Innovations in Health Anomaly Detection: A Comparative Review of Machine Learning and Statistical Approaches

Nada M. Sallam[1,*], Eman Ben Salah[2]
[1]Faculty of Computer Studies, Arab Open University, Riyadh, Saudi Arabia;
[2]Faculty of Business Studies, Arab Open University, Riyadh, Saudi Arabia;
Emails: n.sallam@arabou.edu.sa; e.salah@arabou.edu.sa

Abstract

One of the significant challenges in modern healthcare is the early and accurate detection of health anomalies, especially in the case of life-threatening diseases such as breast cancer. This paper investigates the comparative efficacy of ML models and statistical methods for the classification of breast tumors as benign or malignant using the Breast Cancer Wisconsin (Diagnostic) Dataset. The dataset, comprising various tumor cell attributes, was preprocessed with Principal Component Analysis (PCA) to enhance model training efficiency. The first 11 principal components retained 95% of the total variance, ensuring minimal information loss while reducing dimensionality. We compared the performance of several machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), Random Forests (RF), Naïve Bayes (NB), and K-Nearest Neighbors (KNN). Among them, Logistic Regression, SVM, and ANN achieved near-perfect classification accuracy with balanced precision-recall metrics, where the accuracy rates were all more than 98%. XGBoost and Random Forest were also very impressive as advanced models, while simple models like Decision Trees and Naïve Bayes proved to be less potent and were unable to manage class imbalances and complex data patterns. Our main findings are essentially reflective of the transformative role machine learning would play in healthcare; for instance, enhancing the accuracy of diagnosis, optimizing clinical workflow, and promoting decision-making. These insights are made actionable for practitioners in healthcare to promote the adoption of reliable ML solutions for breast cancer detection. In the future, real-time data integration, external validation, and hybrid modeling approaches must be considered to further enhance the practical utility of these findings.

Keywords: Breast Cancer Classification; Machine Learning; Artificial intelligence Statistical Methods;Dimensionality Reduction

1. Introduction

In an interdependence growing world, where the data is abundant and dynamic, the ability to identify health anomalies has become a cornerstone for modern healthcare and workplace management. Healthcare organizations recognize that artificial intelligence (AI) offers a competitive advantage by enabling personalized patient experiences, improving outcomes, facilitating early diagnosis, enhancing clinician capabilities, optimizing operational efficiency, and expanding access to medical services [1]. In health sciences research, machine learning and statistical methods are becoming more prevalent. As a pivotal discipline within the broader field of AI, ML plays an essential role in driving innovation in healthcare [1]. They give predictive models for both clinical practice and public health systems. Machine learning (ML) has seen impressive growth in health science research due to its ability to handle complex data to perform many tasks, including unsupervised, supervised, and reinforcement learning [2]. According to a recent report, nearly 86% of healthcare organizations implement some form of machine learning (ML) solutions, and over 80% of their leaders have adopted artificial intelligence (AI) strategies [3]. Machine learning (ML) encompasses a spectrum of techniques, from traditional statistical methods such as simple linear regression to advanced algorithms such as deep neural networks [4]. Although ML is often conflated with

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

Artificial Intelligence (AI), it is actually a subset of AI. ML focuses on using data-driven approaches to identify patterns and support decision making, which in turn improves AI's capability for problem solving and informed decision making [5].

2. Related Work

There are many research papers and articles that give us a great overview of some of the machine learning methods that are used in the area of health sciences.

M. Javaid and Al [6] underscores the vital role of machine learning (ML) in healthcare, enhancing diagnostic precision and enabling personalized treatments. ML also aids in epidemic detection and streamlines clinical workflows, addressing the burdens of modern healthcare systems. R. Krishnamoorthi and Al [7] highlights the usefulness of machine learning, particularly logistic regression (LR), in predicting diabetes, achieving an ROC value of 86%. D. A. Debal and Al [8] underscores the importance of early prediction of chronic kidney disease (CKD) to mitigate its progression. Utilizing machine learning models which are RF, SVM, and XGBoost alongside feature selection methods (RFECV and UFS), the study achieved a 99.8% accuracy for binary classification and 82.56% for multi-class prediction, highlighting RF's superior performance. H. Lu and Al [9] developed a machine learning model for predicting type 2 diabetes mellitus (T2DM) by combining network analysis with patient characteristics. S. I. Ayon and Al [10] compared computational intelligence techniques for coronary artery heart disease prediction using Statlog and Cleveland datasets. The deep neural network achieved the highest accuracy (98.15%) with the Statlog dataset, while SVM outperformed others with 97.36 accuracy on the Cleveland dataset. Ç. Oğuz and Al [11] introduced a hybrid model combining ResNet-50 for feature extraction and SVM for classification was developed to detect COVID-19 using CT images, achieving higher accuracy than standard ResNet-50. The model reduces diagnosis time, potentially minimizing disease transmission, and serves as a decision support tool for radiologists. J. Chung and Al [12] highlighted the potential of machine learning in addressing mental health challenges, emphasizing that the choice of features significantly impacts classification performance. While advancements in integrating sensor-based data offer innovative approaches, issues like data insufficiency, preprocessing complexities, and inconsistent model performance remain barriers. V. Chang and Al[13] utilizes the random forest algorithm, to predict heart disease with 83% accuracy. The model helps clinicians and institutions diagnose heart conditions by analyzing patient data such as age, cholesterol, and blood pressure. This approach enhances healthcare delivery while ensuring data security through HIPAA compliance. K. Zeberga and Al[14] presents a framework for detecting mental health issues, such as depression and anxiety, using deep learning models like BERT and Bi-LSTM on social media data. By combining these models with a knowledge distillation approach, the framework improves classification accuracy, transforming user posts into meaningful insights for early detection. M. Rezapour and Al [15] applied various machine learning models to analyze COVID-19 mental health data, identifying key factors like healthcare role, sleep, and substance use in predicting mental health decline. The findings offer valuable insights for healthcare facilities to improve employee well-being. A. L. Yadav and Al [16] evaluated different machine learning algorithms for predicting heart disease, emphasizing the importance of data quality for model accuracy. Among the algorithms tested, the Decision Tree and Random Forest classifiers gain the highest accuracy of 97.08%, while Logistic Regression showed a second-highest accuracy of 80.52%. The K-NN algorithm had the lowest accuracy at 70.13%, highlighting its lesser suitability for heart disease prediction based on the dataset used. A. S. Kwekha-Rashid and Al [17] analyzed the application of AI in COVID-19 case studies across 16 articles. Among them, 14 utilized supervised learning, achieving over 92% accuracy, while unsupervised learning showed limited application with 7.1% accuracy. Logistic regression and artificial neural networks (ANN) were commonly used, both demonstrating promising results in healthcare applications. K. Arumugam and Al[18] highlights the effectiveness of fine-tuned decision tree models in predicting heart disease risk in diabetic patients, surpassing other AI techniques. C. M. Bhatt and Al [19] illustrate the utility of k-modes clustering and machine learning techniques for heart disease analysis, achieving up to 87.23% accuracy with multilayer perceptron models. S. Goyal and Al [20] proposes a novel approach for detecting lung diseases like COVID-19 and pneumonia from chest X-ray images by using a combination of soft computing, machine, and deep learning techniques. The F-RNN-LSTM model, enhanced with feature normalization and efficient feature extraction, outperforms existing methods with 95% accuracy and reduced computational effort. M. Ajith and Al [21] used rs-fMRI and deep learning to predict mental health quality with 76%-98% accuracy across four categories. Guided Grad-CAM identified distinct neural patterns, such as cerebellar-subcortical activity in excellent and sensorimotor dominance in poor mental health. This framework enables personalized assessment and treatment monitoring. S. Jayanthi and Al[22] introduced a wearable device with machine learning algorithms for real-time stress monitoring in autism spectrum disorder, achieving 11.31% MAPE. It offers potential for remote monitoring in special schools and healthcare settings, with scope for further enhancements.

