The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES
20(2): S2: 800-809, 2025 www.thebioscan.com

HOW DIFFERENT FEATURE SELECTION METHODS AFFECT THE PRECISION OF BREAST CANCER PREDICTION MACHINE LEARNING MODELS: A COMPARATIVE STUDY.

Ms. Khamrunissa Hussain[1,2], Prof. Narendra Kumar[3], Dr. Sudheer Kumar Sharma[4]

[1]Research Scholar, NIMS University, India

[2]Lecturer, Ibn Sina National College for Medical Studies, Jeddah, Saudi Arabia

[3]Professor, Department of Applied Sciences, NIMS University, Jaipur, India

[4]Department of Mathematics, NIMS University, Jaipur, India

DOI: 10.63001/tbs.2025.v20.i02.S2.pp800-809

KEYWORDS
Breast Cancer Diagnostic Dataset, Feature Selection, Machine Learning, Breast Cancer.
Received on:

12-04-2025

Accepted on:

15-05-2025

Published on:

21-06-2025

ABSTRACT

Breast cancer is common in developing nations, the early identification of breast cancer is critical for successful treatment. When combined with standard diagnostic data, machine learning techniques can be used to evaluate the risk of acquiring breast cancer. While cancer datasets contain a wealth of patient information, not all data points are useful for predicting cancer outcomes, underscoring the importance of feature selection methods in identifying relevant data.

Numerous studies in this domain have sought to predict various types of breast tumors, as accurate diagnosis is essential for effective breast cancer treatment. The aim of this research is to compare how various feature selection techniques affect the accuracy of different machine learning algorithms currently in use. K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression (LR), Neural Networks (NN), Random Forest (RF), and Naive Bayes (NB) are the seven machine learning methods being assessed in this study. Mutual Information (MI), Spearman Correlation Coefficient, and F-test Feature Selection are among the feature selection methods examined.

The dataset by Wisconsin Diagnostic Breast Cancer (WDBC) is made accessible to the public via the UCI Repository, is used in the studies. According to the results, both the Logistic Regression and Neural Network algorithms outperform other models in terms of accuracy and performance across a wide range of metrics when feature selection is used.

INTRODUCTION

Cancer is among the most lethal illnesses globally. The newest data on breast cancer cases were released in 2023 (Zhou et al., 2024), counting eleven different types of cancer, notably breast cancer in women. About 31% of females of the world are suffering from breast cancer, which makes it the most prevalent kind of malignancy. It ranks sixth globally in terms of overall cancer mortality and is thought to be the main reason of cancer-related deaths amid womenfolk. It killed 685,000 people in 2020, and the World Health Organization (WHO) estimates that it will kill over 963,000 people in 2021, surpassing lung cancer with about 2.3 million new cases (Bray et al., 2024). The incidence of these types of instances of cancer was 25%, whereas the mortality rate among women globally was 17% (Zhou et al., 2024). The atypical proliferation of breast cells is termed a tumor, which is classified into two categories: malignant and benign. The former is malignant, whereas the latter is benign. Although the etiology of carcinoma of the breast in women remains poorly understood, various factors have been identified as potential contributors to the disease, including familial history, intrauterine environmental issues, adolescent exposures, complications during pregnancy, genetic mutations, alcohol and tobacco use, and advanced maternal age, particularly in developing nations (Uddin et al., 2023).

Therefore, it's critical to regularly contact health professionals for early identification, treatment plan, and a precise clinical evaluation in order to reduce the occurrence of breast cancer and prevent deaths among women. However, incorrect diagnosis may occur, reducing the chances of early recovery, or there may be a scarcity of health specialists. Furthermore, the medical examination of the tumor is time-consuming and expensive. As a result, employing techniques like Machine Learning (ML) to automatically identify breast cancer is critical. Various ML classification algorithms, including Naïve Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM), have been utilized to identify malignant breast tumors (Lappeenranta-, 2023) (Ak, 2020).

As an additional point of interest, numerous researchers have utilized a variety of machine learning classification algorithms in order to forecast breast cancer, highlighting the significance of

such methods and illustrating the challenges that are experienced in this field. (Ak, 2020; Mohammed et al., 2020; Nemade et al., 2022; Abunasser et al., 2023; and Ebrahim et al., 2023). Others, such as Chen et al. (2023), Botlagunta et al. (2023), and Laghmati et al. (2024), scrutinized other cancer databases, like the Wisconsin Original Breast Cancer datasets (WDBC) for this purpose and have attained substantial outcomes (Wolberg, 1995).

The purpose of this work is to study the impact that employing the Spearman correlation coefficient, F-test, and Mutual Information (MI) feature selection procedures has on the accuracy of the machine learning algorithm that has currently been selected. This may be performed by comparing the outcomes of both implementations, as well as the results gained by applying the feature selection techniques in tandem. For this experiment, seven distinct machine learning techniques were used: Decision Trees (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Neural Networks (NN), Support Vector Machines (SVM), and Random Forest (RF).

2. Related Works

ML algorithms for breast cancer data categorization have been proposed recently due to medical breakthroughs. Breast cancer data is often utilized for this. Table 1 presents papers on breast cancer prediction, especially the Wisconsin Diagnostic Breast Cancer (WDBC).

