THE EFFECT OF FEATURE SELECTION METHODS ON MACHINE LEARNING MODEL PERFORMANCE: A COMPARATIVE STUDY FOR BREAST CANCER PREDICTION

Diman Siddiq Hassan

Computer Science Department, College of Science, University of Zakho, Zakho, Kurdistan Region, Iraq-
Corresponding author email:diman.hassan@uoz.edu.krd

ABSTRACT:

Developing countries often face a high incidence of breast cancer, making early detection vital for effective treatment. The risk of developing breast cancer can be evaluated using machine learning methods and regular diagnostic data. In cancer datasets, there is a wealth of patient information, but not all of it is valuable for predicting cancer. This highlights the significance of feature selection methods in uncovering the relevant data. In this field, many studies have attempted to predict the different types of breast tumours, since it is important to diagnose breast cancer medication accurately. This paper aims to perform a comparison such that to show the effect of different feature selection methods on the accuracy of various existing machine learning algorithms. The study focuses on seven machine learning algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression (LR), Neural Network (NN), and Random Forest (RF). The feature selection techniques examined include F-test Feature Selection, Mutual Information (MI), and Spearman Correlation Coefficient. The dataset used for the experiments is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is publicly available from the UCI Repository. The findings reveal that when feature selection is implemented, the LR and NN algorithms demonstrate superior accuracy and perform exceptionally well across other metrics compared to the other models.

KEYWORDS: Breast Cancer; Machine Learning; Feature Selection; Breast Cancer Diagnostic Dataset.

1. INTRODUCTION

Cancer is one of the deadliest diseases in the world. The latest statistics about this disease were reported in 2023 (Zhou *et al*., 2024), listing ten types of cancers, including breast cancer diagnosed in women. Breast cancer has been and still is the most common type of cancer that has affected a high percentage of women around the world at approximately 31%. It is considered the first type of cancer that causes deaths in women and is ranked fifth in terms of all cancer deaths around the world. It was the reason for 685,000 deaths in 2020, and that number increased to around 963,000 deaths in 2021, exceeding lung cancer with approximately 2.3 million new cases of this disease, according to the World Health Organization (WHO) (Bray *et al*., 2024). The percentage of these cancer cases was 25%, and the death cases among women were 17% around the world (Zhou *et al*., 2024). The abnormal growth of the breast cell is called a tumour, which is divided into two types: malignant and benign. The former is cancerous, while the latter is non-cancerous. Despite the incomprehension of the causes of breast cancer in women, several factors and attributes were contributed as the reasons for this disease, such as family history, problems in the inside uterine environment, adolescent exposures, pregnancy problems, gene mutation, alcohol and tobacco consumption, and childbearing at advanced maternal ages, specifically in developing countries (Uddin *et al*., 2023).

Consequently, to reduce the rate of breast cancer cases and to prevent mortality in women, it is important to make regular visits to health professionals for screening, treatment, and accurate examination in clinical health. However, misdiagnosis may occur, which reduces the opportunity for early recovery, or it may as well be that there is a shortage in the number of health experts. Also, the medical examining of the tumour is time-consuming and costly. Therefore, implementing techniques such as Machine Learning (ML) that automatically diagnose breast cancer is crucial. There are different examples of ML classification techniques that have been used to determine whether the breast tumour is cancerous or not, such as Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR), among others (Lappeenranta-, 2023) (Ak, 2020).

Furthermore, many researchers have used different ML classification algorithms for the prediction of breast cancer, underlining the importance of using such techniques for predicting the disease and showing challenges in this field (Ak, 2020; Mohammed *et al*., 2020; Nemade *et al*., 2022; Abunasser *et al*., 2023; and Ebrahim *et al*., 2023). Whereas some others, namely (Chen *et al*., 2023) (Botlagunta *et al*., 2023) and (Laghmati *et al*., 2024) have analyzed different breast cancer datasets, such as the Wisconsin Original Breast Cancer and Wisconsin Diagnostic Breast Cancer datasets (WDBC)Click or tap here to enter text. for this purpose and have obtained significant results (Wolberg, 1995).

In this paper, the primary goal is to implement different ML techniques to classify the patients as having cancer or not using the WDBC dataset and to obtain the accuracies of the models. Then, the following goal is to explore the influence of using the F-test, Mutual Information (MI), and the Spearman correlation coefficient feature selection techniques on the accuracy of the selected ML. This can be accomplished by comparing the results of both implementations as well as comparing the results obtained from implementing the feature selection methods with each other. Seven different machine learning algorithms, K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression (LR), Neural Networks (NN), and Random Forest (RF) are used for this investigation. Furthermore, some other methods are used to improve the performance of the selected models. For instance, Synthetic Minority Over-sampling (SMOTE) (Chawla *et al.,* 2002), is used to prevent the issue of imbalanced classes, using feature scaling to ensure that all features contribute equally to the models, and cross-validation to reduce overfitting and enhance model performance. The performance of the ML models is

evaluated using different evaluation metrics such as accuracy, F1 score, precision, recall, ROC AUC, and Matthew's correlation coefficient (MCC).

The rest of the paper is organized as follows: in Section 2, the most relevant works to this study are included, specifically those publications that have compared different ML methods using the WDBC dataset. The methodology and the applications used in the study are presented in Section 3. Section 4 illustrates the results of this study both prior to and after the implementation of the feature selection methods. Section 5 compares the outcomes of implementing the selected feature selection methods for each of the seven models used in this study. Section 6 is the last section that includes the conclusion and future works.

2. RELATED WORKS

Recently, with advances in medical research, different ML algorithms have been suggested to assess the classification of breast cancer data. Breast cancer is one of the common medical data that researchers have used for this purpose. These data can be obtained from breast cancer data repositories. In this section, a review of the publications related to the prediction of breast cancer, specifically the Wisconsin Diagnostic Breast Cancer (WDBC), is presented and surveyed, as shown in Table 1.

Ak, M. F. (2020) accomplished a comparative study to analyze the performance of different machine learning techniques, LR, KNN, SVM, and DT, using a graphical program named CITY for data visualization and samples from breast cancer patients in the WDBC dataset. The results of the study indicated that LR outperformed the other techniques, with the highest classification accuracy at 98.1%. In (Mohammed *et al*., 2020), the performance of three different ML algorithms, DT, NB, and Sequential Minimal Optimization (SMO), was compared using two different breast cancer datasets: the Wisconsin breast cancer and the Original Wisconsin Breast Cancer datasets (WBC). The study included a number of pre-processing steps to improve the performance of the ML techniques further, such as discretization and removing records that have missing data. The results showed that the algorithm SMO outperformed the other two classifiers with an accuracy of 99.56% on the WBC dataset.

Chaurasia *et al*. (2020) proposed a new method called Mode to remove frequent features from the WDBC and then applied an ensemble technique with stacking classifiers to categorize records with all features in comparison to the reduced data subset and to enhance accuracy. Their results showed that their proposed method increased the breast cancer accuracy to over 90%. Moreover, Islam *et al.* (2020) aimed to compare five ML algorithms: SVM, KNN, RF, ANN, and LR to diagnose breast cancer using the Wisconsin Breast Cancer dataset. The results of their study revealed that ANN outperformed other techniques, achieving the highest accuracy, 98.57%.

Naji *et al*. (2021), on the other hand, explored the ability of five ML algorithms to predict cancer in the WDBC dataset. Their results showed that the SVM technique surpassed other models by obtaining the highest accuracy, 97.2%. The authors revealed that the prediction of breast cancer using ML algorithms is possible; however, they acknowledged limitations and planned to explore larger datasets for improved accuracy and ethical implications.