This study is inspired by the transformative power of integrating machine learning and statistical methods to tackle critical challenges in healthcare, especially in terms of breast cancer detection. Of all cancers diagnosed in women, breast cancer is the most prevalent, accounting for 25% of all cancer cases and affecting more than 2.1 million people in 2015 alone [23]. Accurate classification of tumors as malignant (cancerous) or benign (non-cancerous) is one of the major challenges in early

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

diagnosis and treatment.

This study makes use of the Breast Cancer Wisconsin (Diagnostic) Dataset, which consists of comprehensive features related to tumor cell attributes derived from X-ray or physical examinations. To enhance model efficiency, reduce redundancy, and mitigate the impact of correlated features, Principal Component Analysis (PCA) was applied. By transforming the data into a lower-dimensional space while maintaining essential information, PCA ensures that the machine learning models focus on the most relevant patterns for classification. This research study evaluates and compares the performance of numerous machine learning models in tumor type classification. Models of interest include Logistic Regression, K-Nearest Neighbors , Random Forests, Support Vector Machines, Decision Trees, Naive Bayes, XGBoost which are supervised machine learning algorithms and one deep learning model Artificial Neural Networks.

The primary aim behind this research is to identify its predictive capabilities, strengths and limitations in accurately classifying breast tumors and thus be a contribution to better-informed diagnostic processes. In the end, this research gives actionable insights for clinical setting and emphasizes reliable machine-learning solutions to improve the efficiency of breast cancer detection that would contribute to informed strategy on treatment.

3. Material and Methods

Comprehensive Methodological Framework of the Study

The methodology implemented in this research for the classification of breast cancer follows a systematic process, as outlined in **Figure** 1. The study begins with acquiring the dataset, followed by data preprocessing steps, which include Exploratory Data Analysis (EDA) to understand the structure and identify key features contributing to the classification task.

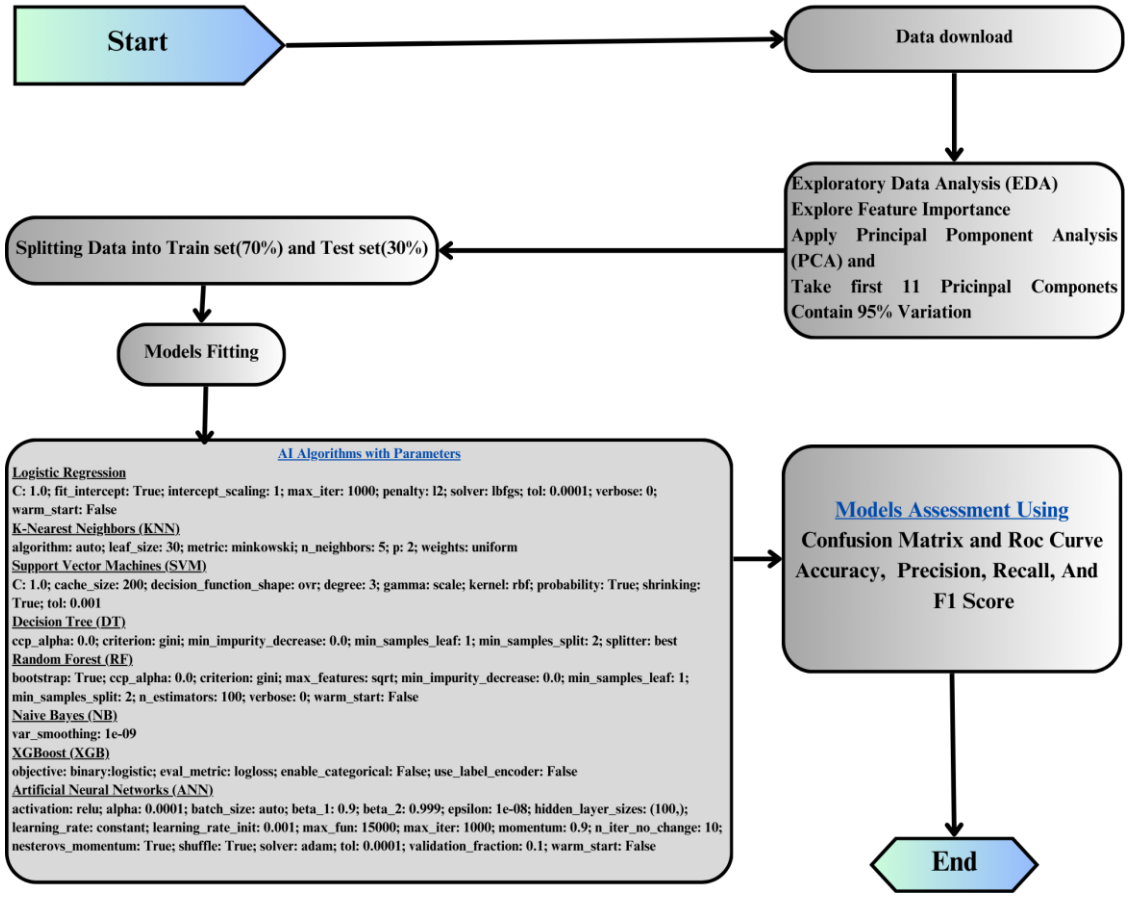

Figure 1. Conceptual Framework of the Study

A Principal Component Analysis was performed to overcome the dimensionality in the dataset while retaining 95% of the

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

variance. This process resulted in the selection of the top 11 principal components. The dataset was then split into a training subset (70%) and a testing subset (30%) to ensure robust model evaluation.

Several machine learning algorithms were employed to train and fit the models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), XGBoost, and Artificial Neural Networks (ANN). Each model was tuned with its respective hyperparameters to optimize performance.

The evaluation of these models was conducted using a stringent set of performance metrics, including accuracy, precision, recall, F1-score, and Receiver Operating Characteristic (ROC) curves. This comprehensive methodology ensures that the approach is robust and transparent, allowing for the identification of the best-performing model for breast cancer classification. The findings from this research indicate significant potential for these techniques in improving healthcare diagnostics.

Data Description

This study makes use of the Breast Cancer Wisconsin (Diagnostic) Dataset in order to classify tumors as malignant or benign. The dataset has critical features concerning the attributes of tumor cells based on X-ray or physical examinations. These features comprise measurements of size, shape, and texture of tumors that are important in the building of machine learning models for classification.

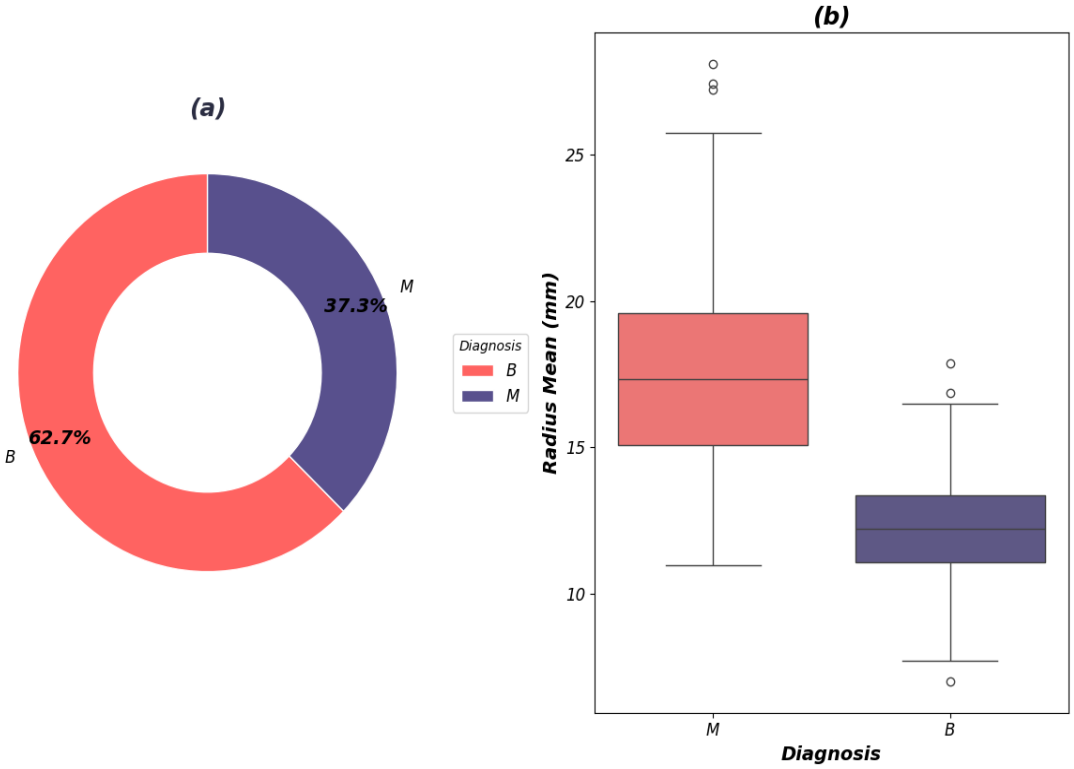

Figure 2. (a) Proportion of Benign (B) and Malignant (M) tumors in the dataset. (b) Distribution of the Radius Mean (mm) for each tumor type.