Rahmanul Hoque et al. [2024] demonstrate the effectiveness of XGBoost, achieving 94.74% accuracy and 95.24% recall on the Wisconsin diagnostic breast cancer dataset. This suggests XGBoost's potential as a strong candidate for breast cancer prediction, although its performance on Kaggle data, which may differ in characteristics from the Wisconsin dataset, remains unknown. The Wisconsin dataset, while widely used as a benchmark, might not fully represent the variability and complexities found in real-world Kaggle datasets.

Sakib et al. (2022) employed ML and Deep Learning to predict and diagnose WDBC breast cancer. They evaluated categorization models using several measures. Accuracy, recall, specificity, precision, FNR, FPR, F1-score, and MCC were measured. Based on accuracy, the RF classifier performed best at 96.66%. Chen et al. (2023) examined K-Nearest Neighbor (KNN), random forest, XGBoost, and logistic regression, for breast cancer classification, stressing need for early detection. Using WDBC from the UCI library, they selected features using Z-score normalization and Pearson correlation and handled data imbalance via hierarchical sampling. In 80:20 and 70:30 splits, the XGBoost model surpassed others with 100% recall, 96.0% precision, 97.4% accuracy, and 98.0% F1-score. The study found split-specific performance variability and diagnostics' general machine learning technique lacking. Various approaches and preprocessing processes are utilized in the literature to compare and optimize model performance. Their accuracy and other metrics are good. This study compares machine learning algorithms and evaluates feature selection approaches on model performance in multiple experimental contexts, making its conclusions different from others.

Guido et al. [2024] provides a broader perspective on Support Vector Machines (SVMs) in healthcare applications, including breast cancer prediction. While they don't present specific performance metrics on a particular dataset, they highlight the importance of hyperparameter selection and data quality for optimal SVM performance. This underscores the critical need for careful parameter tuning and data preprocessing when applying SVMs to Kaggle datasets, which may exhibit different characteristics than the benchmark datasets discussed in the review. The review's discussion of hybrid methods combining SVMs with optimization techniques suggests that further improvements in accuracy might be possible through such combinations. However, the effectiveness of these hybrid approaches would need to be assessed specifically on Kaggle data.

Ak, M. F. (2020) used CITY for data visualization and WDBC breast cancer samples to compare the performance of LR, KNN, SVM, and DT. The analysis found that LR had the greatest classification accuracy at 98.1%. The Original WBCD were used to test the execution of three ML algorithms, NB, SMO and DT, (Mohammed et al., 2020). Pre-processing measures like discretization and deleting missing data improved ML performance in the research. SMO beat the other two classifiers with 99.56% accuracy on the WBC dataset.

Study by Mahesh et al. [2024] highlights the success of a Convolutional Neural Network (CNN) model, achieving 95.2% accuracy in distinguishing between cancerous and non-cancerous tissues. Their methodology emphasizes the importance of diverse datasets and robust training strategies, suggesting that CNNs, particularly with optimizations like Early Stopping and Reduce LR On Plateau, could be highly effective. The adaptability mentioned in the study hints at the potential for good performance on various datasets, including those found on Kaggle, but this remains to be empirically validated with direct testing.

Chaurasia et al. (2020) incorporated Mode in order to get rid of typical features from the WDBC, and then employed an ensemble technique with stacking classifiers in order to classify entries with all characteristics in comparison to the limited data subset in order to enhance accuracy. Their strategy improved breast cancer accuracy to almost 90%. Islam et al. (2020) used the Wisconsin Breast Cancer dataset to assess five ML algorithms for breast cancer diagnosis: SVM, KNN, RF, ANN, and LR. Their investigation found that ANN had the highest accuracy, 98.57%. In contrast, Naji et al. (2021) used ML systems to predict cancer in the WDBC dataset and the SVM had the highest accuracy, 97.2%. The scientists showed that ML algorithms can predict breast cancer, although they acknowledged limits and planned to use larger datasets for better accuracy and ethical issues.

In addition, Ara et al. (2021) examined ML algorithms for breast tumor classification as malignant or not. To increase model performance, the study employed training and testing procedures to minimize characteristics to those most linked with the aim. With 96.5% accuracy, RF and SVM models outperformed the other ML approaches.

Table 1: Relevant studies referred in this Study

Reference/ Authors	ML Algorithms	Feature Selection Methods	Dataset	Methodology Used	Results
Rahmanul Hoque et al. [2024]	XGBoost, CNNs, and SVMs	Data with all the features	WDBC	A comparative study	achieving 94.74% accuracy and 95.24%
Sakib et al. (2022)	LR, SVM, RF, KNN, DT,and a DL in order to categorize utilizing cross-validation algorithm	None	WDBC	A comparative study	Accuracy: 96.66% for RF
Guido et al. [2024]	Support Vector Machines (SVMs)	Hyperparameter selection and data quality for optimal SVM performance	Kaggle datasets	Parameter tuning and data preprocessing	Accuracy: 95.33%
Ak, M. F. (2020)	LR, KNN, SVM, and DT	Characteristics used after constructing three datasets: • Feature-complete data, • Feature-heavily correlated	WDBC	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