Furthermore, Ara *et al*. (2021) explored ML algorithms for categorizing breast tumours as cancer or not using the WDBC Dataset. Training and testing techniques were used in the study, and the number of features was reduced, keeping only the highly correlated features to the target to improve the model's performance. Among the ML techniques used, the researchers concluded that RF and SVM models outperformed the other models by obtaining an accuracy of 96.5%.

Sakib *et al*. (2022), used two types of prediction techniques to predict and diagnose breast cancer in the WDBC: ML and Deep Learning (DL). They used different evaluation metrics to assess the performance of the models used for classification. The metrics used were accuracy, recall, specificity, precision, false-negative rate (FNR), false-positive rate (FPR), F1-score, and Matthews Correlation Coefficient (MCC). The results showed that the performance of the RF classifier was the highest based on the accuracy obtained, 96.66%. Chen *et al*. (2023), studied machine learning algorithms — XGBoost, random forest, logistic regression, and K-Nearest Neighbour (KNN) — for breast cancer classification, emphasizing recall for early detection. Using a dataset named WDBC from the UCI repository, they applied Z-score standardization and Pearson correlation for feature selection and addressed data imbalance through hierarchical sampling. Evaluating model performance with 80:20 and 70:30 splits, the XGBoost model outperformed others at 80:20, achieving a recall of 100%, precision of 96.0%, accuracy of 97.4%, and F1-score of 98.0%. The study noted performance variability across splits and the limitations of a universal machine learning approach in diagnostics.

The literature uses various methods and preprocessing procedures to compare and obtain the best performance of the models used. Their results are satisfactory in terms of accuracy and other metrics obtained. However, in this study, several different experimental settings are implemented to compare various machine learning algorithms and assess the impact of feature selection methods on model performance, making the findings of this study different from existing ones.

Table 1: A Survey of the Related Research Used in This Study

Reference/ Authors	ML Algorithms	Feature Selection Methods	Dataset	Methodology Used	Results
(Ak, 2020)	LR, KNN, SVM, and DT	Features selected by creating 3 datasets 1: data with all the features 2: with highly correlated features 3: with low correlated features	WDBC	A comparative analysis and new data visualization technique (CITY)	Accuracy: 98.1% for LR for dataset1 97.4 for the dataset2 and 95.6% for the dataset 3
(Mohammed *et al*., 2020)	DT, NB, and SMO	None	WBC and original breast cancer dataset	A comparison, discretization, and removing records with missing data	Accuracy: 99.56% for SMO using WBC dataset
(Chaurasia *et al*., 2020)	ensemble technique: AdaBoost, Gradient Boosting Classifier, RF, Extra Tree (ET)	Statistical method of feature selection 'Mode' to reduce the dataset to have 12 features only out of 32 features	WDBC	proposed method named Mode to reduce the dataset features	Accuracy over 90%

	Bagging and Extra Gradient Boost (XGB). stacking classifiers LR, DT, SVC, KNN, RF and NB				
(Islam *et al.*, 2020)	SVM, KNN, RF, ANN and LR	None	WBC	A comparison study	Accuracy: 98.57%. for ANN, precision of 97.82% and F1 score of 98.90%
(Naji *et al.*, 2021)	SVM, RF, LR, DT and KNN	Feature extraction method with no details	WDBC	A comparison study	Accuracy: 97.2% for SVM
(Ara *et al.*, 2021)	SVM, LR KNN, DT, NB and RF	All the dataset features used	WDBC	A comparison study and the correlation between different features of the dataset has been analyzed for feature selection	Accuracy: 96.5% for RF and SVM
(Sakib *et al.*, 2022)	SVM, DT, LR, RF, KNN, and a DL for classification using cross-validation.	None	WDBC	A comparative study	Accuracy: 96.66% for RF
(Chen *et al.*, 2023)	XGBoost, RF, LR, and KNN	Z-score for standardization and Pearson correlation for feature selection	WDBC	Predicting and classifying along with data preprocessing and feature selection	Accuracy of 97.4%, Recall of 100%, precision of 96.0%, and F1-score of 98.0%.

3. METHODS AND APPLICATIONS

In this section, the dataset used for conducting the experiments is explained. Then, each of the ML techniques used in this study is presented, followed by the section presenting the feature selection methods used. All of these techniques are presented to show the most effective combinations of them for predicting breast cancer and to evaluate the performance obtained using such combinations. The evaluation is accomplished using the evaluation metrics to assess the effectiveness of each model to provide valuable insight into enhancing predictive capabilities for breast cancer diagnosis and treatment.

Dataset Description

The dataset used in this study is called Breast Cancer Wisconsin Diagnostic (WDBC) (Wolberg, 1995). It can be accessed from UCI Machine Learning Repository. WDBC dataset consists of 569 samples designed for binary classification that is distributed between 357 benign and 212 malignant breast tumours collected from fine needle aspiration (FNA) biopsy images. In other words, the distribution of the dataset is 62.7% non-cancerous and 37.3% cancerous breast lesions. The dataset has 30 attributes, which represent the measurements of the shape of cell nuclei, such as radius, roughness, and smoothness. Table 2 shows the dataset's features and their description as illustrated by (Kumar *et al.*, 2021). This paper uses this dataset because its attributes can describe the symptoms effectively. Therefore, it is considered as a good resource for diagnosing breast cancer and examining the feature selection and model performance. However, the classes of the WDBC dataset are imbalanced; therefore, they need to be resampled. In this work, the Synthetic Minority Over-sampling (SMOTE) statistical technique is used for resampling the dataset, and the results are shown in Figure 1. The figure shows the distribution of the attributes to the target before and after resampling, which is equal to 50% for each class. Moreover, a smaller dataset is created by removing the attributes that have a weak correlation to the target to avoid noise and the models' inaccurate prediction, resulting in each record having a patient ID, a diagnosis, and 23 real-valued attributes. To ensure that all the features of the new dataset contribute equally to the prediction of the models, the attributes are scaled using the feature scaling method 'Standard Scaler' to have zero mean and a standard deviation.

Table 2: Summary of Wisconsin Diagnostic Breast Cancer (WDBC) Dataset (Kumar *et al.,* 2021).

|	Measurement range			
Attributes	**Mean**	**Standard Deviation**	**Maximum**	**Attribute description**
Radius	6.99–28.12	0.121–2.923	7.95–37.01	Calculated as the average of distances from the center to points on the perimeter
Texture	9.80–40.02	0.37–4.90	112.10–50.01	Calculated as the standard deviation of Gray-scale values.
Perimeter	44.02–189.09	0.80–22.01	50.48–252.03	The total distance between consecutive points in a contour or outline.
Area	144.04–2503.01	6.90–543.10	186.01–4255.00	Calculated Number of conductive points in an outline
Smoothness	0.054–0.164	0.003–0.035	0.072–1.102	calculated as the local variation in radius lengths
Compactness	0.020–0.350	0.002–0.138	0.030–1.060	Calculated as the ratio of perimeter squared to area minus 1
Concavity	0.001–0.501	0.000–0.400	0.000–1.255	The severity of concave portions of the contour.
Concave Points	0.0001–0.202	0.000–0.055	0.000–1.296	Number of concave portions of the contour.
Symmetry	0.108–0.305	0.009–0.080	0.158–0.668	
Fractal dimension	0.051–0.098	0.001–0.031	0.057–0.210	Coastline approximation minus 1
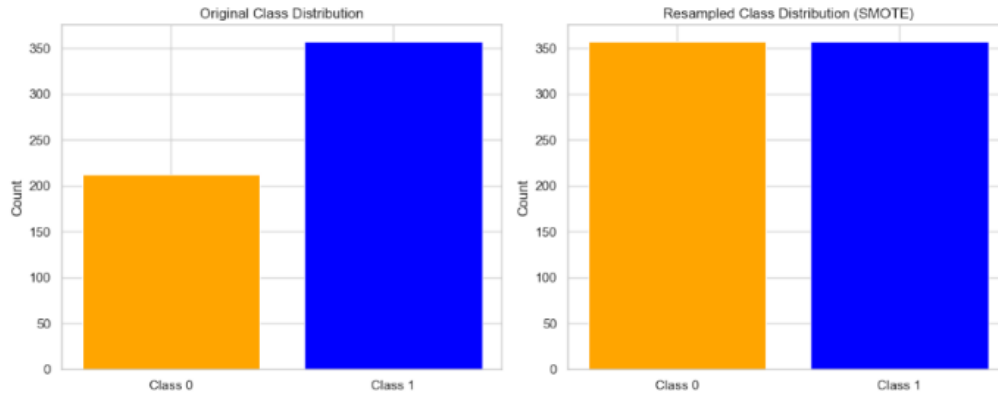
Figure 1: The Distribution of the WDBC before and after Resampling across the Target Variable.