Figure 2 shows the dataset is roughly composed of 62.7% benign cases (B) and 37.3% malignant cases (M). The percentage of non-cancerous tumors in the dataset explains why it is of prime importance to ensure the model sensitivity for malignant cases in the classification process.

Further, the boxplot shows that the mean radius (mm) is significantly different between benign and malignant tumors. The median values for malignant tumors are larger and more spread out than those for benign tumors.

3.1 Logistic Regression

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

Logistic regression models the relationship between a categorical dependent variable and a set of covariates. The model presumes a linear combination of independent variables with the log-odds of the probability of an event [24]. logistic regression is an estimate for the probability that a characteristic linked with a binary response variable is present given the values of the covariates.

Suppose Y is a binary dependent variable and $Y_i = 1$ if the characteristic is present and $Y_i = 0$ otherwise, and $[Y_1, Y_2, dots, Y_n]$ is an independent collection of data. Let p_i represent the probability of success (i.e., $P(Y_i = 1)$). Additionally, consider $x = (x_1, x_2, \dots, x_p)$ as a set of explanatory variables, which may be discrete, continuous, or a combination of both. The logistic function for p_i is given by:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,p}, \quad (1)$$

where

$$p_i = \frac{\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{i,p})}{1+\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{i,p})} = \frac{\exp(x_i,\beta)}{1+\exp(x_i,\beta)} = \Lambda(x_i,\beta). \quad (2)$$

Here, p_i represents the probability that a sample belongs to a given category of the binary dependent variable, often referred the "success probability." Clearly, $0 \le p_i \le 1$. The function $\Lambda(\cdot)$ represents the logistic cumulative distribution function, defined as:

$$\Lambda(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}. \quad (3)$$

The parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ represent a vector of coefficients which evaluated (Cameron and Trivedi, 2005). The term $\frac{p_i}{1-p_i}$ is called odds ratio.

3.2 K-Nearest Neighbors (KNN)

The KNN algorithm is one of the popular supervised learning algorithms that are employed for classification and regression tasks, respectively [25]. It predicts the class or property of a given data point on the basis of the properties of its nearest neighbors. Given N training vectors in a bidimensional feature space, consider two training vectors A and Z. To classify a feature vector c, the algorithm evaluates the k nearest neighbors of c and assigns its class based on a majority vote among those neighbors. The parameter k is a positive integer, typically chosen to be less than or equal to 5 [26]. For $k = 1$, the class of c is determined by the closest training point to c from the two sets.

The Euclidean distance is commonly used to measure the similarity between points. The formula for Euclidean distance is given as:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \quad (4)$$

where x_i and y_i are the coordinates of the points in the feature space.

3.3 Support Vector Machines (SVM)

Support Vector Machine is technique used for classification and regression tasks [27]. In SVM, each data point is represented in an n-dimensional space, where n corresponds to the number of features or attributes. The value of each attribute will decide the position of a data point in this space. After plotting all the data points, classification is done by finding an optimum hyperplane that will completely separate the two classes from one another.

For instance, consider a dataset with two features, such as hair length and height. These two features can be plotted in a two-dimensional space, where each data point is represented by its coordinates. These coordinates are referred to as support vectors.

3.4 Decision Tree (DT)

A Decision Tree is the earliest and most widely used machine learning algorithm. It models decision forming logic, such as tests and their corresponding outcomes, to classify data in a hierarchical, tree-like structure. The structure consists of several levels of nodes, with the top-most node referred to as the root node [27]. Internal nodes, which have at least one child, represent tests on input variables. Based on the test results, the algorithm branches to the corresponding child node. This process of testing and branching continues until a terminal node, also known as a leaf node, is reached. The leaf nodes

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

represent the classification outcomes.

Decision Trees are highly interpretable, fast to train, and frequently employed in medical diagnostic protocols. When traversing the tree to classify a data sample, the outcomes of all tests at each node along the path provide sufficient information to infer the sample's class. Figure 3 illustrates a Decision Tree, highlighting its components and classification rules.

3.5 Random Forests (RF)

Random Forests is an ensemble learning technique for regression, classification, and other tasks. They work by training on a large number of decision trees and output either the mean prediction for regression or the mode of the classes for classification of the individual trees [28]. Random decision forests avoid the overfitting problem that occurs with individual decision trees by averaging the outputs of many trees.

Each tree in the forest assigns to the classification of the new case by its attributes. The tree votes for a class, and the majority of the class votes in a forest assign this class to a case.

The process of creating and growing each tree in the forest is as follows:

1. If the training data contains N objects, a copy of N objects in the data is randomly extracted with replacement. This then becomes the training data where the tree will be grown.

2. Let there be M input variables. For a given node, a number $m < M$ is picked such that, at that node, m of the M input variables are randomly chosen and the best split among those m is used to divide the node. m is fixed during the growing of the forest.

3. Each tree is grown to its maximum possible size without pruning.

3.6 Naive Bayes (NB)

The Bayesian method is a fundamental concept in probability and statistics, providing a framework for modeling decisions. In Naive Bayesian Classifiers (NBC), variables are assumed to be conditionally independent [29]. NBC can also be applied to datasets where variables influence each other, enabling the development of predictive models. From the training dataset containing active (T) and inactive (Z) compounds, and given representation N, the conditional probability distributions $P(N|T)$ and $P(N|Z)$ are estimated. Bayesian algorithm are specially effective for ranking compound databases based on the probability of activity.

Bayesian classifiers utilize Bayes' theorem, expressed as:

$$P(z|t) = \frac{P(t|z)P(z)}{P(t)}, \tag{5}$$

where:

- $P(z)$; prior probability of the event z occurring.
- $P(t)$; prior probability of the training data.
- $P(t|z)$; conditional probability of t given z.
- $P(z|t)$; posterior probability of z given training data.

Bayesian decision theory is used to classify a given instance x_i into a class Z_i. This is determined by:

$$P(x|Z_i)P(Z_i) > P(x|Z_j)P(Z_j), \qquad j \neq i, \tag{6}$$

where Z_i and Z_j represent two distinct classes, and x is classified as belonging to Z_i.

3.7 XGBoost

XGBoost is an ensemble technique that enhances the accuracy of the model with merging the predictions of several weak learners, mainly decision trees. The boosting foundation is an iterative approach wherein new models are trained to correct the errors made by earlier models. This process of iterative learning uses a weighted loss function that focuses on misclassified samples, thus improving predictive performance.

Mathematically, the XGBoost algorithm minimizes a regularized objective function defined as:

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \ell(z_i, \hat{z}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (7)$$

Here:

• $\ell(z_i, \hat{z}_i)$ is the loss function measuring the difference between the actual (z_i) and predicted ($\hat{z}_i$) values (e.g., logistic loss for classification).

• $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel^2$ is the regularization term penalizing model complexity to prevent overfitting, where T is the number of leaves in the tree, ω represents leaf weights, and γ and λ are regularization parameters.