		data, and • Feature-poor data			dataset 3
Mahesh et al. [2024]	Convolutional Neural Network (CNN) model	Robust training strategies like early Stopping and Reduce LR On Plateau	Kaggle datasets	A comparative analysis	Accuracy over 95.2%
Chaurasia et al. (2020)	Combination method: Gradient Boosting Classifier, RF, AdaBoost, Extra Tree (ET)	A statistical feature selection method called "Mode" was used to narrow down the dataset from 32 features to just 12.	WDBC	Mode to decrease the dataset characteristics	Accuracy over 90%
Naji et al. (2021)	SVM, RF, LR, DT and KNN	Attribute extraction technique with no details	WDBC	A comparative study	Accuracy: 97.2% for SVM
Chen et al., (2023)	LR, KNN, RF, and XGBoost	Z-score is used for standardization and Pearson correlation is used for feature selection	WDBC	Data preparation, feature selection, classification, and prediction	Accuracy of 97.4%, Recall of 100%, precision of 96.0%, and F1-score of 98.0%.

3. METHODS AND APPLICATIONS

Dataset Description

From fine needle aspiration (FNA) biopsy images of 357 benign and 212 carcinogenic breast cancers, the WDBC dataset includes 569 binary classification samples. The dataset contains thus 37.3% cancerous and 62.7% non-cancerous breast lesions. The collection comprises 30 characteristics for cell nuclei shape measurements such diameter, roughness, and smoothness. Kumar et al. (2021) describe the dataset's properties in Table 2. However, WDBC dataset classes are uneven and must be resampled. Figure 1 shows the dataset resampled using Synthetic Minority Over-sampling (SMOTE). The graphic illustrates the 50% attribute distribution to the goal prior to and following resampling for each class. To reduce noise and model error, attributes with a poor association to the goal are removed, leaving each record with a patient ID, being diagnosed, and 23 real-valued attributes. Achieving a standard deviation and zero mean for the attributes is done using the 'Standard Scaler' feature scaling process. This ensures that all characteristics in the new dataset contribute equally to the models' prediction.

The Decision Tree (DT) method (Mohammed et al., 2020) is a supervised machine learning technique commonly employed for categorization and regression. This approach relies heavily on the input node a lot.. The DT has a root node at the top, an inner node representing the input characteristics, and an outer node indicating the decision node or dataset class at the bottom. The DT has many nodes at various levels, and subtrees are miniature trees that may be retrieved from the main tree. Data categorization becomes harder as the tree grows owing to overfitting and data splitting. These issues may be solved utilizing pruning, cross-validation, and ensemble methods to combine numerous trees.

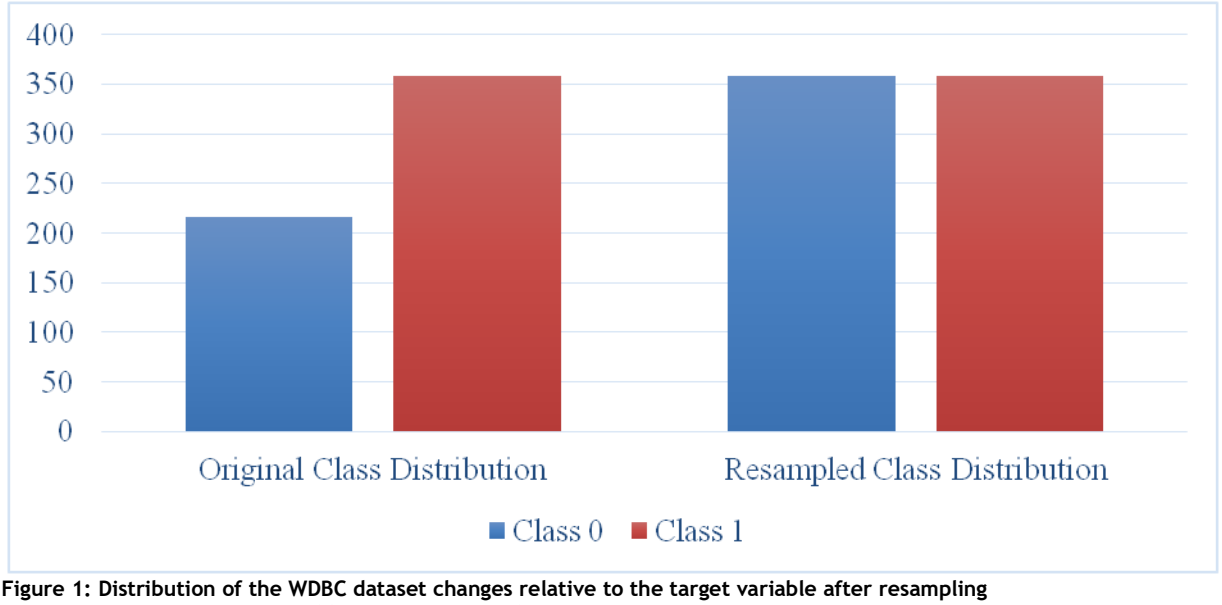

Figure 1: Distribution of the WDBC dataset changes relative to the target variable after resampling

Table 2: Overview of the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset

Attribute	Mean	Standard Deviation	Maximum	Attribute Description
Radius	6.909-28.11	0.121-2.923	7.95-37.10	Determined by averaging the distances from the center to the points along the perimeter.
Texture	9.80-40.00	0.37-4.90	112.10-50.01	Calculated as the standard deviation of grayscale values.
Perimeter	44.02-189.09	0.80-22.01	50.48-252.03	The total distance between successive points along a contour or outline.

