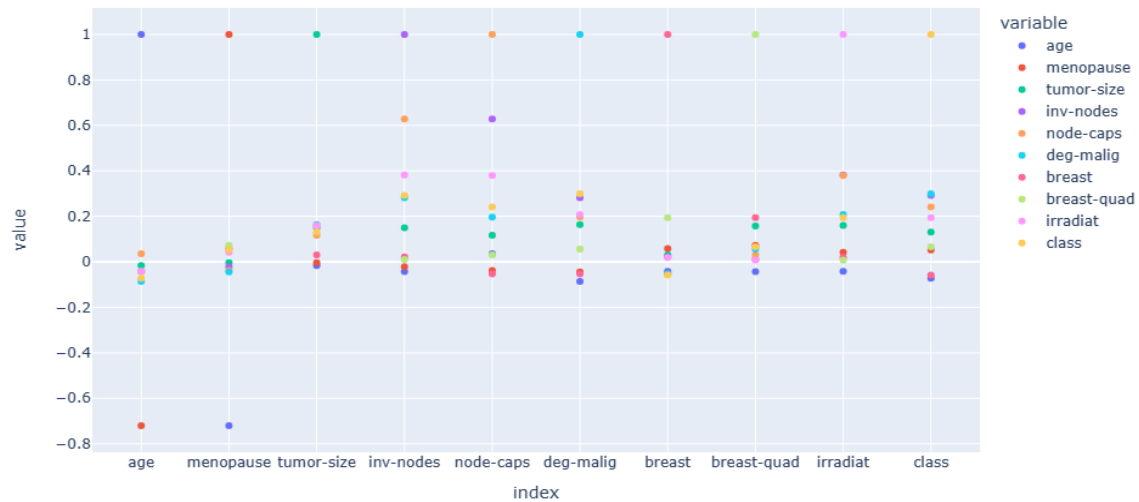


Figure 4. Scatter plot of breast cancer dataset

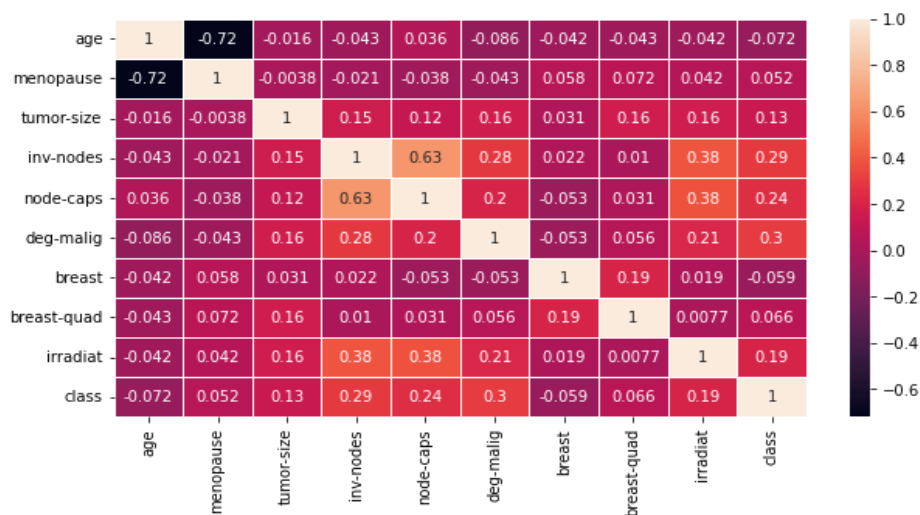


Figure 5. Heatmap of breast cancer dataset

2.4. Model selection

Seven ML algorithms, "logistic regression (LR), support vector machine (SVM), K-nearest neighbors' classifier (KNN), decision tree classifier (DT), random forest classifier (RF), Naïve Bayes (NB), and artificial neural network (ANN)" were selected for WBC and breast cancer datasets. These machine-learning algorithms were selected based on how well they handled intricate interactions between variables. Train test split is a model validation method that predicts a model's performance on new data. Train test split of our models: i) Using the WBC dataset, the training set contains 80% of the data, whereas the testing set contains 20%. ii) Using the breast cancer dataset, we put 70% of the data into the training set and 30% into the testing set.