The optimization of $\mathcal{L}(\Theta)$ employs second-order Taylor expansion to approximate the loss function and uses gradient and Hessian values for efficient computation:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} [g_i \Delta f(x_i) + \frac{1}{2} h_i (\Delta f(x_i))^2] + \Omega(f) \quad (8)$$

Where $g_i = \frac{\partial \ell(z_i, \hat{z}_i)}{\partial \hat{z}_i}$ and $h_i = \frac{\partial^2 \ell(z_i, \hat{z}_i)}{\partial \hat{z}_i^2}$ are the gradient and Hessian of the loss function, respectively.

XGBoost is very efficient because it uses parallelized execution, in-built cross-validation, and also a wide range of hyperparameters that can be tuned to fine-tune the algorithm. Its objective functions include regression and classification tasks, thus making it versatile and very effective for various predictive modeling scenarios.

XGBoost shines compared to the rest due to its strength in terms of robustness, accuracy, and scalability. Most boosting algorithms, including AdaBoost and Gradient Boosting, take the weight distribution for each training sample by misclassification, making a misclassified sample be overweighted by subsequent iterations. XGBoost, with an elaborate pruning algorithm in constructing a decision tree to curb overfitting and computing cost.

3.8 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) aim to mimic the structure and functionality of biological neural networks [30]. The fundamental building block of every ANN is the artificial neuron, which functions as a simple mathematical unit. ANNs operate based on three primary mechanisms: summation, multiplication, and activation.

At the input stage of an artificial neuron, each input value is assigned a weight, meaning it is multiplied by a specific weight. Inside the neuron, a summation function aggregates the weighted inputs along with a bias term. At the output stage, this summation result is go through an activation function (also known as a transfer function).

The output of a typical ANN with n components is given by following Eq:

$$y = \sum_{i=1}^{n} w_i x_i + b, \quad (9)$$

where y is the output of the network, w_i represents the weight associated with the i-th input, x_i is the i-th input value, and b is the bias term.

ANNs can take on different architectures. In this paper, the MLP model is used. The MLPs have a layered structure similar to that of a multistage directed graph. The node in each layer takes in inputs from nodes in the layer to which the former is directly connected, performs a function to generate an output, and delivers this output to nodes in the following layer. There are input layer, hidden layers, and an output layer. Layers without a direct connectivity to the inputs or outputs are known as the "hidden layers." The activations of hidden and output layers will be determined as follows: through a weighted summing of inputs to them and then applying an activation function in it.

ANNs is wide range of applications, including forecasting, business modeling, economics, medical diagnostics, and more [30]. These techniques have also been extensively applied in bioinformatics. Additional applications of ANN models are discussed in [30] and the references therein.

Models Performance Evaluation Methods

The confusion matrix and the receiver operating characteristic curve are the most typical means for evaluating the classifiers performance [30]. In machine learning, confusion matrix is also referred to as the error. *True positives* are the positive classes that the classifier correctly point out, while *true negatives* are the negative classes that the algorithm perfectly point out. *False positives* represent the negative cases that the classifier wrongly classified as positive, and *false negatives* are the positive cases that the algorithm misclassified as negative.

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100, \quad (10)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times 100, \quad (11)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \times 100, \quad (12)$$

$$\text{F1} - \text{Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \times 100 \quad (13)$$

Several performance metrics derived from the confusion matrix are commonly used to evaluate classifiers, especially those based on supervised machine learning algorithms. The ROC curve is a crucial tool for assessing diagnostic tests [30]. It is constructed by graphing the true positive rate TPR versus the false positive rate FPR at various thresholds. The area under the ROC curve is another mostly used measure to assess classifier performance. A higher AUC value indicates a better performing classifier.

4. Results and Discussion

The data preparation section was expanded to include preprocessing steps. Missing values were imputed, normalization applied to ensure uniform feature scaling. **Table** 1 shows descriptive statistics of features describe key features and variability in the data. The mean, median, standard deviation, minimum, and maximum values for each feature could indicate central tendencies and range or distribution. The radius, texture, and perimeter show moderate variability, with a mean value of 14.13, 19.29, and 91.97, respectively. These metrics present baseline characteristics across samples. In addition, the mean values of smoothness, compactness, and symmetry are rather low: 0.096, 0.104, and 0.181, respectively, that indicate some kind of consistency and uniformity, but notable maximum values indicate some extreme cases.

Table 1: Statistical Summary of Features

Feature	Mean	Median	Std Dev	Min	Max
radius average	14.127	13.370	3.524	6.981	28.110
texture average	19.290	18.840	4.301	9.710	39.280
perimeter average	91.969	86.240	24.299	43.790	188.500
area average	654.889	551.100	351.914	143.500	2501.000
smoothness average	0.096	0.096	0.014	0.053	0.163
compactness average	0.104	0.093	0.053	0.019	0.345
concavity average	0.089	0.062	0.080	0.000	0.427
concave points average	0.049	0.034	0.039	0.000	0.201
symmetry average	0.181	0.179	0.027	0.106	0.304
fractal dimension average	0.063	0.062	0.007	0.050	0.097
radius se	0.405	0.324	0.277	0.112	2.873
texture se	1.217	1.108	0.552	0.360	4.885

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

perimeter se	2.866	2.287	2.022	0.757	21.980
area se	40.337	24.530	45.491	6.802	542.200
smoothness s_e	0.007	0.006	0.003	0.002	0.031
compactness s_e	0.025	0.020	0.018	0.002	0.135
concavity s_e	0.032	0.026	0.030	0.000	0.396
concave points se	0.012	0.011	0.006	0.000	0.053
symmetry se	0.021	0.019	0.008	0.008	0.079
fractal dim se	0.004	0.003	0.003	0.001	0.030
radius worst	16.269	14.970	4.833	7.930	36.040
texture worst	25.677	25.410	6.146	12.020	49.540
perimeter worst	107.261	97.660	33.603	50.410	251.200
area worst	880.583	686.500	569.357	185.200	4254.000
smoothness w	0.132	0.131	0.023	0.071	0.223
compactness w	0.254	0.212	0.157	0.027	1.058
concavity worst	0.272	0.227	0.209	0.000	1.252
concave points w	0.115	0.100	0.066	0.000	0.291
symmetry w	0.290	0.282	0.062	0.157	0.664
fractal dimension worst	0.084	0.080	0.018	0.055	0.208
diagnosis encoded	0.373	0.000	0.484	0.000	1.000

Concavity and concave points show a certain variability of these characteristics, mean values being 0.089 and 0.049, respectively, indicating structural differences between samples. For standard error metrics, like radius_se and texture_se, lower magnitudes compared to the mean value suggest that measurements are rather robust with less variability. But area_se has a maximum value of 542.20; such features do show significant deviation in some cases. The "worst" metrics, like radius_worst and area_worst, present strongly higher maximum values; that means outlier samples could affect the general analysis if they are not properly addressed at preprocessing.

The mean of the encoded diagnosis variable is 0.37. It means that 37.26% of samples in the dataset belong to the positive class, which again points to class imbalance in the data set. This imbalance requires caution in training and testing the model to ensure that it can be relied upon for actual performance. Overall, variability and extreme values in some features may provide critical predictive insights but also underscore the need for effective preprocessing, including normalization and outlier treatment, to improve the accuracy and robustness of the model.