Area	144.04-2503.01	6.90-543.10	186.01-4255.00	The total number of concave points in an outline.
Smoothness	0.054-0.164	0.003-0.035	0.072-1.102	Measured as the local variation in radius lengths.
Compactness	0.020-0.350	0.002-0.138	0.030-1.060	Is the ratio of the square of the perimeter to the area, minus 1.
Concavity	0.001-0.501	0.000-0.400	0.000-1.255	The intensity of concave sections of the contour.
Concave Points	0.0001-0.202	0.000-0.055	0.000-1.296	The total number of concave sections of the contour.
Fractal dimension	0.051-0.098	0.001-0.031	0.057-0.210	Coastline approximation minus 1.

Logistic regression (LR): ML approach Logistic Regression (LR) (Dhanya R, 2019, Ak, 2020; and Hossin et al., 2023) foresees two values, 0 or 1, and classifies linear combination data. Linear regression and logistic regression produce parameter coefficients. Logistic regression can be done using gradient descent. To lessen bias and overfit, the LR approach employs regularization and cross-validation. Classification problems are easily and successfully resolved using LR.
Naïve Bayes (NB) is a powerful observed categorization method that can classify massive, complicated data with minimal training data (Kadhim et al., 2023; Hossin et al., 2023; Dhanya R, 2019). The Bayes theorem presupposes conditional independence between two attributes and a class, making it simple. Calculating probability theory is easier using NB. It can help reduce data noise and overfitting due to high independence assumptions.
Random Forest (RF): Random Forest provides classification and regression via ensemble learning. Instead of building a single tree during preprocessing, a 'Random Forest' makes decision trees from subsets of training data randomly (Dhanya R, 2019; Hossin et al., 2023; Kadhim et al., 2023). The formed group minimizes data noise, decreases overfitting, enhances model performance and generalization, and increases accuracy. Therefore, RF is a top choice for many ML applications.
K-Nearest Neighbour (KNN) is a supervised learning technique that, like others, uses labeled training data to predict outcomes on unlabeled data. To get a new input's label or value, the algorithm finds the K data points that are geographically nearest to it (its neighbors) and uses their labels or values to make a prediction. (Ak, 2020; Hossin et al., 2023). When making predictions in a dataset using consensus voting for categorization or aggregating for regression, the numerical value "k" represents the data points that are closest together. A model's performance level is also affected by the actual amount of the "k." Overfitting happens when the value is too low because every signal in the information is caught, whereas generalization and accurate forecasting happen when the value is too high. Distance metrics, such as the one used to calculate the difference between each pair of data points in a dataset (Euclidean distance), form the basis of KNN predictions. This helps to limit noise and the likelihood of overfitting. KNN is a straightforward and adaptable technique that needs precise tuning to get optimal model performance. An illustration of the KNN approach is presented in Figure 2 (Ak, 2020). Figure 2: A KNN Classifier Example (Ak, 2020).

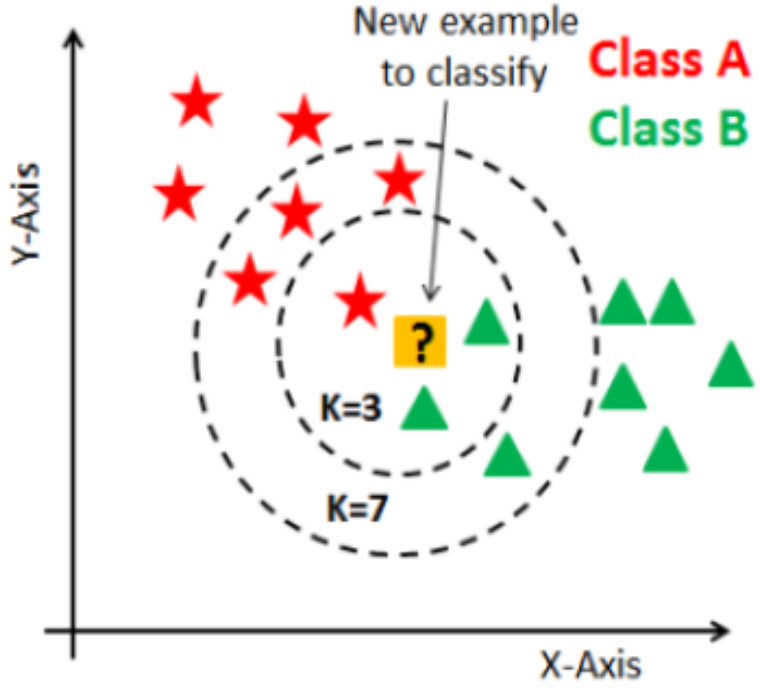

Figure 2: KNN Classifier (Ak, 2020).