2.4.1. Logistic regression

Logistic regression is a machine learning algorithm most frequently employed under supervised learning [16]. It's used to predict categorical dependent variables (0 or 1, yes or no, true or false) from a set of independent variables [17]. The prediction is made by converting the unobserved data to the built-in logit function. Predict 0 and 1 for the logistic regression modeling utilizing the standard logistic function and linear probability function (1).

$$p(x) = \frac{e^{ax+b}}{1+e^{ax+b}} = \frac{1}{1+e^{-ax+b}} \quad (1)$$

Logistic regression gives linear classifier results, predicting $y = 1$ when $p \geq \frac{1}{2}$ and $y = 0$ when $< \frac{1}{2}$. Logistic function in general (2).

$$f(x) = \frac{1}{1+e^{-ax+b}} \quad (2)$$

2.4.2. Support vector machine

Support vector machines, one of the most well-known supervised learning algorithms, are used in various biological applications, including medical data classification [18]. SVMs identify the best hyperplane in a dataset to divide the multiple classes.

$$y = \text{sign}(wx_i - b) \quad (3)$$

Where the sign is an operator, any input of positive values yields a result of 1, and any input of negative values yields a result of -1, known as a hard margin.

$$\text{If } y = 1, wx_i - b \geq 1 \quad (4)$$

$$\text{If } y = -1, wx_i - b \leq -1 \quad (5)$$

2.4.3. K-nearest neighbors' classifier

It's a straightforward and popular supervised machine learning algorithm. By employing a distance metric such as Manhattan, Minkowski, or Euclidean distance, the classification of an observation can be predicted by considering the classes of its k nearest neighbors. In our study, we use Minkowski distance and set n-neighbors = 5. The Minkowski distance metric equations are as (6).

$$d\text{Minkowski}(x_i, x_j) = [\sum_{k=1}^d |x_{ik} - x_{jk}|^p]^{1/p} \quad (6)$$

These equations express the kth attributes of x_i and x_j in a d-dimensional space, respectively, by x_{ik} and x_{jk} . A data point's class is determined by taking the majority class among its k closest neighbors, which are found using these distance measures [19]-[21].

2.4.4. Decision tree classifier

Decision tree is one of the most used classification methods. The classifier is tree-structured, where the internal nodes correspond to the dataset's properties, and each leaf node signifies the classification result. Decision trees classify depending on the values of the features. The information gain approach determines which aspect of the dataset provides the most information, designates that as the root nodes, and so on until they can classify each dataset entity. Using the (7), you can compute it.

$$\text{Information Gain} = \text{Entropy}(S) - [(\text{Weighted Avg}) * \text{Entropy (Each feature)}] \quad (7)$$

Entropy is a metric used to quantify the impurity in a particular characteristic. It describes data randomness. Calculating entropy is as (8).

$$\text{Entropy}(S) = -P(0)\log_2 P(0) - P(1)\log_2 P(1) \quad (8)$$

Where,

- S is the total number of samples.
- P(0) is probability of Benign.
- P(1) is a probability of malignant.

2.4.5. Random forest classifier

The random forest algorithm, which was put out by Breiman [22], is an ensemble learning technique. Random forest uses predictions from several decision trees rather than just one, basing its forecast of the result on the majority votes of predictions. The random forest classifier uses an average of the predictions made by each decision tree on several subsets of the input dataset to improve the overall accuracy of the predictions. The algorithm determines which attribute at each node of the trees offers the most significant decrease in uncertainty by calculating the entropy of each attribute [23].

$$\text{Entropy}(p) = -[\sum_{n=1}^N P_n \log_2(P_n)] \quad (9)$$

The log to base 2 of the probability of a category (P_n) represents the uncertainty or impurity.

2.4.6. Naïve Bayes

Naïve Bayes is one of the most efficient yet straightforward classifiers. It is based on the Bayes theorem, which describes how event probability is calculated using prior knowledge of circumstances that could be pertinent to the occurrence.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (10)$$

A and B stand for separate occurrences in this equation. The likelihood that event A will occur, given that event B has occurred, is determined by this equation. The likelihood of identifying a specific collection of feature values in response to a class label can be determined using the Gaussian probability density function [24].

2.4.7. Artificial neural network

A branch of AI inspired by biological principles and designed to mimic brain function is known as artificial neural networks (ANNs) (Figure 6). The brain's architecture is based on biological neural networks. On the other hand, a computer network based on biological neural networks is frequently called an artificial neural network.