Machine Learning Algorithms

Decision Tree (DT): the DT algorithm (Mohammed *et al.,* 2020) is a supervised ML algorithm that is mainly used for classification and regression. The input node is the main feature of this technique. Its structure consists of a root node, where it is at the top of the tree, an internal node representing the input features, and a leaf node representing the decision node or the class of the dataset located at the bottom of the DT, as shown in Figure 2. The hierarchical structure of the DT is made up of a number of nodes at different levels, and the small trees that can be extracted from the main tree are called subtrees. The larger the tree, the more difficult the classification of data accurately is due to problems such as overfitting and data splitting. These problems can be tackled by using techniques such as pruning, cross-validation, and ensemble techniques to integrate multiple trees.
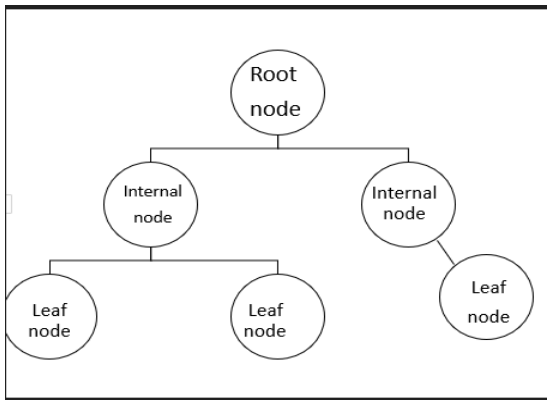

Figure 2: The Structure of Decision Tree Algorithm (Mohammed et al., 2020)

Logistic Regression (LR): Logistic Regression (LR) (Ak, 2020; Dhanya R, 2019; and Hossin *et al*., 2023) is an ML technique used for predicting two values, 0 or 1, and classifying data derived from a linear combination of data. The parameter coefficient values can be calculated using both linear regression and logistic regression. In the case of logistic regression, gradient descent can be used for this purpose. The LR algorithm utilizes some techniques to overcome the problems of overfitting and bias, such as cross-validation and regularization. Generally, LR is a simple and strong technique in solving classification problems.

Naïve Bayes (NB): Naive Bayes (NB) is a strong supervised classification technique used for classifying large and complex data using a small size of training data (Dhanya R, 2019; Hossin *et al*., 2023; and Kadhim *et al*., 2023). It is an easy method and is based on the theorem called Bayes, which assumes conditional independence between each two features and a given class. NB calculates the probability theory in a simpler way. It can also tackle the risk of data noise and overfitting for its reliance on strong independence assumptions. The NB equation can be represented as follows:

$$y = argmax_y \, P(y) \prod_{i=1}^{n} P(x_i|y)$$
(1)

Since the features of WDBC data are integers and follow a normal distribution, in this study Gaussian Naive Bayes type is used (Eq. 2).

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(- \frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$
(2)

Random Forest (RF): Random Forest is an ensemble learning technique that is used for classification and regression (Dhanya R, 2019; Hossin *et al*., 2023; and Kadhim *et al*., 2023). The term 'Random Forest' refers to the group of decision trees that are created from subsets of training data randomly instead of creating a single tree during the preprocessing step. The created group helps to tackle noises in data, which in turn reduces the effect of overfitting, improves the performance and the generalization of the models, and obtains better accuracy results. Therefore, RF is considered one of the best solutions for many ML applications.

Support Vector Machine (SVM): Support Vector Machine is a supervised ML learning method utilized for classification and regression problems (SVM) (Ak, 2020; Hossin *et al*., 2023). It is also known as a powerful method to detect outliers and noises in data. It works by finding an n-dimensional separation hyperplane that helps to classify data inputs into a similar and non-similar class, as shown in Figure 3. The maximum the margin in the SVM classifier between classes, the better the hyperplane to compare more than two features for classification and then produce accurate findings. Furthermore, the closer the support vectors are to the hyperplane, the more the ability of the SVM classifier to reduce overfitting, ensuring the generalization of the model to new data properly.

$$L(\omega) = \sum_{i=1} \max(0.1 - y_i \, [\omega^T x_i + b]) + \gamma\|\omega\|_2^2$$
(3)

where $\sum_{i=1} \max(0.1 - y_i \, [\omega^T x_i + b])$ is the loss function and $\gamma\|\omega\|_2^2$ is the regularization.
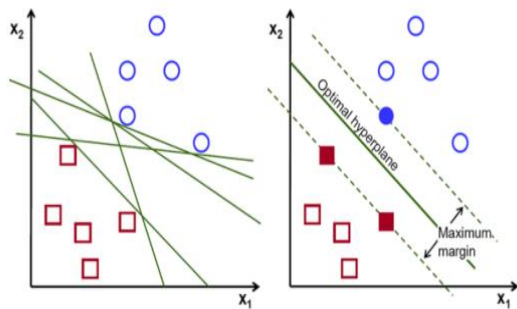
Figure 3: The Illustration of SVM Algorithm (Ak, 2020)

K-Nearest Neighbour (KNN): K-Nearest Neighbors (KNN) is also a supervised learning technique used for classification and regression tasks (Ak, 2020; Hossin *et al*., 2023). The term 'nearest neigbors' means the numerical value of 'k', which represents the nearest data points determined in a dataset for prediction using majority voting for classification or averaging for regression. The value of the 'k' also determines the degree of the model performance, and its low value causes overfitting due to the noise capturing in the data, whereas the high value leads to the model generalization and produces an accurate prediction. KNN predictions are based on distance metrics such as Euclidean distance, which calculates the distance between each of the two data points in the datasets to reduce noise and the risk of overfitting. KNN is a direct and flexible method that requires accurate tuning to obtain the best model performance. Figure 4 shows an example of the KNN technique (Ak, 2020).
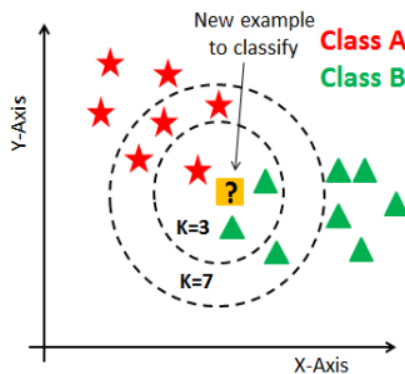

Figure 4: An Example of KNN Classifier (Ak, 2020).