Neural Networks (NN): Another supervised machine learning method for categorization issues is neural networks (Mahesh, 2020). Figure 3 illustrates the basic edifice of NN, which is composed of an input layer, one hidden layer, and an output layer (Yadav et al., 2022). The NN layers, which are made up of artificial neurons that collaborate to solve problems akin to those in the human brain and are employed to identify patterns in data, are connected to one another. Neurons in the NN layers process and assess data, and then their output is sent to the hidden layer, which further evaluates it before passing it to the output layer. By computing mistakes in the estimate process, a function known as a loss function is utilized to assess the NN's performance. The more successful the NN prediction, the smaller the loss function value. A number of strategies, including regularization, dropout, and early termination, can be employed to prevent noise and overfitting in data. Figure 3: An Artificial Neural Network's Structure (Yadav et al., 2022).

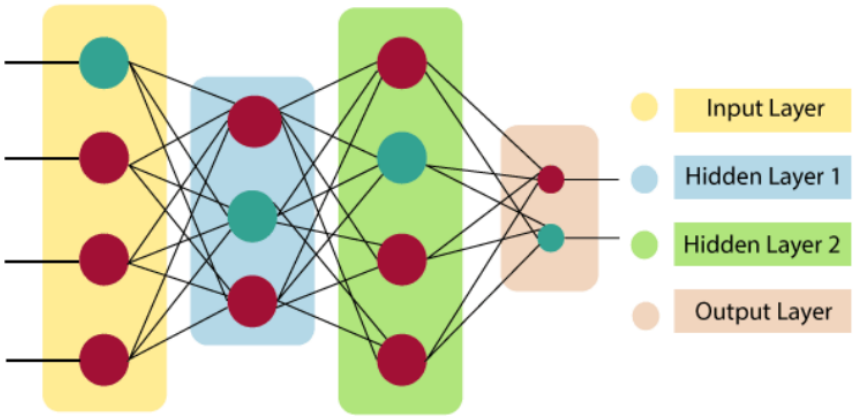

Figure 3: Artificial Neural Network Structure (three layers)

Feature Selection Methods

F-test: To determine how significantly different two or more sets of data are from one another, statisticians utilize the F-test, a feature selection tool. (Dhal et al., 2022). It is well recognized that the F-test works well when dealing with high dimensionality data, like medical datasets. Thus, in a dataset such as breast cancer, this approach is helpful for choosing parameters like tumor characteristics to ascertain whether the patient has cancer or not. Additionally, the models' speed is improved and the overfitting of the classes is decreased by using the F-test to choose just pertinent attributes. To compute F-test, we divide the square of standard deviation of the first set of samples by the square of standard deviation of the second set of samples, where s12 and s22 are the respective square of standard deviation.

Another statistical feature selection technique for identifying linear and non-linear connections in complex datasets, including medical datasets, is mutual information (MI) (Dhal et al., 2022). As a result, it is a helpful method in many domains, including healthcare for diagnosis and treatment and machine learning for modeling. While 0 MI indicates no dependency is present, one feature in MI offers useful information about another feature, i.e., it gauges how reliant each of the two variables is. A reliance between the two attributes is shown by a value greater than zero (Vergara et al., 2015). MI offers the features that are most relevant to the dataset's goal, which improves the performance of the ML model. This procedure is crucial for elucidating the relationship between the largely confusing dataset variables in medical datasets such as breast cancer.

The strength of two-variable correlations is characterized by Spearman Correlation Coefficients, which can have weak or strong values and be either positive or negative (Dhal et al., 2022). According to statistics, it is a non-distribution rank measure, meaning that it assesses the correlation between variables independently of the data's distribution (Hauke et al., 2011). In medical datasets, where the data distribution's frequency is ignored, the Spearman correlation coefficients come in handy due to this property. In other words, it's a breeze to figure out how the input and output variables are related. Accordingly, the Spearman correlation coefficients approach is a useful tool for medical researchers to determine patient diagnosis and treatment.

Evaluation Metrics

Models' performance, strengths, and weaknesses are essential to be evaluated. Consequently, this study assesses the chosen categorization models by utilizing various assessment metrics that are frequently found in the literature. The metrics used are:

- Accuracy (Patro et al., 2021) , mathematically shown as follows:
  $Accuracy$ = $((TP+TN)/(TP+TN+FP+FN))\times 100$ (%) ----------(1)
- F1 score (Lichtenwalter et al., 2010), mathematically shown as follows:
  $F1\ score = 2 * (precision * recall)/(precision + recall)$ ---------- (2)
  where: $precision = true\ positives/(true\ positives + false\ positives)$ --------(3)
  $recall = true\ positives/(true\ positives + false\ negatives)$ ----------(4)
- Matthew's correlation coefficient (MCC) (Ali et al., 2021),
  $MCC=((TP\times TN-FP\times FN)/\int((TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN))$ ----------(5)
- Another statistic used to evaluate models is the area under the curve (AUC) of receiver operating characteristics (ROC) (Shiny Irene et al., 2020). In order to measure how well classification models work, it uses threshold values between 0 and 1, where 1 represents very poor predictions and 0 represents very good ones.