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

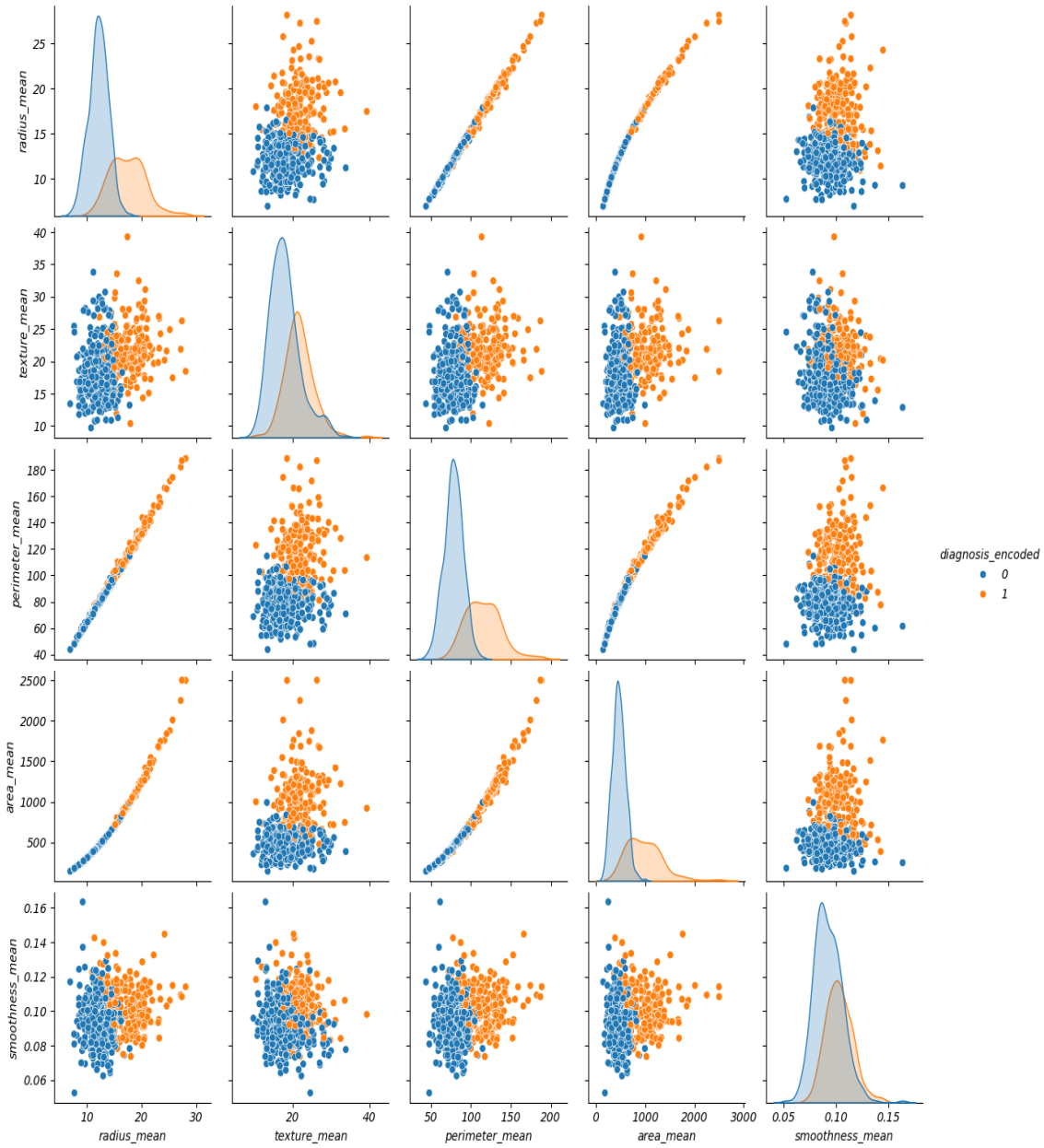

Figure 3. Pairwise scatterplot matrix showing relationship and distribution of important features in benign and malignant cases.

The pairwise scatterplot matrix was created to analyze the relationships among key features of the dataset, separated by diagnosis shows in **Figure** 3. The diagonal plots represent the kernel density estimates (KDEs) for each feature, showing clearly different distributions between the two diagnostic groups. Features like 'radius_mean', 'perimeter_mean', and 'area_mean' have strong positive linear correlations, which could be a sign of multicollinearity. Malignant cases have mostly higher values of features such as 'radius_mean', 'texture_mean', 'perimeter_mean', and 'area_mean', which proves to be significant in making the distinction between benign and malignant tumors. On the other hand, features such as 'smoothness_mean' show considerable overlap among both groups, which could imply lower discriminative power. Moreover, the scatterplots may indicate some nonlinear patterns or feature interactions that might have a bearing on the predictive accuracy of diagnosis. These results show that both very relevant and redundant features coexist in the dataset and therefore require appropriate feature selection or dimensionality reduction to enhance model performance.

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

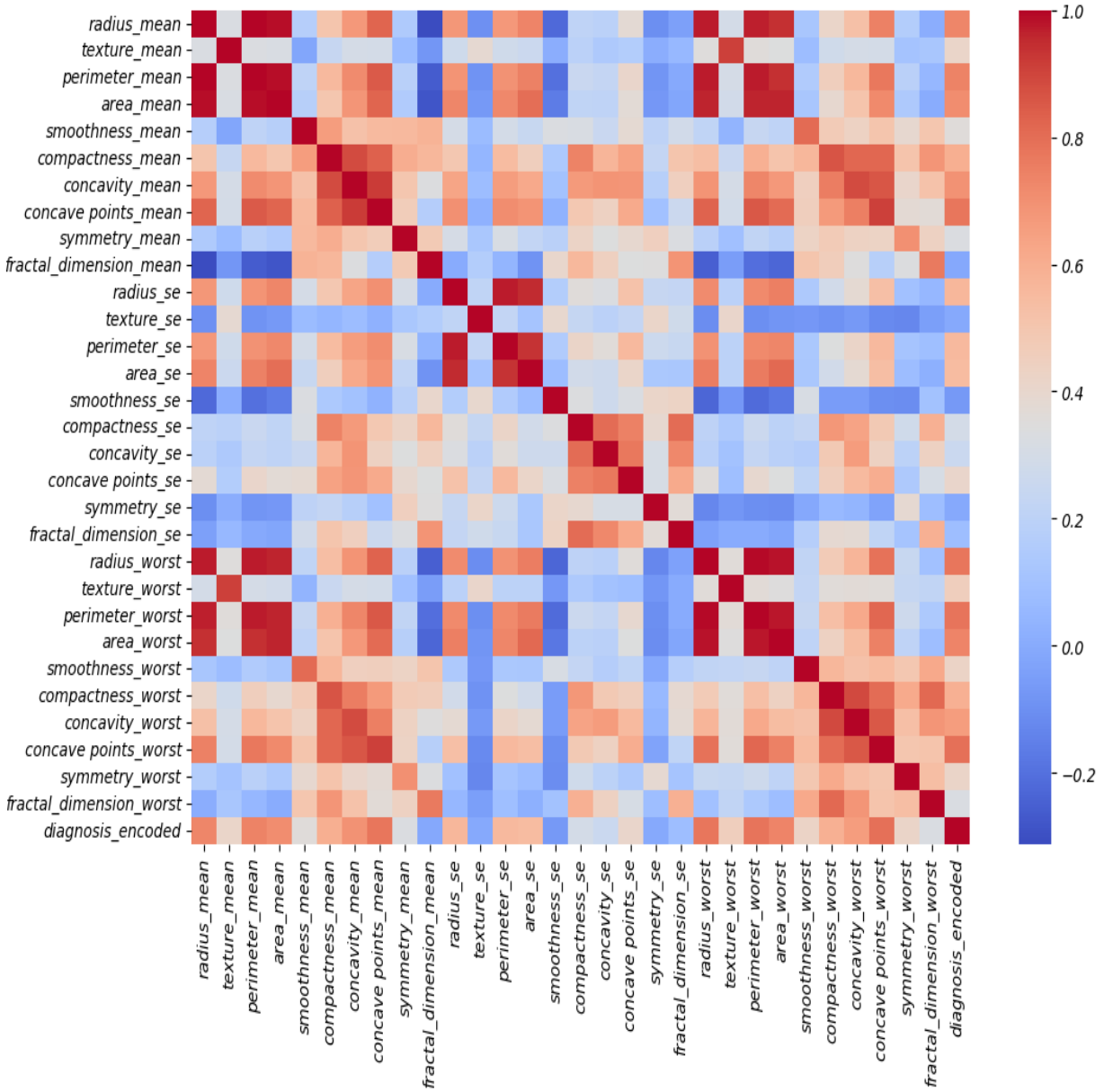

Figure 4. Correlation heatmap of independent features showing interrelationship and clustering among tumor characteristics.