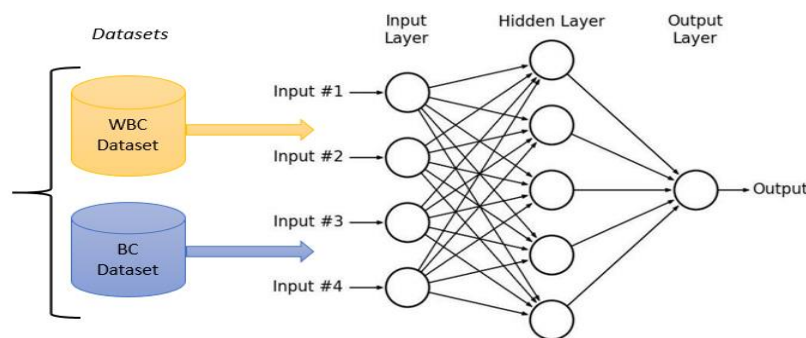


Figure 6. General structure of ANN

2.5.8. ROC-AUC curve

The receiver operating characteristic (ROC) curve illustrates the relationship between the true and false positive rates when the categorization thresholds change. The area under the curve (AUC) aggregates performance across all thresholds on a 0 to 1 scale, 1 being perfect classification. ROC-AUC provides a threshold-independent visual and numerical evaluation of classification model performance. Higher AUC signals better overall diagnostic ability. We employ a two-layer network with 5 and 2 neurons, an MLP classifier with the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) solver, L2 regularization (α), and a particular random state. Next, the model is trained using the fit technique on the training set of data [25].

3. MODEL EVALUATION

The measurement of the model's performance can be analyzed based on a confusion matrix (Table 3). That helps us to find accuracy, precision, sensitivity or recall, specificity, FP rate, FN rate, F1-score, and ROC-AUC curve.

Table 3. The confusion matrix

Actual	Predicted	
	Predicated positive	Predicated negative
Actual positive	True positive (TP)	False negative (FN) Type II Error
Actual negative	False positive (FP) Type I Error	True Negative

3.1. Confusion matrix

The confusion matrix or error matrix is a specific table used to measure the performance of our seven algorithms. In the confusion matrix, model predictions are categorized into four groups: True Positives (TP) are positive situations that are accepted correctly, while true negatives (TN) are negative cases that are rejected correctly. When positive occurrences are incorrectly labelled as negative, this is known as a false negative (FN), whereas negative occurrences are erroneously labelled as a false positive (FP). In order to accurately evaluate performance, this analysis provides a helpful comprehension of various classification errors and accomplishment categories.

Cases that are appropriately labelled as true and true are known as true positives (TP). This indicates that the patient has cancer, as predicted by the model. In cases where the erroneous classification is correct, we say that it is a true negative (TN). This finding confirms the model's prediction that the patient does not have cancer. When something is incorrectly labelled as true when it is actually false, this is called a false positive (FP). It indicates that the patient's condition is not cancerous, while the model indicates otherwise. When true cases are mistakenly labelled as false, this is known as a false negative (FN). In other words, the patient actually has cancer, even though the model says otherwise.

3.2. Accuracy

Total number of correct predictions. This gives the overall rate of accurate diagnoses for both positive and negative cases.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (11)$$

3.3. Precision

The ratio of true positives (TP) to the total number of positive predictions made by the model. It measures the rate of correct positive predictions out of all cases where the model predicted the positive class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (12)$$

3.4. Sensitivity

Sensitivity measures the rate of correct positive predictions made out of all positive samples. It indicates the likelihood that the model correctly diagnoses an actual positive case.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (13)$$

3.5. Specificity

Specificity measures the rate of correct negative predictions made out of all negative samples. It indicates the likelihood that the model correctly rules out an actual negative case.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (14)$$

3.6. False positive rate

The false positives (FP) ratio is the total number of actual negative cases. It indicates the likelihood that the model incorrectly classifies an actual negative case as positive.

$$\text{FP rate} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (15)$$

3.7. False negative rate

The false negatives (FN) ratio to the total number of positive cases. It indicates the likelihood that the model incorrectly rules out an actual positive case as negative.