Neural Networks (NN): Neural networks is another supervised ML technique used for classification problems (Mahesh, 2020). The simple structure of NN consists of an input layer, one hidden layer, and an output layer, as shown in Figure 5 (Yadav *et al*., 2022). If more than one hidden layer exists in the NN algorithm, then it will be defined as a deep learning algorithm. The NN layers are connected to each other, which consist of artificial neurons that work together to find a solution to a problem similar to that in the human brain and are used to detect patterns in data. The neuron of the NN layers works by processing and analyzing the data and then passing its output to the hidden layer, which further processes the incoming output and passes its output to the output layer. The final layer might have more than one output neuron, depending on the problem solved. A function named a loss function is used to evaluate the performance of the NN by calculating errors in the estimation process. The lower the value of the loss function, the more effective the NN prediction is. To avoid noises and overfitting in data, several techniques can be used, such as dropout, early stopping, and regularization.
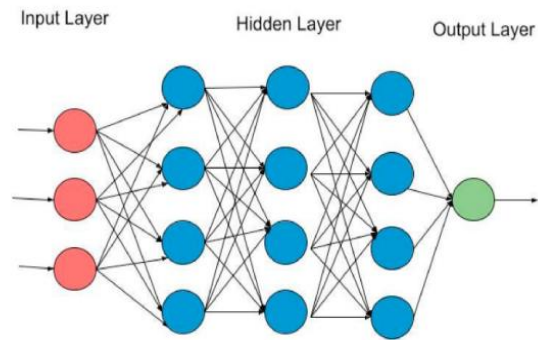

Figure 5: The Structure of an Artificial Neural Network (Yadav Et Al., 2022).

Feature Selection Methods

F-test: In statistics, the F-test is a statistical feature selection technique that computes the difference amount between two or more subsets of data (Dhal *et al*., 2022). The F-test is known to be an effective method to deal with data with high dimensionality, such as medical datasets. Therefore, this method is useful to be used for selecting features such as tumour characteristics to determine whether the patient has cancer or not in a dataset like breast cancer. Moreover, selecting the only relevant attributes by F-test helps to reduce the classes' overfitting and increase the models' performance. The F-test formula is $(s1^2/s2^2)$, where $s1^2$ is the variance of the first sample set, and $s2^2$ is the variance of the second sample set.

Mutual Information (MI): Mutual Information (MI) is another statistical feature selection method that is used to find linear and non-linear associations in sophisticated datasets such as medical datasets (Dhal *et al*., 2022). Therefore, it is a useful approach in many fields, such as ML for modeling and healthcare for diagnosing and treatment. In MI, one feature provides valuable information about another feature, i.e., it measures how dependent each of the two variables is, whereas zero MI means no dependency is available. More than zero means a dependency between the two attributes (Vergara *et al*., 2015). MI provides the most related features to the target of the dataset used which helps to increase the ML model performance. In medical datasets like breast cancer, this process is important in clarifying the relation between the dataset variables, which are mostly ambiguous.

Spearman Correlation Coefficients: In Spearman Correlation Coefficients, the robustness of two-variable correlations is specified such that it can have positive or negative and weak or strong values (Dhal *et al*., 2022). Statistically, it is a non-distribution rank measure, i.e., it measures the correlation between variables without considering the distribution of the data (Hauke *et al*., 2011). This feature makes the Spearman correlation coefficients valuable, specifically in medical datasets where the frequency of the data distribution is not considered. In other words, it can simply find the relation between the input variables and the target variable. Based on that, the Spearman correlation coefficients method is a helpful measure for researchers working in medical fields to decide treatment and diagnosis for patients.

Evaluation Metrics

Models' performance, strengths, and weaknesses are essential to be evaluated. Hence, in this study, the selected classification models using different evaluation metrics commonly used in the literature are evaluated. The metrics used are:

- Accuracy (Patro et al., 2021) , mathematically shown as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \ (\%) \qquad (4)$$

- F1 score (Lichtenwalter et al., 2010), mathematically shown as follows:

$$F1 \ score = \frac{2*(precision*recall)}{(precision+recall)} \qquad (5)$$

where:

$$precision = \frac{true \ positives}{(true \ positives + false \ positives)} \qquad (6)$$

$$recall = \frac{true \ positives}{(true \ positives + false \ negatives)} \qquad (7)$$

- Matthew's correlation coefficient (MCC) (Ali et al., 2021),

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \qquad (8)$$

- The area under the curve (AUC) of receiver operating characteristics (ROC) is another statistic that is used to evaluate models (Shiny Irene *et al.*, 2020). It assesses the performance of classification models using threshold values ranging from 0 to 1, signifying poor to excellent predictions.

Accuracy is the proportion of participants accurately predicted by the classification model relative to the total number of tested subjects. Precision and recall (Eqs. 6 and 7) are combined into the F1 score (Eq. 5), which is frequently used in binary classification problems. Precision is defined as the ratio of true positive forecasts to all positive predictions, whereas recall is the ratio of true positive predictions to all predictions of positive data that are actually seen. The F1 score ranges within the interval [0, 1], with a score of 1 indicating perfect precision and recall.

MCC is an important statistical measure for assessing the accuracy of binary categorization. It only gives an outstanding grade if the prediction performs well in all four elements of the confusion matrix: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). As a result, it is frequently viewed as a balanced metric that may be used even when the classes are of vastly different sizes. Its value ranges from -1 to 1, with -1 indicating that all tested participants were wrongly or correctly predicted and 0 indicating that the model prediction is no better than a random guess.

4. EXPERIMENTAL RESULTS

In this part of the study, two different types of experiments are conducted on the created dataset from the WDBC dataset as described in section 3.1. In the first experiment, the selected ML algorithms are implemented on the dataset without the incorporation of the feature selection methods F-test, MI, and Spearman Correlation and without resampling the data. In the second experiment, the ML models are implemented using the three feature selection methods and the resampling method SOMTE. Moreover, the performance of the models is analyzed and compared in Section 5 using the feature selection methods as illustrated in Tables 7-13. Different assessment metrics are used for the models' performance evaluation: accuracy, F1 score, precision, recall, ROC AUC, and MCC.

In this work, all the experiments are accomplished using a computer system having the following features: Intel(R) Core (TM) i3-2310M CPU @ 2.10GHz, 4 GB of RAM, and 64-bit Windows 10 Pro OS. Also, the Python programming language is used to conduct the experiments with its libraries or packages such as NumPy, Pandas, and Scikit-Learn. For each experiment, the dataset is divided into ten parts of the same size in order to train and test the models efficiently and to make the models more reliable for prediction. The process of dividing the dataset is called cross-validation, which splits the data into 10 equal parts. Moreover, the default hyperparameters are used for each model, such as setting the value of k to 5 in KNN; one hidden layer and a maximum of 5000 iterations are set in NN; and lastly, in LR, 1000 iterations are established.

Results without feature selection

In this section, the results of the implementation of the selected ML models on the reduced WDBC dataset are presented. This is done without using the resampling technique and the selected feature selection methods. The findings of the classification are illustrated in Table 3 and Figure 6 for the selected evaluation metrics used in this study.

As shown in Table 3, the performance of the NN and LR models shows the highest accuracy values of 0.978910 and 0.977153 compared to those of the NB model, which is poor. Also, the two models provide a high balance between true positive and false positive scoring, with high F1 values of 0.978869 and 0.977096, respectively. In terms of precision and recall, NN, LR, and SVM achieve the highest discrimination of the class labels. Moreover, the highest ROC AUC value is also for NN, with a score of 0.975530 compared to the DT's, which is the lowest value of 0.907424. The obtained results reveal that NN and LR are the best models in performing classification on the WDBC dataset compared to the other models. Lastly, the MCC values show that NN and LR also outperform the other models, specifically NB and DT, in providing a balanced score, achieving 0.954827 and 0.951067, respectively, while NB and DT have the lowest values of MCC.