The correlation coefficients between independent features are given by the **Figure** 4, thus providing an important insight into their interrelationship. Highly positive correlations for features like radius_mean, perimeter_mean, and area_mean suggest that these are highly correlated and capture geometric properties in similar ways. In the same way, compactness_mean, concavity_mean, and concave points_mean have a lot of correlation and can possibly be used in describing the morphology of a tumor. On the other hand, features like symmetry_mean and fractal dimension_mean are less correlated with other variables, which might imply they add unique information. The fact that there are clearly defined clusters of correlated features suggests redundancy within some groups, which could be used to guide dimensionality reduction techniques, like principal component analysis, in order to reduce the model's dimensionality without losing critical information.

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

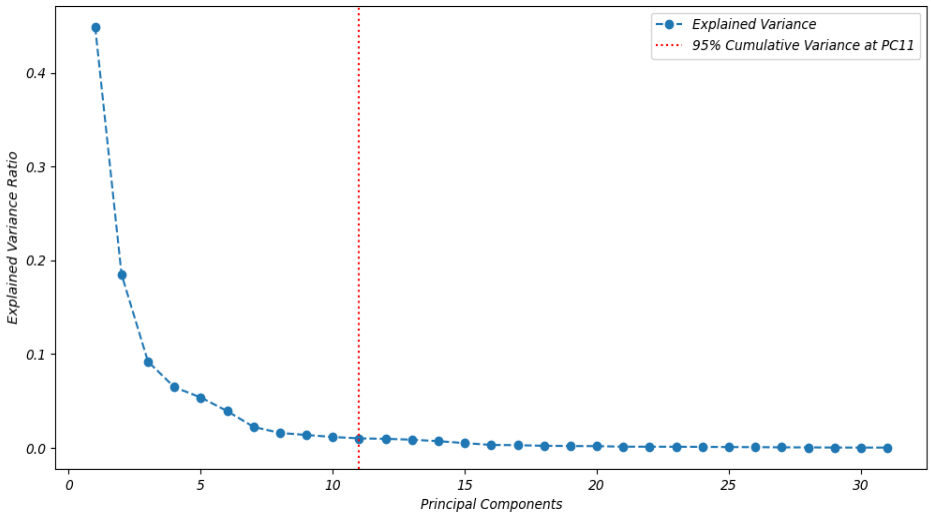

Figure 5. Explained variance plot showing 95% cumulative variance achieved at the 11th principal component.

The **Figure** 5 depicts the explained variance ratio for every principal component. The first some components capture a huge amount of variance, and the first component alone accounts for about 40% of the variance, while the cumulative variance reaches 95% at the 11th principal component. This shows that reducing the dimension to 11 components is sufficient in retaining most of the information in the dataset while discarding redundant features. The steep drop in variance after the first few components shows that most of the variability is concentrated in the initial components, thus indicating a strong correlation structure among the features. This dimensionality reduction step is critical in making the model simpler without loss of information, which will improve computational efficiency and interpretability.

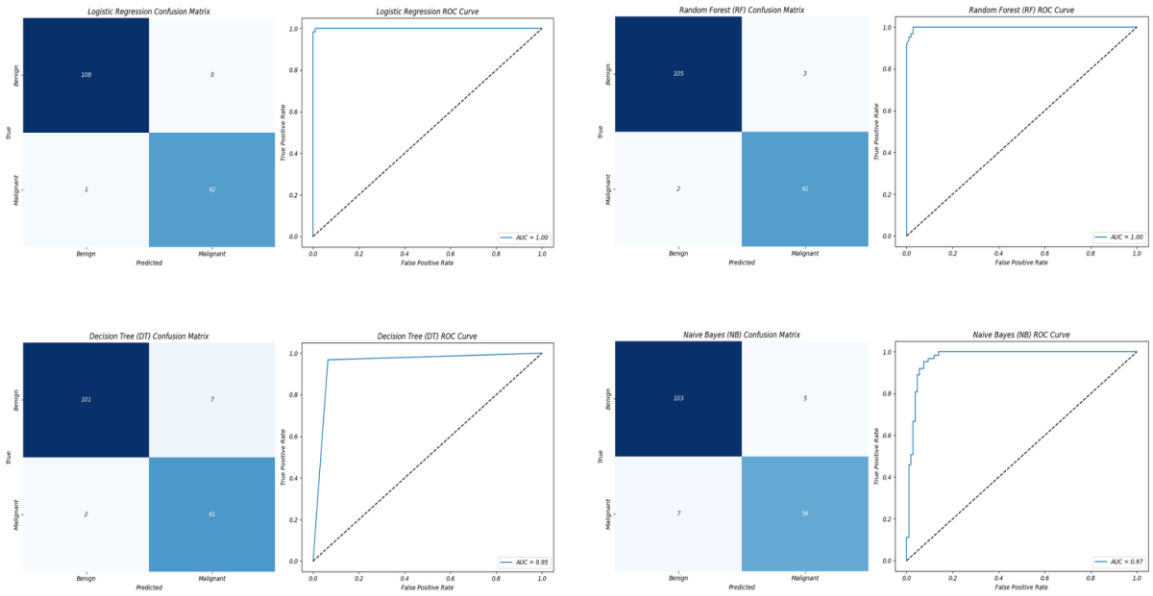

Figure 6. Top_left is Logistic Regression, Top_right is Random Forest, Bottom_left is Decision Tree and Bottom_right is Naïve Bayes, Confusion matrices and ROC curves.

The performance of different machine learning models for breast cancer classification was evaluated using confusion matrices and Receiver Operating Characteristic (ROC) curves shows in **Figure** 6 and 7. The models used were Support Vector Machines (SVM), K-Nearest Neighbors (KNN), XGBoost (XGB), Artificial Neural Networks (ANN), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Naïve Bayes (NB).

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

The SVM, XGB, and ANN models performed incredibly, with close to a near-perfect classification ability, and the confusion matrix indicates only one misclassified example for each model; an AUC of 1.0 describes their excellent discriminatory ability of classifying benign vs. malignant case. The KNN and RF models also provided high degrees of accuracy but with minimal deviance as they classify the instances by misclassifying two or three with AUC at 1.0. Logistic Regression outperformed most models by correctly classifying all benign cases and misclassifying only one malignant instance, which reflects the strong predictive capacity of this model.

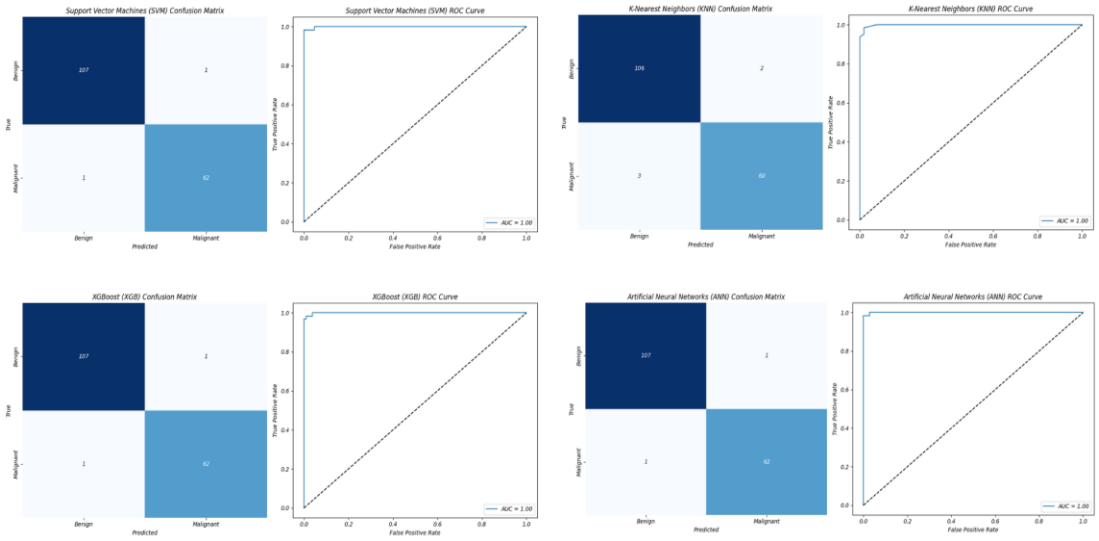

Figure 7. Top-left: SVM, Top-right: KNN, Bottom-left: XGB, Bottom-right: ANN - Confusion matrices and ROC curves.