$$\text{FN rate} = \frac{\text{FN}}{\text{TP} + \text{FN}} \quad (16)$$

3.8. F1-score

The F1-score combines precision and sensitivity from the confusion matrix into a harmonic mean, balancing positive predictive value and true positive rate. It provides a singular metric of a model's overall accuracy, capturing false positives and negatives. This consolidated measure is especially useful for imbalanced classification cases where both precision and sensitivity matter.

$$F1 - score = \frac{(2 \times (Precision \times Recall))}{(Precision + Recall)} \quad (17)$$

4. RESULTS AND DISCUSSION

Breast cancer prediction can be enhanced using machine learning, which is vital for early diagnosis and improved prognosis. This study analyzes tabular datasets to demonstrate accurate classification performance by machine learning models for enhanced breast cancer diagnosis and prognosis without relying on complex neural networks. Seven algorithms, "logistic regression, support vector machine, K-NN, decision tree, random forest, Naïve Bayes, and neural networks," were evaluated on two distinct breast cancer datasets.

4.1. Results using the WBC dataset

We apply seven machine learning algorithms on the WBC dataset to see how the models perform. We analyze models' performance based on accuracy, precision, and more (Table 4). We analyze and find that the K-nearest neighbor model outperforms with 99% accuracy in our predictive analysis. The random forest model is the second-best performer with 98% accuracy. Support vector machine, Naïve Bayes, and logistic regression all perform well, achieving 96% accuracy. The artificial neural network achieves 93% accuracy, and the Decision tree is the lowest performer with 92% accuracy.

The area under the ROC curve or AUC-ROC measures a classification model's effectiveness and potential classification thresholds. The categorization threshold changes from 0 to 1, illustrating the trade-off between genuine and false positive rates. A perfect model has an AUC-ROC value of 1, whereas a mediocre model has an AUC-ROC value of 0.5 [26]. The ROC curves in Figures 7 and 8 demonstrate the relationship between the true positive rate and the false positive rate. In both the figures, since curves of K-nearest neighbor, support vector machine, Naïve Bayes, and logistic regression are closely following the left and the top border of ROC space, it can be said that these classifiers are comparatively more accurate than decision tree and artificial neural network for the data set under study for this research

Table 4. Performance of algorithms using the WBC dataset

ML Models	Accuracy	Specificity	Precision	Recall	F1-score	FP rate	FN rate	AUC
KNN	0.99	0.98	0.99	0.99	0.99	0.02	0.01	0.98
RF	0.98	0.98	0.98	0.98	0.98	0.02	0.02	0.97
SVM	0.96	0.98	0.97	0.95	0.96	0.02	0.04	0.95
NB	0.96	0.92	0.96	0.97	0.96	0.08	0.01	0.98
LR	0.96	0.98	0.97	0.94	0.95	0.02	0.05	0.94
ANN	0.93	0.95	0.94	0.90	0.92	0.05	0.07	0.90
DT	0.92	0.90	0.92	0.90	0.91	0.09	0.07	0.90