Table 3: Evaluation Metrics Comparison for Wisconsin Diagnostic Breast Cancer Without Feature Selection

| Model	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
DT	0.913884	0.913842	0.913807	0.913884	0.907424	0.815636
KNN	0.964851	0.964707	0.964976	0.964851	0.958578	0.924663
LR	0.977153	0.977096	0.977202	0.977153	0.973171	0.951067
NB	0.933216	0.933015	0.933036	0.933216	0.925704	0.856551
NN	0.978910	0.978869	0.978931	0.978910	0.975530	0.954827
RF	0.968366	0.968303	0.968343	0.968366	0.964253	0.932183
SVM	0.973638	0.973599	0.973618	0.973638	0.970370	0.943512
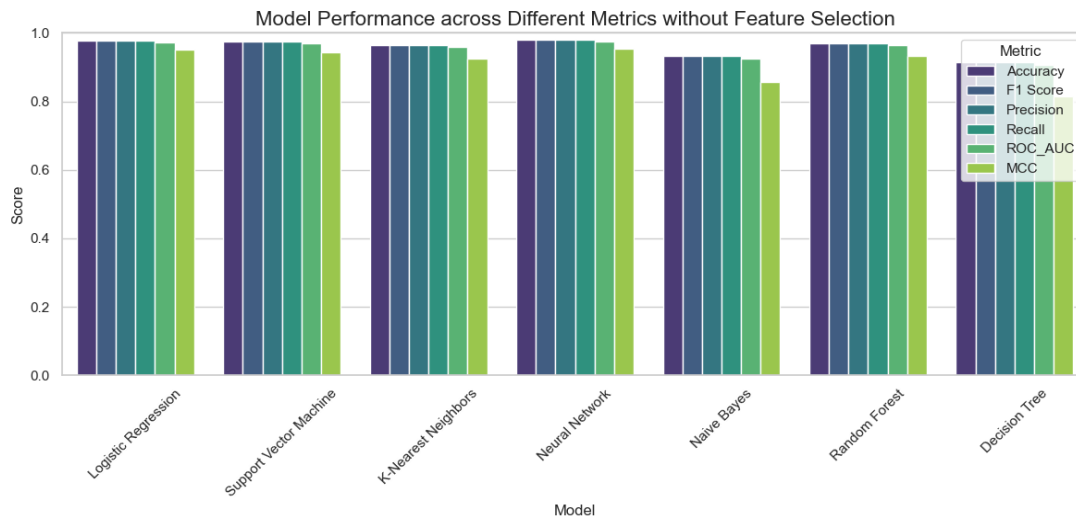
Figure 6: Model Performance across Different Metrics without Feature Selection Methods.

Results with Feature Selection

F-test: This section presents the results of combining F-test feature selection methodology with machine learning techniques. The dataset is resampled and scaled, and the features are selected. The classification results show accuracy, precision, recall, MCC, and ROC-AUC assessment metrics, as presented in Table 4 and Figure 7.

Table 4 shows a detailed analysis of how well different models classify data based on the metrics produced from the F-test feature selection procedure. Among the models tested, LR and NN have the maximum accuracy, scoring 0.985994 and 0.981793, with the F-test technique selecting only 20 and 23 features, respectively. The obtained high performance of these two models reveals that both models are reliable for accurate prediction of the classes and also that the number of selected attributes is enough to obtain such good performance, which means it is not necessary to use all the features of the dataset. On the other hand, the F test method selected 18, 20, and 20 features for the KNN, RF, and SVM models of 0.976190, 0.976190, and 0.977591, respectively. This also indicates that the selected features are good enough to obtain such good performance;

however, these performances are still lower in accuracy compared to those in NN and LR. In terms of precision and recall, both LR and NN outperform the other models in detecting the true positives and reducing the false negatives, with LR's recall of 0.988796 and precision of 0.983287, providing a balance to distinguish between positive and negative records.

The NB model, on the other hand, underperforms the other models by obtaining the lowest accuracy of 0.946779 and F1 score of 0.947514. This indicates that NB is not suitable enough for predicting cancer in the WDBC dataset.

Moreover, LR and NN provide a strong F1 score, which balances between false positives and false negatives, with LR receiving the best score of 0.986034 and NN, with 0.981818. Furthermore, LR, NN, and three other models KNN, RF, and SVM produce a good score in terms of ROC AUC values. The metrics values are over 90%, indicating that the models are capable of distinguishing between classes. The final metric, MCC, is considered as a reliable metric for the prediction process and for considering all categories of the confusion matrix since it provides the highest values for LR with 0.972004 and NN with 0.963589.

Table 4: Evaluation Metrics Comparison for Wisconsin Diagnostic Breast Cancer with F-test Feature Selection Method.

| Model	No. of Selected Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
DT	16	0.959384	0.958982	0.968571	0.949580	0.959384	0.918944
KNN	18	0.976190	0.976224	0.974860	0.977591	0.976190	0.952385
LR	20	0.985994	0.986034	0.983287	0.988796	0.985994	0.972004
NB	3	0.946779	0.947514	0.934605	0.960784	0.946779	0.893908
NN	23	0.981793	0.981818	0.980447	0.983193	0.981793	0.963589
RF	20	0.976190	0.976157	0.977528	0.974790	0.976190	0.952385
SVM	20	0.977591	0.977591	0.977591	0.977591	0.977591	0.955182
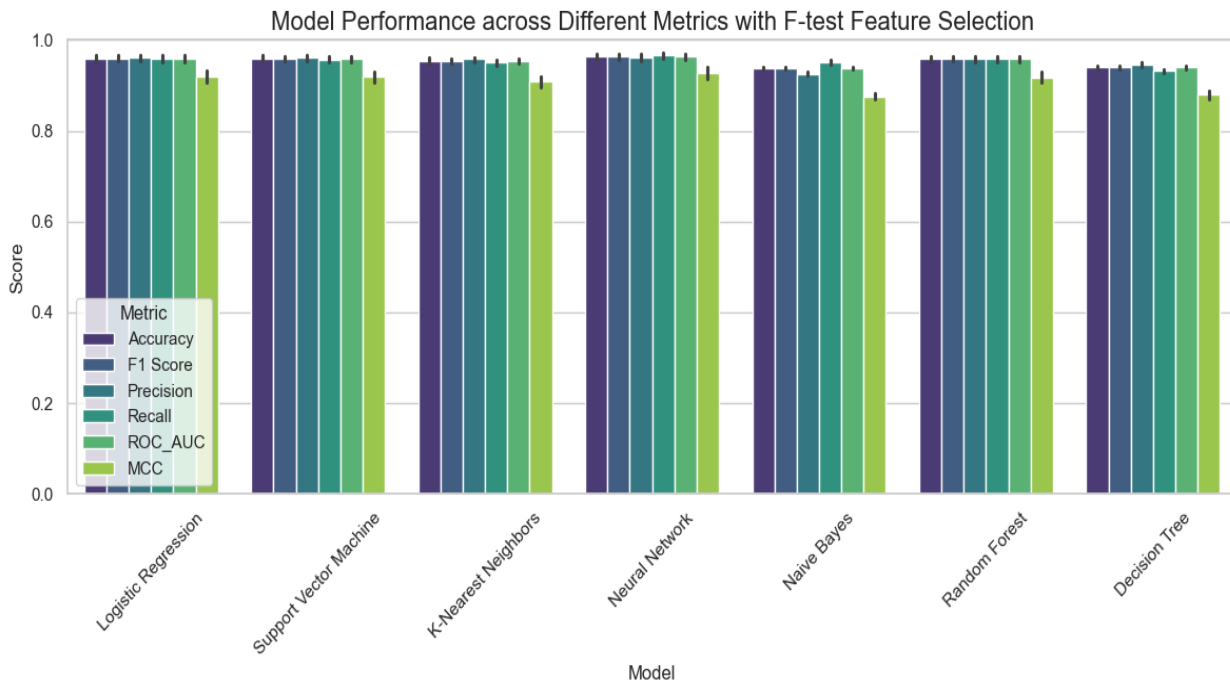
Figure 7: Model Performance Across Different Metrics with F-test Feature Selection Method