The proportion of participants that the categorization model properly predicted to the absolute number of subjects evaluated is known as accuracy. Often employed in binary classification tasks, the F1 score (Eq. 2) incorporates precision and recall (see Eqs. 3 and 4). The Matthews correlation coefficient (MCC) is a central statistical measure for evaluating the accurateness of binary classification. It provides an excellent score only if the predictions are accurate across all four components of the confusion matrix: true negative (TN), true positive (TP), false positive (FP), and false negative (FN). Consequently, it is often regarded as a balanced metric suitable for use even when class sizes vary significantly. The MCC value ranges between -1 and 1, where -1 indicates that all predictions were incorrect or correct, and 0 signifies that the model's predictions are no better than random chance.

4. EXPERIMENTAL RESULTS

The WDBC dataset developed in above section is used for two types of experiments in this phase of the study. The specified ML algorithms are applied to the dataset without feature selection

techniques F-test, MI, and Spearman Correlation or data resampling in the first trial. The second experiment implements ML models utilizing three feature selection approaches and SOMTE resampling. Model performance is assessed using accuracy, F1 score, precision, recall, ROC AUC, and MCC.

This work uses a computer system with Python libraries like NumPy, Pandas, and Scikit-Learn are used to run experiments. The dataset is separated into 10 equal-sized portions for each experiment to train and test the models effectively and improve prediction accuracy. The data is splits into 10 equal pieces using Cross-validation. Each model uses default hyperparameters, such as setting k to 5 in KNN, one hidden layer and 5000 iterations in NN, and 1000 iterations in LR.

No-feature selection results

This section presents the selected ML model implementation results on the reduced WDBC dataset. This avoids resampling and feature selection Figure 4 show the categorization results for the study's assessment measures.

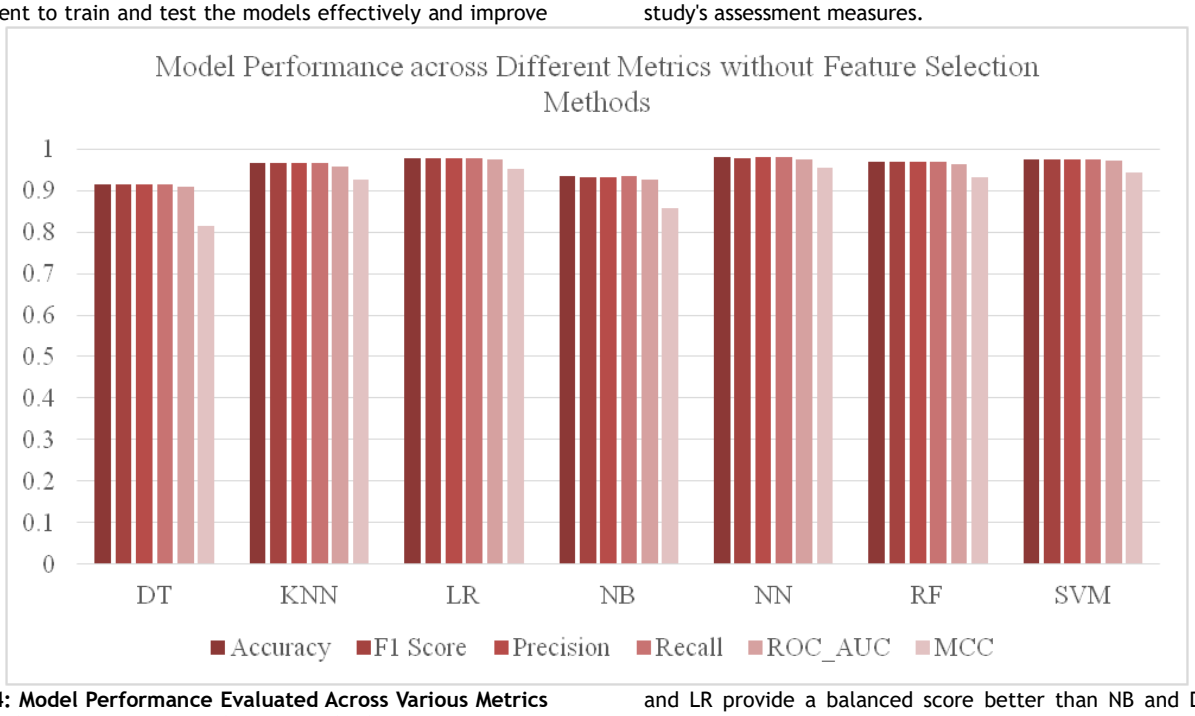

Figure 4: Model Performance Evaluated Across Various Metrics Without the Use of Feature Selection Methods.

Figure 4 reveals that the NN and LR models have the greatest accuracy scores of 0.978910 and 0.977153, whereas the NB model is bad. Both models balance true positive and false positive scores with high F1 values of 0.978869 and 0.977096. NN, LR, and SVM have the highest class label discrimination precision and recall. The maximum ROC AUC value is 0.975530 for NN, compared to the lowest at 0.907424 for DT. The findings show that NN and LR classify the WDBC dataset well. Finally, NN and LR provide a balanced score better than NB and DT, with MCC values of 0.954827 and 0.951067, respectively. NB and DT have the smallest MCC values.

Feature Selection Results

F-test: This section shows the outcomes of integrating F-test feature selection with machine learning. Features are chosen after resampling and scaling the dataset. Figure 5 provide classification accuracy, precision, recall, MCC, and ROC-AUC metrics.