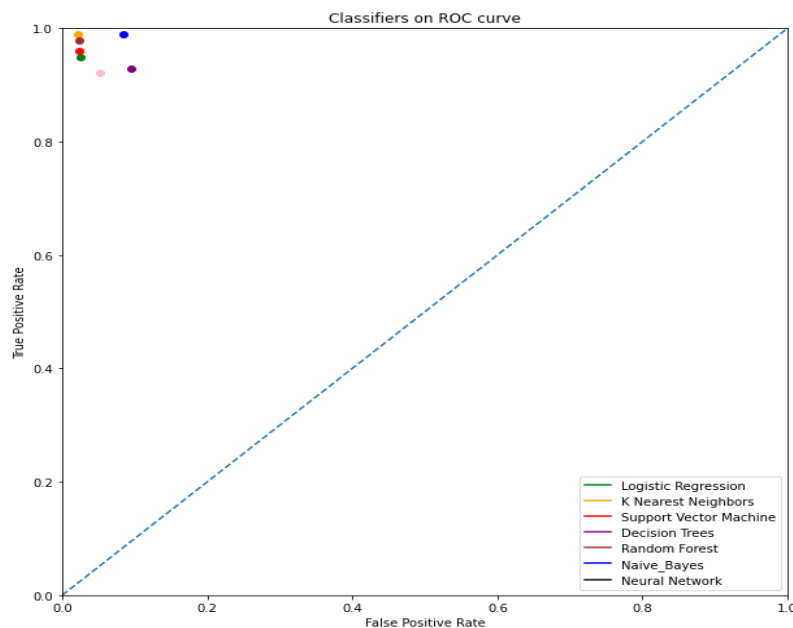


Figure 7. Classifiers as points on ROC curve

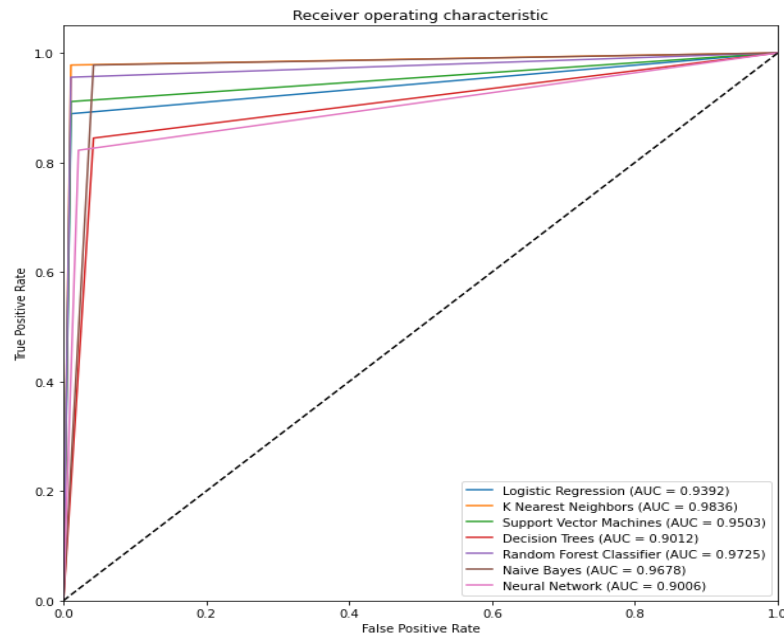


Figure 8. ROC-AUC for different classifiers

4.2. Results using the breast cancer dataset

We apply seven machine learning algorithms on the Breast Cancer dataset to see how the models perform. We analyze models' performance based on accuracy, precision, and more (Table 5). The logistic regression model outperforms 83% accuracy in our predictive analysis. Artificial neural network is the second-best performer with 81% accuracy. Naïve Bayes is the third-best performer with 78% accuracy, while the Support vector machine is the fourth-best performer with 74% accuracy. Random forest is the fifth-best performer with 70% accuracy. K-nearest neighbor performs quite less, achieving 67% accuracy, and the Decision tree is the lowest performer overall with 65% accuracy.

Plotting false positive rates on the x-axis compared to true positive rates are receiver operator characteristic (ROC) curves (as seen in Figures 9 and 10). Given that the curves of Naïve Bayes, support vector machine, artificial neural network, and logistic regression closely follow the top and left borders of ROC space in both figures, it can be concluded that these classifiers are relatively more accurate than Decision Tree and K-nearest neighbor for the data set being studied for this research.

Pre-processing the Wisconsin dataset by imputing missing values with median enhanced data robustness. The nonparametric simplicity of k-NN achieved the highest accuracy of 99%, excelling in complex nonlinear relationships. Conversely, the decision tree attained minimum accuracy, likely due to overfitting tendencies. Meanwhile, label encoding enabled compatibility with ML algorithms for the categorical Breast Cancer dataset. However, overlooking intricate inter-category relationships can limit performance. Advanced encodings better retain feature correlations vital for cancer data. Logistic regression attained a maximum accuracy of 83% by effectively modeling binary tumor outcomes. Again, the decision tree underperformed.

Overall, nonlinear models suit the Wisconsin data better. However, linear models fit the Breast Cancer data well for tumor classification tasks. Our findings suggest that tailored pre-processing and model selection customized to problem and dataset intricacy are key to optimizing accuracy for breast cancer diagnosis and prognosis using ML.