Mutual Information (MI): The second feature selection method used in this study is MI, which is combined in this section with the ML techniques to show its impact on the performance of the models. The prediction process includes resampling the dataset, scaling, and then using the MI for selecting features. The classification results are then evaluated using the evaluation metrics selected for this study, as shown in Table 5 and Figure 8. In the context of classification tasks, each metric in Table 5 provides a different perspective on the model's ability to make accurate predictions. The best-performing model is LR, which achieves the highest scores across practically all metrics using all the dataset's attributes, as selected using the MI approach. With an accuracy of 0.983193 and an F1 score of 0.983287, LR demonstrates outstanding ability to balance precision with recall, which is critical for effective classification. LR's precision of 0.977839 and recall of 0.988796 shows that it not only generates accurate positive predictions but also correctly recognizes a large number of true positive events. The ROC_AUC score of

0.983193 highlights its ability to discriminate across classes, while the MCC of 0.966447 indicates a significant connection between predicted and actual results. KNN is a reliable model, with an accuracy of 0.964986; however, it falls short of LR in terms of overall effectiveness. The NN model achieves an impressive accuracy of 0.978992, demonstrating its ability to sustain high precision (0.975000) and recall (0.983193). The SVM and RF models perform well, with SVM reaching an accuracy of 0.976190 and RF scoring 0.970588. Despite their competitiveness, these models do not outperform the other ones in any specific category, implying that they may be more appropriate for certain contexts rather than being the top choices overall. Furthermore, the investigated models indicate that the DT model has the lowest performance, with an accuracy of 0.959384. On the other hand, the NB model dramatically underperforms, with the lowest scores across most metrics and an accuracy of 0.929972. This suggests that NB may not be appropriate for this dataset.

Table 5: Evaluation Metrics Comparison for Wisconsin Diagnostic Breast Cancer with MI Feature Selection Method

| Model	No of Selected Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
DT	23	0.959384	0.959327	0.960674	0.957983	0.959384	0.918771
KNN	23	0.964986	0.965132	0.961111	0.969188	0.964986	0.930005
LR	23	0.983193	0.983287	0.977839	0.988796	0.983193	0.966447
NB	23	0.929972	0.931694	0.909333	0.955182	0.929972	0.861039
NN	23	0.978992	0.979079	0.975000	0.983193	0.978992	0.958017
RF	23	0.970588	0.970547	0.971910	0.969188	0.970588	0.941180
SVM	23	0.976190	0.976290	0.972222	0.980392	0.976190	0.952415
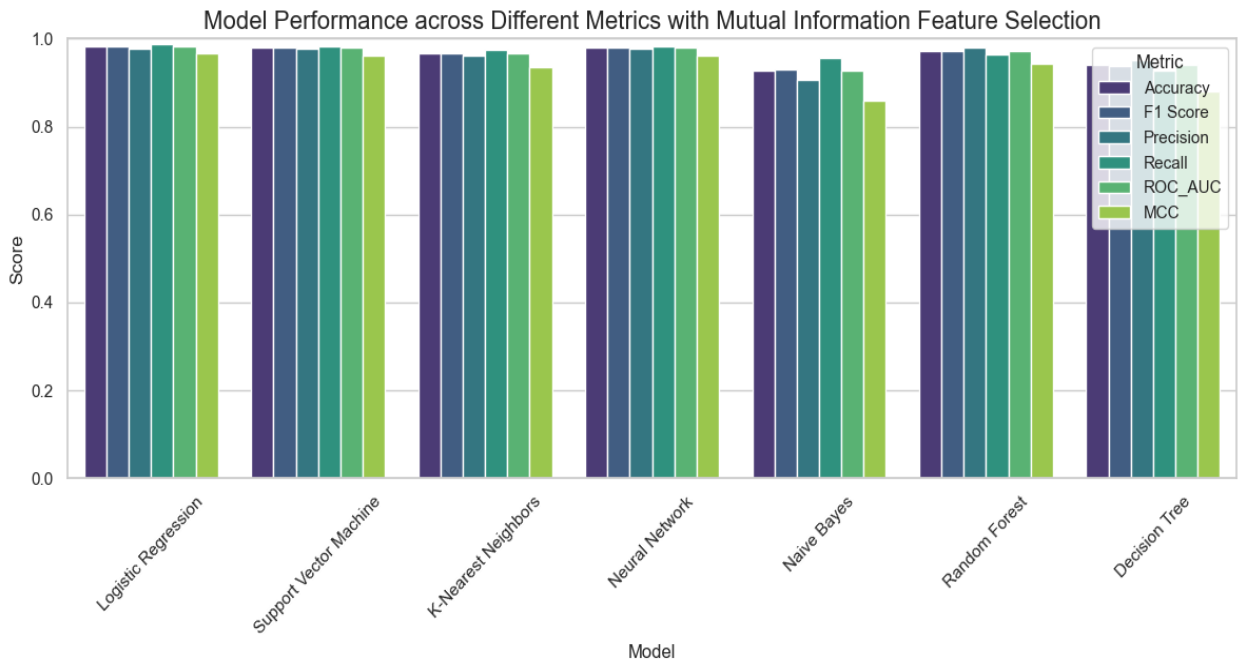
Figure 8: Model Performance Across Different Metrics with MI Feature Selection Method

Spearman Correlation Coefficient: This section presents the results of combining Spearman Correlation Coefficient feature selection methodology with machine learning techniques. The dataset is resampled, scaled, and the features selected. The classification results show accuracy, precision, recall, MCC, and ROC-AUC assessment metrics, as shown in Table 6 and Figure 9. In Table 6, the results of the assessment metrics used to evaluate the models' performance in this study are presented. This is to show the impact of combining the Spearman method with the ML models. In the table, the LR model outperforms the other models, obtaining the highest accuracy and F1 score using 23 features. The accuracy is 0.984594, and the F1 score is 0.984658. In terms of other metrics such as precision and recall, as well as the MCC, LR show competitive values with a precision of 0.980556, a recall of 0.988796, and an MCC of 0.969222. These findings indicate that the capability of LR to correctly classify the positive cases and almost all of them are correct. They also provide a strong association between the predicted and real outcomes. Also, LR demonstrates the high capability of differentiation between the target classes by providing a ROC AUC value of 0.984594.

On the other hand, the table shows the results of both NN and KNN, which are considered to be close to each other in terms of accuracy and F1 score. NN has the highest values of accuracy of 0.981793 and F1 score of 0.981818 compared to KNN, which has the accuracy value of 0.970588 and F1 score of 0.970711. This indicates that KNN provides a good balance between precision and recall according to the value of the F1 score as well as the capability of NN to provide a stronger balance between precision of 0.980447 and recall of 0.983193. Despite the good results of both NN and KNN, they still fall below the performance of LR. Regarding the RF and SVM models, their assessment metrics results show that they are accurately measured in terms of accuracy and F1 score. Furthermore, the results of the DT model highlight that its performance is the lowest compared to the other models, with an accuracy of 0.950980 and an F1 score of 0.950495, indicating that DT is not suitable for undertaking the process of modelling incorporating the Spearman feature selection method. However, the performance of the NB model is the worst compared to DT and other models, achieving an accuracy of 0.938375 and lower precision and recall values, leading to the least favorable overall metrics.