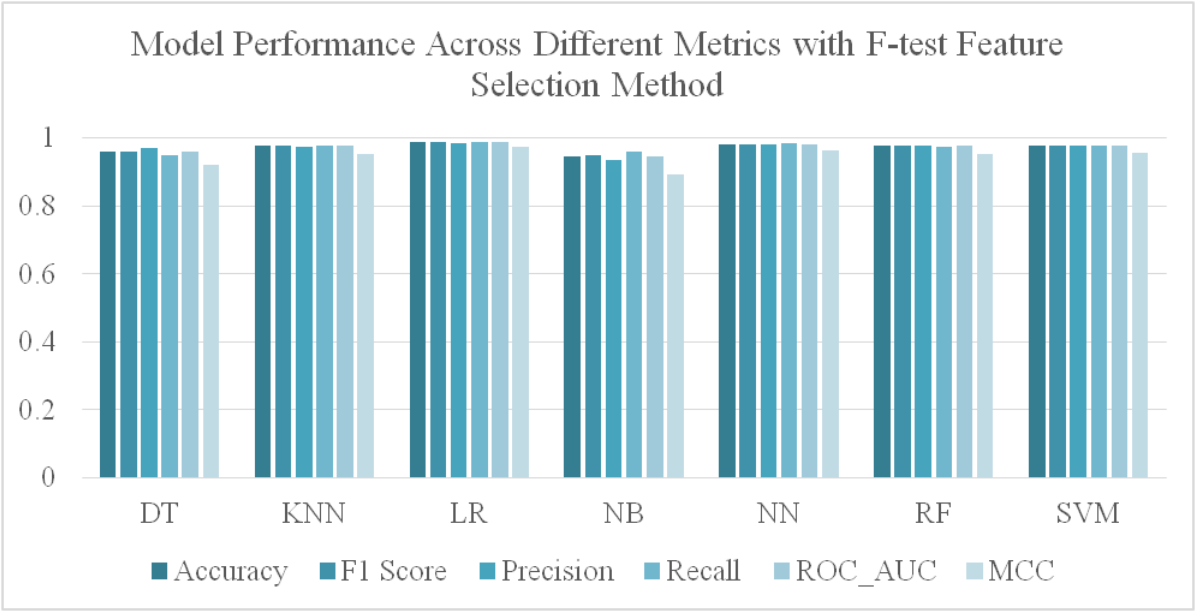

Figure 5: Model Performance Evaluated Across Various Metrics With F-test Feature Selection Methods.

Figure 5 details how successfully different models categorize data using F-test feature selection measures. LR and NN have the highest accuracy, scoring 0.985994 and 0.981793, with the F-test picking 20 and 23 characteristics, respectively. The strong performance of these two models shows that they are dependable for precise class prediction and that the quantity of

selected characteristics is enough to achieve such outstanding performance, therefore not all dataset features are needed. However, the F test technique picked 18, 20, and 20 features for the models of 0.976190, 0.976190, and 0.977591. This also shows that the selected features are good enough to achieve such excellent performance, yet these results are still less accurate than NN and LR. With a recall of 0.988796 and an accuracy of 0.983287, LR successfully balances positive and negative data, while NN also outperforms the other models in identifying true positives and removing false negatives.

The NB model has the lowest accuracy (0.946779) and F1 score (0.947514). This suggests that NB cannot predict cancer in the WDBC dataset. LR has the best F1 score of 0.986034 and NN 0.981818, balancing false positives and negatives. Additionally, LR, NN, KNN, RF, and SVM have acceptable ROC AUC values. The metrics are over 90%, showing class distinction by the models. The final measure, MCC, has the greatest values for LR with 0.972004 and NN with 0.963589, making it dependable for prediction in all confusion matrix categories.

Mutual Information MI, another feature selection approach utilized in this study, is paired with MI methods to demonstrate its effect on model performance. Resampling, scaling, and feature selection using the MI are the prediction steps. Table 3 demonstrates the study's assessment metrics for categorization outcomes. The model's accuracy in classification tasks is assessed by each measure in Table 3. LR is the best model, scoring top across almost all criteria utilizing all dataset features picked using the MI technique. LR balances precision and recall well with an accuracy of 0.983193 and an F1 score of 0.983287, which is essential for classification. LR's accuracy of 0.977839 and recall of 0.988796 demonstrates its ability to make accurate positive predictions and identify many real positive occurrences. The ROC_AUC score of 0.983193 shows its class discrimination, while the MCC of 0.966447 shows a significant correlation between expected and actual outcomes. With an accuracy of 0.964986, KNN is trustworthy but less effective than LR. With 0.978992 accuracy, the NN model maintains good precision (0.975000) and recall (0.983193). RF and SVM models present well, with SVM scoring 0.976190 and RF 0.970588. Despite their competition, these models do not exceed the others in any area, suggesting they may be better for certain situations rather than the best overall. With an accuracy of 0.959384, the DT model performs worse. The NB model performs poorly, scoring lowest in most criteria and with an accuracy of 0.929972. NB might not succeed for this dataset.

Table 3: Evaluation Metrics for WDBC Using the MI Feature Selection Method.