Table 5. Performance of algorithms using breast cancer dataset

ML Models	Accuracy	Specificity	Precision	Recall	F1-score	FP rate	FN rate	AUC
LR	0.83	0.85	0.84	0.72	0.74	0.15	0.17	0.71
ANN	0.81	0.83	0.82	0.70	0.72	0.17	0.19	0.69
NB	0.78	0.61	0.73	0.72	0.73	0.40	0.16	0.72
SVM	0.74	0.55	0.68	0.66	0.67	0.45	0.20	0.66
RF	0.70	0.45	0.63	0.63	0.63	0.54	0.21	0.62
KNN	0.67	0.33	0.53	0.52	0.51	0.66	0.27	0.52
DT	0.65	0.40	0.60	0.61	0.60	0.60	0.21	0.60

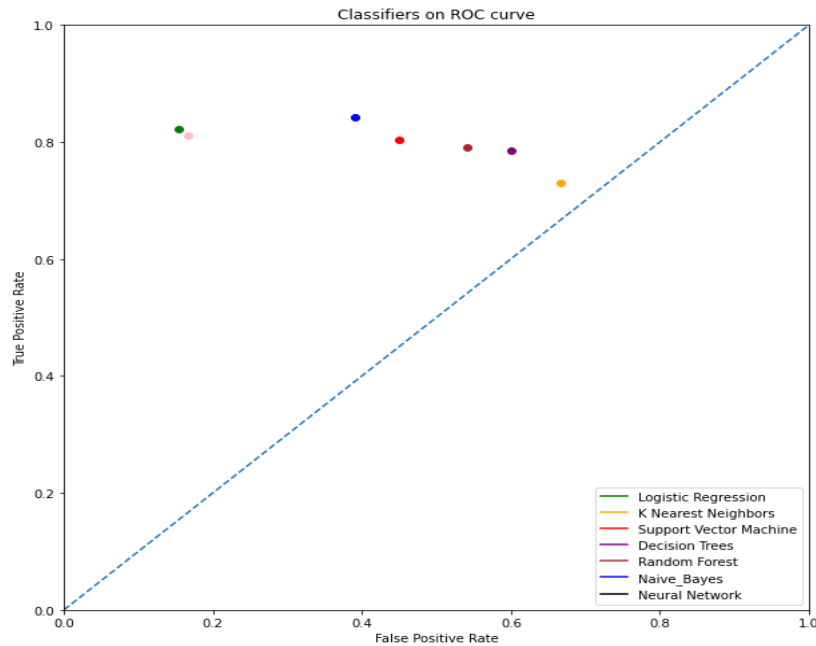


Figure 9. Classifiers as points on ROC curve

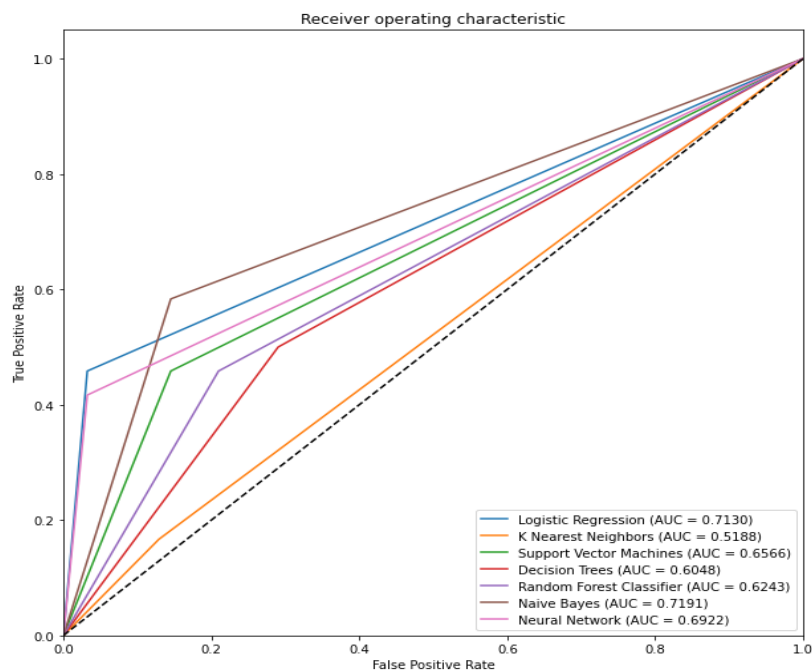


Figure 10. ROC-AUC for different classifiers

5. CONCLUSION

This research analyzed seven machine learning models, "logistic regression, support vector machines, K-Nearest Neighbor, decision trees, random forests, Naïve Bayes, and neural networks," to assess efficacy in breast cancer diagnosis and prognosis using two distinct datasets. Careful pre-processing, like balancing and imputation, enabled robust evaluation. The Wisconsin dataset saw top performance from KNN at 99% accuracy, leveraging adaptability to complex relationships. However, Logistic Regression better handled the categorical tumor predictions on the broader dataset. Varying results emphasize the influence of nuanced dataset differences on model effectiveness for this critical domain. Continued advancement of tailored machine learning techniques can substantially augment expert detection and decision-making

regarding interventions. Yet translating promising techniques into clinical practice requires addressing interpretability and context-awareness to produce equitable and accurate AI tools that earn practitioner trust. Overall, this analysis highlights that machine learning is maturing rapidly but deep learning algorithms can be applied for better predictions.

ACKNOWLEDGEMENTS





The authors wish to express their gratitude to Prince Sultan University for facilitating the publication of this article through the Theoretical and Applied Sciences Lab.

REFERENCES





- [1] A. Rudolph, J. Chang-Claude, and M. K. Schmidt, "Gene-environment interaction and risk of breast cancer," *British Journal of Cancer*, vol. 114, no. 2, pp. 125–133, Jan. 2016, doi: 10.1038/bjc.2015.439.
- [2] S. Shiovitz and L. A. Korde, "Genetics of breast cancer: A topic in evolution," *Annals of Oncology*, vol. 26, no. 7, pp. 1291–1299, Jul. 2015, doi: 10.1093/annonc/mdv022.
- [3] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.
- [4] S. Łukasiewicz, M. Czezelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, "Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—An updated review," *Cancers*, vol. 13, no. 17, p. 4287, Aug. 2021, doi: 10.3390/cancers13174287.
- [5] Z. Ullah, M. N. Khan, Z. U. Din, and S. Afaq, "Breast Cancer Awareness and Associated Factors Amongst Women in Peshawar, Pakistan: A Cross-Sectional Study," *Breast Cancer: Basic and Clinical Research*, vol. 15, p. 117822342110253, Jan. 2021, doi: 10.1177/11782234211025346.
- [6] M. Chadha *et al.*, "Optimal adjuvant therapy in older (≥ 70 years of age) women with low-risk early-stage breast cancer," *npj Breast Cancer*, vol. 9, no. 1, p. 99, Dec. 2023, doi: 10.1038/s41523-023-00591-6.
- [7] S. Zaheer, N. Shah, S. A. Maqbool, and N. M. Soomro, "Estimates of past and future time trends in age-specific breast cancer incidence among women in Karachi, Pakistan: 2004-2025," *BMC Public Health*, vol. 19, no. 1, p. 1001, Dec. 2019, doi: 10.1186/s12889-019-7330-z.