Table 6: The Comparison of Evaluation Metrics for WDBC dataset using Spearman Feature Selection Method

| Model	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
DT	21	0.950980	0.950495	0.960000	0.941176	0.950980	0.902134
KNN	23	0.970588	0.970711	0.966667	0.974790	0.970588	0.941210
LR	23	0.984594	0.984658	0.980556	0.988796	0.984594	0.969222
NB	23	0.938375	0.939560	0.921833	0.957983	0.938375	0.877426
NN	21	0.981793	0.981818	0.980447	0.983193	0.981793	0.963589
RF	20	0.978992	0.979021	0.977654	0.980392	0.978992	0.957987
SVM	23	0.978992	0.979021	0.977654	0.980392	0.978992	0.957987
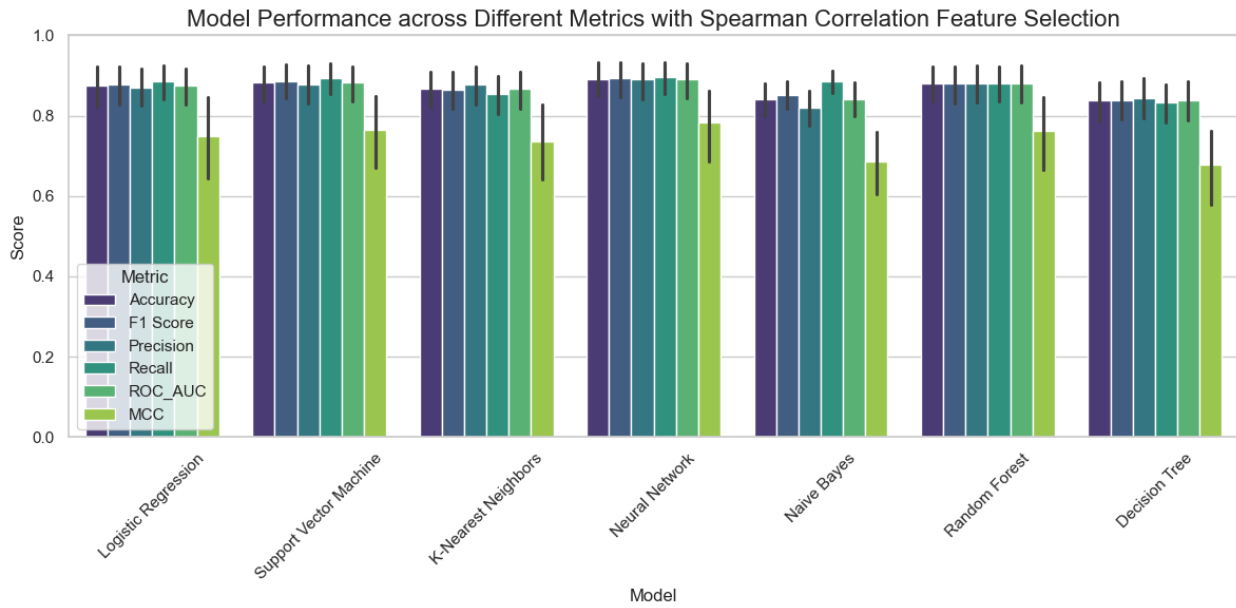
Figure 9: Model Performance Across Different Metrics with Spearman Feature Selection Method

5. COMPARISON OF THE ML MODELS USING THE FEATURE SELECTION METHODS

In this section, Tables 7-13 illustrate the results of comparing the performance of the ML models - DT, KNN, LR, NB, NN, RF, and SVM that are analyzed and selected for this study. The analysis and comparison are done using the feature selection techniques: F-test, MI, and Spearman, and are evaluated using the performance metrics: accuracy, F1 score, precision, recall, ROC_AUC, and Matthews Correlation Coefficient (MCC).

Table 7: The Performance of Decision Tree Classifier Using F-test, MI, and Spearman Feature Selection Methods

FS method/ DT	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
F-test	16	0.959384	0.958982	0.968571	0.949580	0.959384	0.918944
MI	23	0.959384	0.959327	0.960674	0.957983	0.959384	0.918771
Spearman	21	0.950980	0.950495	0.960000	0.941176	0.950980	0.902134

Table 8: The Performance of KNN Classifier Using F-test, MI, and Spearman Feature Selection Methods

FS method/ KNN	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
F-test	18	0.976190	0.976224	0.974860	0.977591	0.976190	0.952385
MI	23	0.964986	0.965132	0.961111	0.969188	0.964986	0.930005
Spearman	23	0.970588	0.970711	0.966667	0.974790	0.970588	0.941210

Table 9: The LR Classifier Performance Using F-test, MI, and Spearman Feature Selection Methods

FS method/ LR	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
F-test	20	0.985994	0.986034	0.983287	0.988796	0.985994	0.972004
MI	23	0.983193	0.983287	0.977839	0.988796	0.983193	0.966447
Spearman	23	0.984594	0.984658	0.980556	0.988796	0.984594	0.969222

Table 10: The NB Classifier Performance Using F-test, MI, and Spearman Feature Selection Methods

FS method/ NB	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
F-test	3	0.946779	0.947514	0.934605	0.960784	0.946779	0.893908
MI	23	0.929972	0.931694	0.909333	0.955182	0.929972	0.861039
Spearman	23	0.938375	0.939560	0.921833	0.957983	0.938375	0.877426

Table 11: The Performance of NN Classifier Using F-test, MI, and Spearman Feature Selection Methods

FS method/ NN	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
F-test	23	0.981793	0.981818	0.980447	0.983193	0.981793	0.963589
MI	23	0.978992	0.979079	0.975000	0.983193	0.978992	0.958017

| Spearman | 21 | 0.981793 | 0.981818 | 0.980447 | 0.983193 | 0.981793 | 0.963589 |

Table 12: The Performance of RF Classifier Using F-test, MI, and Spearman Feature Selection Methods

FS method/ RF	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
F-test	20	0.976190	0.976157	0.977528	0.974790	0.976190	0.952385
MI	23	0.970588	0.970547	0.971910	0.969188	0.970588	0.941180
Spearman	20	0.978992	0.979021	0.977654	0.980392	0.978992	0.957987

Table 13: The SVM Classifier Performance Using F-test, MI, and Spearman Feature Selection Methods

FS method/ SVM	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
F-test	20	0.977591	0.977591	0.977591	0.977591	0.977591	0.955182
MI	23	0.976190	0.976290	0.972222	0.980392	0.976190	0.952415
Spearman	23	0.978992	0.979021	0.977654	0.980392	0.978992	0.957987

By utilizing the F-test, the method in Table 9 achieves an impressive accuracy of 0.985994 and an F1 score of 0.986034, proving its robustness. In addition, it has comparable metrics to MI and Spearman. LR exhibits a balanced performance in classification tasks, as evidenced by its consistently high precision and recall. Likewise, the NN exhibits impressive results, especially when utilizing F-test (achieving an accuracy of 0.981793) as depicted in Table 11. It consistently achieves high precision and recall scores in both MI (accuracy of 0.978992) and Spearman (accuracy of 0.981793), validating its capability to handle complex patterns.

Table 8 demonstrates that KNN produces impressive results, particularly when compared to F-test (with an accuracy of 0.976190). However, it exhibits more variability when used with MI and Spearman (yielding accuracies of 0.964986 and 0.970588, respectively). While KNN demonstrates good performance, it does not outperform the top two models (LR and NN). RF algorithm consistently demonstrates strong performance, especially when paired with F-test (achieving an accuracy of 0.976190). Despite showing strong metrics, it still falls behind LR and NN, but performs similarly to KNN with F-test. In Table 13, SVM demonstrates good performance in both F-test and Spearman, with an accuracy of approximately 0.977591 and 0.978992, respectively. However, the accuracy achieved by MI is slightly lower – 0.976190. Additionally, SVM's robust recall scores validate its effectiveness in accurately detecting positive instances. On the other hand, when comparing the models, the DT is found to be the least effective performer, as indicated in Table 7. It achieves an accuracy of 0.959384 using F-test. Despite achieving high accuracy scores of 0.959384 in MI and 0.950980 in Spearman, the performance of the model does not match that of the other models. NB performs weakest among all models, consistently scoring below 0.943 in accuracy across all feature selection methods. The performance metrics as presented in Table 10 reveal a significant drop in precision and F1 scores, indicating that NB is less suited for this classification task compared to other models.