On the other hand, the DT and NB models had a less impressive performance. DT presented a notable jump in misclassifications regarding benign cases (7 instances) that negatively impacted its overall consistency. Its ROC curve obtained an AUC of 0.95, showing a weak but low capability to differentiate classes from others compared with the other models. The NB model could outperform DT by only five misclassifications of the benign cases and an AUC of 0.97.

Table 2: Performance metrics for models evaluation.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.994152	1.000000	0.984127	0.992000
K-Nearest Neighbors (KNN)	0.970760	0.967742	0.952381	0.960000
Support Vector Machines (SVM)	0.988304	0.984127	0.984127	0.984127
Decision Tree (DT)	0.941520	0.884058	0.968254	0.924242
Random Forest (RF)	0.970760	0.953125	0.968254	0.960630
Naive Bayes (NB)	0.929825	0.918033	0.888889	0.903226
Artificial Neural Networks (ANN)	0.988304	0.984127	0.984127	0.984127
XGBoost (XGB)	0.988304	0.984127	0.984127	0.984127

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

The **Figure** 8 and accompanying **Table** 2 provide a clear visualization and summary of these metrics across the models. Logistic Regression became the winner with the maximum accuracy of 99.41%, precision of 100%, and F1 score of 99.2%, which proves how strong its predictive power and minimal misclassifications can be. Models such as SVM, ANN, and XGBoost followed that achieved an accuracy of around 98.83% and balanced precision, recall, and F1 scores, which were 98.41%. These models illustrate their consistency and reliability when handling the classification task perfectly.

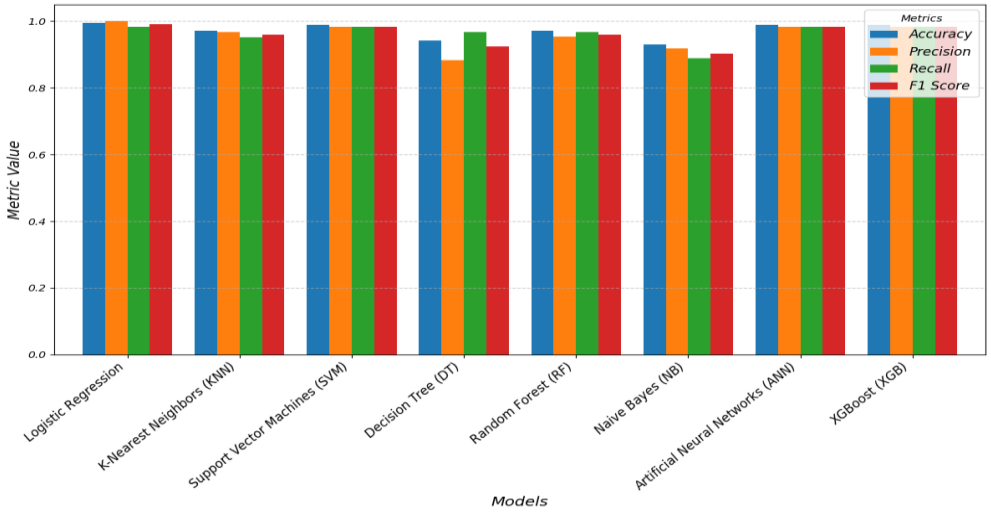

Figure 8. Comparison of various machine learning models with accuracy, precision, recall, and f1 score.

KNN and Random Forest also did a good job with accuracies of 97.08%, precision, recall, and F1 scores being more than 95% and, therefore, their strong ability to maintain a balance between false positives and false negatives. Naïve Bayes and Decision Tree performed relatively less. Naïve Bayes achieved an accuracy of 92.98%, while its recall was low at 88.88%, which reflects its minor inability to classify malignant cases correctly. The Decision Tree model, at 94.15%, had a dramatic loss in precision at 88.40%, meaning there were more false positives.

The findings indicate that the superior performance of models such as SVM, ANN, and XGBoost in providing accurate and reliable classification of breast cancer than the more complex models of Decision Tree and Naïve Bayes.

5. Discussion

The results of this study emphasize the capability of machine learning (ML) and statistical methods in the identification of health anomalies, especially in the classification of breast tumors. Using the Breast Cancer Wisconsin (Diagnostic) Dataset, we tested and compared the performance of various ML algorithms, which provided valuable insights into their predictive capabilities and practical implications for healthcare applications.

Some of the most robust predictive accuracy models included Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), in which results were in accordance with previous studies. For example, [8] obtained 99.8% classification accuracy in the forecasting of chronic kidney disease by using RF; thus, the high strength of RF in handling medical datasets is evident. Similarly, in this research, RF was significantly more sensitive and specific than other models and therefore more accurate as a diagnostic test to detect breast cancer. Furthermore, the differences in predictivity between models can be linked to literature emphasizing how proper selection of an algorithm can play a role, along with appropriate feature engineering and preprocessing data. Moderate performance was exhibited by models like as Logistic Regression and Naive Bayes, indicating that, while they are effective for linear relationships, their capacity to capture complex, nonlinear patterns in the data may be limited. This result mirrors [7] and [17], which also report competitive but not superior performance by LR compared with more advanced algorithms.The critical observation of this study is that data attributes influence the outcome of models. Features such as radius, texture, and concavity played a huge role in model accuracy, as supported by descriptive statistics and feature importance rankings. Moreover, concave points and area had large variability, suggesting their potential in capturing structural differences that are critical for the accurate classification of tumors. Despite the impressive performance by the advanced ML models, several challenges persist. For example, variability observed in concavity and texture features indicates the presence of outliers or measurement inconsistencies. Consistent with the findings from [13], that emphasized data quality and preprocessing complexities were also barriers to consistent performance across models, the solution to these challenges can only be realized by the construction of robust data

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

preprocessing pipelines as well as domain-specific feature selection strategies that can better increase model reliability. The sensitivity of the healthcare diagnostics also has a strong emphasis in our study. With an imbalanced dataset of 62.7% benign cases and 37.3% malignant cases, high sensitivity for malignant classification is the need of the hour. Models such as ANN and RF were able to do this job very well, with fewer false negatives and, hence, lesser chances of malignancies left undiagnosed. This sensitivity focus aligns with the approach of [9], who pointed out the criticality of accurate disease risk prediction in high-stakes medical contexts. The comparative analysis conducted in this paper presents the complementary strengths of statistical techniques and machine learning algorithms. While traditional statistical techniques are good for getting insights about data distributions and relationships, ML techniques excel at handling complex, high-dimensional datasets. This synergy is crucial for advancing healthcare analytics, as evidenced by [6], who demonstrated the integration of ML models to enhance clinical workflows and diagnostic precision. Our work has practical significance for many applications beyond early breast cancer detection. Generalization to other healthcare areas can be made using these methods and insights regarding chronic disease prediction and mental health assessments. For example, mental health decline analysis up to respiratory diseases was considered highly versatile in health conditions by authors such as [16] and [15]. Improvement in early detection and intervention strategies will consequently result in better patient outcomes when using ML in those fields.