- [8] Q. Min *et al.*, "Differential diagnosis of benign and malignant breast masses using diffusion-weighted magnetic resonance imaging," *World Journal of Surgical Oncology*, vol. 13, no. 1, p. 32, Dec. 2015, doi: 10.1186/s12957-014-0431-3.
- [9] D. A. Omondigabe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," *IOP Conference Series: Materials Science and Engineering*, vol. 495, no. 1, p. 012033, Jun. 2019, doi: 10.1088/1757-899X/495/1/012033.
- [10] M. M. Hossin, F. M. Javed Mehedi Shamrat, M. R. Bhuiyan, R. A. Hira, T. Khan, and S. Molla, "Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 4, pp. 2446–2456, Aug. 2023, doi: 10.11591/eei.v12i4.4448.
- [11] S. Sakib, N. Yasmin, A. K. Tanzeem, F. Shorna, K. Md. Hasib, and S. B. Alam, "Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms," in *Lecture Notes in Electrical Engineering*, vol. 844, 2022, pp. 703–717. doi: 10.1007/978-981-16-8862-1_46.
- [12] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthcare (Switzerland)*, vol. 8, no. 2, p. 111, Apr. 2020, doi: 10.3390/healthcare8020111.
- [13] S. S. Olofintuyi, "Breast Cancer Detection With Machine Learning Approach," *Fudma Journal of Sciences*, vol. 7, no. 2, pp. 216–222, Apr. 2023, doi: 10.33003/fjs-2023-0702-1392.
- [14] L. R. Borges, "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection," *Proceedings of XI Workshop de Visão Computacional*, no. December, pp. 15–19, 2015, [Online]. Available: <https://www.researchgate.net/publication/311950799>
- [15] M. Saleh *et al.*, "Enhancing Breast Cancer Detection and Classification Using Advanced Multi-Model Features and Ensemble Machine Learning Techniques," *Life*, vol. 13, no. 10, pp. 1–20, 2023.
- [16] O. Karasoy and S. Ballı, "Spam SMS Detection for Turkish Language with Deep Text Analysis and Deep Learning Methods," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 9361–9377, Aug. 2022, doi: 10.1007/s13369-021-06187-1.
- [17] F. S. de Menezes, G. R. Liska, M. A. Cirillo, and M. J. F. Vivanco, "Data classification with binary response through the Boosting algorithm and logistic regression," *Expert Systems with Applications*, vol. 69, pp. 62–73, Mar. 2017, doi: 10.1016/j.eswa.2016.08.014.
- [18] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)," *Informatics in Medicine Unlocked*, vol. 16, p. 100200, 2019, doi: 10.1016/j.imu.2019.100200.
- [19] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model," *Big Data and Cognitive Computing*, vol. 7, no. 3, p. 144, Aug. 2023, doi: 10.3390/bdcc7030144.
- [20] 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, doi: 10.1109/ACCESS.2023.3304242.
- [21] D. Tsetso *et al.*, "Multi-Input Deep Learning Approach for Breast Cancer Screening Using Thermal Infrared Imaging and Clinical Data," *IEEE Access*, vol. 11, pp. 52101–52116, 2023, doi: 10.1109/ACCESS.2023.3280422.
- [22] L. Breiman, "Random Forests," *Machine learning*, vol. 45, pp. 5–32, 2001.
- [23] G. Biau and E. Scomet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
- [24] N. S. Alfaiz and S. M. Fati, "Enhanced Credit Card Fraud Detection Model Using Machine Learning," *Electronics (Switzerland)*, vol. 11, no. 4, p. 662, Feb. 2022, doi: 10.3390/electronics11040662.
- [25] Y. Li, G. Yuan, and Z. Wei, "A limited-memory BFGS algorithm based on a trust-region quadratic model for large-scale nonlinear equations," *PLoS ONE*, vol. 10, no. 5, p. e0120993, May 2015, doi: 10.1371/journal.pone.0120993.
- [26] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization," *Algorithms*, vol. 16, no. 6, p. 308, Jun. 2023, doi: 10.3390/a16060308.