Based on the feature selection techniques used, the overall analysis of Tables 7–13 identifies NN and LR as the best models. SVM, RF, and KNN are reliable alternatives; however, they fall short of LR and NN in terms of performance. On the other hand, it is evident that DT and NB have limits when it comes to accurately capturing the underlying data complexity, which further confirms their lower suitability for this particular problem domain.

CONCLUSION AND FUTURE WORKS

In this study, a comparison was conducted to show the effect of various feature selection methods, F-test, MI, and Spearman correlation coefficients, to the performance of seven different machine learning techniques: KNN, NB, DT, SVM, LR, NN, and RF. The study evaluated the results using different evaluation metrics such as accuracy, F1 score, precision, recall, ROC AUC,

and MCC with the breast cancer dataset WDBC. It is concluded that the high performance of the models can be obtained by not using all the features of the dataset for prediction and improving modelling. Thus, feature selection methods were employed to select the features that really influence the model's performance and can predict whether a patient has cancer or not. Tables 7-13 presented the number of the selected features using the feature selection methods that yielded the best performance according to the evaluation metrics used.

The results indicate that LR and NN are the best models. SVM, RF, and KNN are reliable alternatives; however, they do not match the performance of LR and NN. On the other hand, it appears that DT and NB are not as effective in this particular problem domain as they could be when it comes to accurately capturing data complexity.

Future research needs to explore more feature selection methods, deep learning models, hyperparameter optimization, and diverse data types to improve breast cancer prediction and enhance machine learning model effectiveness in healthcare. Furthermore, the future work also needs to propose new hybrid algorithms from the ML models based on the results obtained in this paper. This can be accomplished by integrating different ML models to enhance the prediction of breast cancer in the healthcare sector and then use different feature selection methods to select the optimal features from the breast cancer datasets.

REFERENCES

Abunasser, B. S., AL-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2023). Convolution Neural Network for Breast Cancer Detection and Classification Using Deep Learning. *Asian Pacific Journal of Cancer Prevention*, *24*(2), 531–544. DOI: 10.31557/APJCP.2023.24.2.531

Ak, M. F. (2020). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. *Healthcare (Switzerland)*,*8*(2).DOI:10.3390/healthcare8020111

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. DOI: 10.1016/j.compbiomed.2021.104672

Ara, S., Das, A., & Dey, A. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. *2021 International Conference on Artificial Intelligence, ICAI 2021*, 97–101. DOI: 10.1109/ICAI52203.2021.9445249

Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classification and diagnostic prediction of breast cancer metastasis on clinical

data using machine learning algorithms. *Scientific Reports*, *13*(1). DOI: 10.1038/s41598-023-27548-w

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*,*74*(3),229263.DOI:10.3322/caac.21834

Chaurasia, V., & Pal, S. (2020). Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer. *SN Computer Science*, *1*(5). DOI: 10.1007/s42979-020-00296-8

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In Journal of Artificial Intelligence Research (Vol. 16).

Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. *Computational Intelligence and Neuroscience*, *2023*, 1–9. DOI: 10.1155/2023/6530719

Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, *52*(4), 4543–4581. DOI: 10.1007/s10489-021-02550-9

Dhanya R, I. R. P. S. S. A. M. S. and J. J. N. (2019). *A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection*.

Ebrahim, M., Sedky, A. A. H., & Mesbah, S. (2023). Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer. *Data*, *8*(2). DOI: https://doi.org/10.3390/data8020035

Hauke, J., & Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*,*30*(2),8793.DOI:10.2478/v10117-011-0021-1

Hossin, M. M., Javed Mehedi Shamrat, F. M., Bhuiyan, M. R., Hira, R. A., Khan, T., & Molla, S. (2023). Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. *Bulletin of Electrical Engineering and Informatics*, *12*(4), 2446–2456. DOI: 10.11591/eei.v12i4.4448

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*, *1*(5). DOI: 10.1007/s42979-020-00305-w

Kadhim, R. R., & Kamil, M. Y. (2023). Comparison of machine learning models for breast cancer diagnosis. *IAES International Journal of Artificial Intelligence*,*12*(1),415421.DOI:10.11591/ijai.v12.i1.pp415-421

Kumar, S., & Singh, M. (2021). Breast Cancer Detection Based on Feature Selection Using Enhanced Grey Wolf Optimizer and Support Vector Machine Algorithms. *Vietnam Journal of Computer Science*,*8*(2),177197.DOI:10.1142/S219688882150007X

Laghmati, S., Hamida, S., Hicham, K., Cherradi, B., & Tmiri, A. (2024). An improved breast cancer disease prediction system using ML and PCA. *Multimedia Tools and Applications*, *83*(11), 33785–33821. DOI: 10.1007/s11042-023-16874-w

Lappeenranta-. (2023). *BREAST CANCER DIAGNOSTIC USING MACHINE LEARNING Applying Supervised Learning Techniques to Coimbra and Wisconsin Datasets*.

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 243–252. DOI: 10.1145/1835804.1835837

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386. DOI: 10.21275/art20203995

Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. *Communications in Computer and Information Science*, *1234 CCIS*, 108–117. DOI: 10.1007/978-981-15-7205-0_10

Naji, M. A., Filali, S. El, Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. *Procedia Computer Science*,*191*,487492.DOI:10.1016/j.procs.2021.07.062

Nemade, V., & Fegade, V. (2022). Machine Learning Techniques for Breast Cancer Prediction. *Procedia Computer Science*, *218*, 1314–1320. DOI: 10.1016/j.procs.2023.01.110

Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Informatics in Medicine Unlocked*, *26*, 100696. DOI: 10.1016/j.imu.2021.100696

Sakib, S., Yasmin, N., Tanzeem, A. K., Shorna, F., Md. Hasib, K., & Alam, S. B. (2022). Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms. *Lecture Notes in Electrical Engineering*, *844*, 703–717. DOI: 10.1007/978-981-16-8862-1_46

Shiny Irene, D., Sethukarasi, T., & Vadivelan, N. (2020). Heart disease prediction using hybrid fuzzy K-medoids attribute weighting method with DBN-KELM based regression model. *Medical Hypotheses*,*143*(March),110072.DOI:10.1016/j.mehy.2020.110072

Uddin, K. M. M., Biswas, N., Rikta, S. T., & Dey, S. K. (2023). Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. *Computer Methods and Programs in Biomedicine Update*,*3*.DOI:10.1016/j.cmpbup.2023.100098

Vergara, J. R., & Estévez, P. A. (2015). *A Review of Feature Selection Methods Based on Mutual Information*. DOI: 10.1007/s00521-013-1368-0

Wolberg, W. M. S. and S. (1995). *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository. DOI: 10.24432/C5DW2B

Yadav, R. K., Singh, P., & Kashtriya, P. (2022). Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey. *Procedia Computer Science*,*218*,14341443.DOI:10.1016/j.procs.2023.01.122

Zhou, S., Hu, C., Wei, S., & Yan, X. (2024). Breast Cancer Prediction Based on Multiple Machine Learning Algorithms. *Technology in Cancer Research and Treatment*, *23*. DOI: 10.1177/15330338241234791