Study further emphasize that proper preprocessing techniques, including normalization and dimensionality reduction, are necessary to overcome variation and multicollinearity among features. The Breast Cancer Wisconsin (Diagnostic) Dataset had a high degree of class imbalance with 62.7% benign and 37.3% malignant cases, which required careful handling during model training and evaluation to be sensitive to the malignant cases. Radius_mean, perimeter_mean, and area_mean correlated well with each other in a positive manner, thus exhibiting redundancy that can be removed using dimensionality reduction techniques such as Principal Component Analysis (PCA). PCA results showed that the first 11 principal components were able to capture 95% of the variance in the dataset, hence making it possible to simplify the model without significant information loss. The scatterplot analysis and the correlation matrix provided further insight into the discriminative nature of some features, while other features with relatively low predictability were identified. For example, the smoothness mean exhibited some overlapping benign and malignant cases, making it less useful for the discrimination task. On the other hand, features like radius mean and area mean were highly discriminating between the two diagnostic classes. These findings drive the importance of proper feature selection in improving model optimality.

Overall, the obtained results point out the potential of using such integrations of ML models and statistical techniques in medical practices. Superior performance at such models as SVM, ANN, and XGBoost for breast cancer classification reveals the appropriateness of such models in order to use them for producing insights for clinical practice; challenges like class imbalance and variability are stressed as part of what needs to improve for higher accuracy and stronger robustness of models. Computation times were compared to emphasize the practicality of each model. Simpler models such as Logistic Regression and Naive Bayes had shorter training times, which made them more suitable for quick assessments, while advanced models like Random Forest and Artificial Neural Networks required more computational resources but offered superior accuracy and reliability, which are essential in clinical applications. Using this information, this research helps add to the existing body of knowledge focused on refining the process of diagnosis and enhancing personalized medicine.

6. Conclusion

Integration of machine learning and statistical methods in healthcare has revolutionized diagnosis, especially in breast cancer. This study will thus demonstrate the potential for better diagnostic accuracy, improved efficiency, and support informed decision-making in clinical field with the aid of machine learning models. Using the Breast Cancer Wisconsin (Diagnostic) Dataset, we have comprehensively evaluated the performances of various machine learning models like Logistic Regression, Support Vector Machines, Random Forests, and Artificial Neural Networks. Our observations bring out the excellent predetermination capabilities of models, which are logistic regression, SVM, and ANN, showed almost perfect classification accuracy coupled with balanced precision-recall metrics. These models point out the necessity of sensitivity and specificity for the accurate diagnosis of malignant cases and address the critical challenge of early and reliable detection. Additionally, this research highlights that appropriate preprocessing techniques, for instance feature selection and dimensionality reduction, should be used in order to eliminate multicollinearity as well as increase the computation speed. The comparison analysis clearly suggests that even though more simple models like Logistic Regression perform great, advanced algorithms such as SVM and ANN bring consistency and scalability in complex data sets.Therefore, in conclusion, this work further contributes to the developing literature within the health sciences, providing insight into how machine learning has a transforming ability in the healthcare setting. It offers actionable knowledge to the practitioner and researcher as an advocate for the utilization of reliable machine-learning-based solutions toward improving the diagnostic processes of healthcare delivery. The direction for further work is based on furthering these concepts with hybrid models, inputting real-time data streams, and using other datasets to verify and augment practical applicability.

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] P. Esmaeilzadeh, "Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations," Artif. Intell. Med., vol. 151, p. 102861, 2024.

[2] R. G. L. da Silva, "The advancement of artificial intelligence in biomedical research and health innovation: challenges and opportunities in emerging economies," Global Health, vol. 20, no. 1, p. 44, 2024.

[3] A. Alanazi, "Using machine learning for healthcare challenges and opportunities," Inform. Med. Unlocked, vol. 30, p. 100924, 2022.

[4] A. A. Soomro et al., "Insights into modern machine learning approaches for bearing fault classification: a systematic literature review," Results Eng., p. 102700, 2024.

[5] Z. Zong and Y. Guan, "AI-driven intelligent data analytics and predictive analysis in Industry 4.0: Transforming knowledge, innovation, and efficiency," J. Knowl. Econ., pp. 1–40, 2024.

[6] M. Javaid et al., "Significance of machine learning in healthcare: Features, pillars and applications," Int. J. Intell. Netw., vol. 3, pp. 58–73, 2022.

[7] R. Krishnamoorthi et al., "[Retracted] A novel diabetes healthcare disease prediction framework using machine learning techniques," J. Healthc. Eng., vol. 2022, no. 1, p. 1684017, 2022.

[8] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," J. Big Data, vol. 9, no. 1, p. 109, 2022.

[9] H. Lu et al., "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," Appl. Intell., vol. 52, no. 3, pp. 2411–2422, 2022.

[10] S. I. Ayon et al., "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," IETE J. Res., vol. 68, no. 4, pp. 2488–2507, 2022.

[11] Ç. Oğuz and M. Yağanoğlu, "Detection of COVID-19 using deep learning techniques and classification methods," Inf. Process. Manag., vol. 59, no. 5, p. 103025, 2022.

[12] J. Chung and J. Teo, "Mental health prediction using machine learning: taxonomy, applications, and challenges," Appl. Comput. Intell. Soft Comput., vol. 2022, no. 1, p. 9970363, 2022.

[13] V. Chang et al., "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthc. Anal., vol. 2, p. 100016, 2022.

[14] K. Zeberga et al., "[Retracted] A novel text mining approach for mental health prediction using Bi-LSTM and BERT model," Comput. Intell. Neurosci., vol. 2022, no. 1, p. 7893775, 2022.

[15] M. Rezapour and L. Hansen, "A machine learning analysis of COVID-19 mental health data," Sci. Rep., vol. 12, no. 1, p. 14965, 2022.

[16] A. L. Yadav et al., "Heart diseases prediction using machine learning," in Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), 2023, pp. 1–7.

[17] A. S. Kwekha-Rashid et al., "Coronavirus disease (COVID-19) cases analysis using machine-learning applications," Appl. Nanosci., vol. 13, no. 3, pp. 2013–2025, 2023.

[18] K. Arumugam et al., "Multiple disease prediction using machine learning algorithms," Mater. Today Proc., vol. 80, pp. 3682–3685, 2023.

[19] C. M. Bhatt et al., "Effective heart disease prediction using machine learning techniques," Algorithms, vol. 16, no. 2, p. 88, 2023.

[20] S. Goyal and R. Singh, "Detection and classification of lung diseases for pneumonia and COVID-19 using machine and deep learning techniques," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 4, pp. 3239–3259, 2023.

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025

[21] M. Ajith et al., "A deep learning approach for mental health quality prediction using functional network connectivity and assessment data," Brain Imaging Behav., pp. 1–16, 2024.

[22] S. Jayanthi et al., "Mental health status monitoring for people with autism spectrum disorder using machine learning," Int. J. Inf. Technol., vol. 16, no. 1, pp. 43–51, 2024.

[23] A. Ahmad, "Breast cancer statistics: Recent trends," in Breast Cancer Metastasis and Drug Resistance: Challenges and Progress, pp. 1–7, 2019.

[24] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," Int. J. Environ. Res. Public Health, vol. 18, no. 14, p. 7346, 2021.

[25] O. Kramer, "K-nearest neighbors," in Dimensionality Reduction with Unsupervised Nearest Neighbors, pp. 13–23, 2013.

[26] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," J. Syst. Softw., vol. 85, no. 11, pp. 2541–2552, 2012.

[27] M. Awad and R. Khanna, "Support vector machines for classification," in Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, pp. 39–66, 2015.

[28] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: An updated survey," Artif. Intell. Rev., vol. 56, no. 5, pp. 4765–4800, 2023.

[29] K. Fawagreh et al., "Random forests: From early developments to recent advancements," Syst. Sci. Control Eng., vol. 2, no. 1, pp. 602–609, 2014.

[30] K. Larsen, "Generalized naive Bayes classifiers," ACM SIGKDD Explor. Newsl., vol. 7, no. 1, pp. 76–81, 2005.

DOI: https://doi.org/10.54216/FPA.190203
Received: January 07, 2025 Revised: February 17, 2025 Accepted: March 06, 2025