BIOGRAPHIES OF AUTHORS







Muhammad Zeerak Awan     has completed a Master of Science in Mathematics from COMSATS University Islamabad. His research interests include healthcare and analytics. He is a resaeacher in the department of AI and Bigdata Center at Namal University Mianwali Pakistan. He can be contacted at email: zeerak7151@gmail.com.







Muhammad Shoaib Arif     received his Ph.D. in applied mathematics from Beijing Institute of Technology, Beijing, China, in 2015. He is an associate professor at Air University Islamabad, Pakistan. He is currently a Senior Researcher with the TAS Laboratory at Prince Sultan University, Riyadh, Saudi Arabia. His research interests include numerical analysis, mathematical biology, optimization, and machine learning techniques. He can be contacted at email: marif@psu.edu.sa.



Mirza Zain Ul Abideen     is a M.Phil. Biotechnology student at Quaid-i-Azam University, Islamabad. His research area is Genetic Diseases and Oncology. He can be contacted at email: fedralarea@gmail.com.



Kamaleldin Abodayeh     received the M.Sc. degree in functional analysis from University College Dublin, the Ph.D. degree from University College Cork, Ireland, in 1997, and the Ph.D. degree from the Department of Process Engineering, University College Cork. Since 2001, he has been with Prince Sultan University, Saudi Arabia. He has published more than 60 articles in various areas of pure and applied mathematics. His research interests include functional analysis, theoretical physics, discrete potential theory, fixed point theory, and quality monitoring and statistical hypothesis testing. He can be contacted at email: kamal@psu.edu.sa.