Model	No of Selected Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
DT	23	0.95938	0.95933	0.96067	0.95798	0.95938	0.91877
KNN	23	0.96499	0.96513	0.96111	0.96919	0.96499	0.93001
LR	23	0.98319	0.98329	0.97784	0.9888	0.98319	0.96645
NB	23	0.92997	0.93169	0.90933	0.95518	0.92997	0.86104
NN	23	0.97899	0.97908	0.975	0.98319	0.97899	0.95802
RF	23	0.97059	0.97055	0.97191	0.96919	0.97059	0.94118
SVM	23	0.97619	0.97629	0.97222	0.98039	0.97619	0.95242

Spearman Correlation Coefficient: This section shows Spearman Correlation Coefficient feature selection outcomes using machine learning. Resampling, scaling, and feature selection. Table 4 provide classification accuracy, precision, recall, MCC, and ROC-AUC metrics. This study's assessment measures for model performance are shown in Table 4. Combine the Spearman approach with ML models to see the effect. The table 4 shows that the LR model has the uppermost accuracy and F1 score employing 23 features. The accuracy and F1 score are 0.984594 and 0.984658. LR has competitive accuracy, recall, and MCC scores of 0.980556, 0.988796, and 0.969222. These results show that LR can categorize affirmative instances almost entirely accurately. They also strongly correlate projected and actual results. LR has a ROC AUC of 0.984594, indicating strong target class distinction.

Table 4: Assessment Metrics using Spearman Feature Selection Method

Model	No of Features	Accuracy	F1 Score	Precision	Recall	ROC_AUC	MCC
DT	21	0.95098	0.950495	0.96	0.941176	0.95098	0.902134
KNN	23	0.970588	0.970711	0.966667	0.97479	0.970588	0.94121
LR	23	0.984594	0.984658	0.980556	0.988796	0.984594	0.969222
NB	23	0.938375	0.93956	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

However, the table 4 shows NN and KNN outcomes, which are similar in accuracy and F1 score. KNN has 0.970588 accuracy and 0.970711 F1 score, whereas NN has 0.981793 and 0.981818. The F1 score shows that KNN has a decent balance between accuracy and recall, while NN can give a greater balance between precision and recall of 0.980447 and 0.983193. NN and KNN

perform well, while LR performs better. Accuracy and F1 score are correct for the RF and SVM models. DT's accuracy of 0.950980 and F1 score of 0.950495, the lowest of the models, indicate that it is unsuitable for Spearman feature selection modeling. NB performs poorer than DT and other models, with an accuracy of 0.938375 and lower precision and recall values, resulting in the lowest metrics.

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

Table 5 (LR) F-test technique has an exceptional accuracy of 0.985994 and an F1 score of 0.986034, indicating its robustness. It also offers MI and Spearman-like measures. The great accuracy and recall of LR show its balanced categorization performance. As shown in Table 5 (NN), the NN performs well, especially with F-test accuracy of 0.981793. It can handle complicated patterns with good precision and recall in MI (accuracy of 0.978992) and Spearman (accuracy of 0.981793).

Table 5: Evaluating the performance of the ML models - DT, KNN, LR, NB, NN, RF, and SVM Using F-test, MI, and Spearman Feature Selection Methods

		No of Features	Accuracy	F1 Score	Precision	Recall	ROC_ AUC	MCC
FS method DT	F-test	16	0.959384	0.958982	0.968571	0.94958	0.959384	0.918944
	MI	23	0.959384	0.959327	0.960674	0.957983	0.959384	0.918771
	Spearman	21	0.95098	0.950495	0.96	0.941176	0.95098	0.902134
FS method LR	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
FS method / KNN	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
FS method/ NB	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
FS method/ RF	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
FS method/ NN	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
FS method/ SVM	F-test	20	0.977591	0.977591	0.977591	0.977591	0.977591	0.955182
	MI	23	0.97619	0.97629	0.972222	0.980392	0.97619	0.952415
	Spearman	23	0.978992	0.979021	0.977654	0.980392	0.978992	0.957987

Table 5 shows that KNN outperforms F-test (0.976190 accuracy). It is more variable with MI and Spearman (accuracies of 0.964986 and 0.970588). KNN performs well but falls short of LR and NN. RF method performs well, especially with F-test (0.976190 accuracy). In spite of good metrics, it trails LR and NN but performs like KNN with F-test. Table 5 shows that SVM performs well in F-test and Spearman with accuracy of 0.977591 and 0.978992, respectively. MI has somewhat poorer accuracy—0.976190. SVM's high recall scores prove its accuracy in finding positive cases.

Table 5 shows that the DT is the least effective model. Its F-test accuracy is 0.959384. The model performs poorly despite strong MI and Spearman accuracy scores of 0.959384 and 0.950980, respectively. NB is the least accurate model, scoring below 0.943 across all feature selection techniques. Table 5 shows NB's accuracy and F1 scores decline significantly, showing it is less suitable for this classification job than other models.

Tables 5 show that NN and LR are the best models based on feature selection. SVM, RF, and KNN are dependable but not as good as LR and NN. However, DT and NB struggle to correctly capture data complexity, confirming their lesser fit for this problem area.

CONCLUSION

This study assessed the functioning of KNN, NB, DT, SVM, LR, NN, and RF with diverse feature selection methods, MI, F-test, and Spearman correlation coefficients. The breast cancer dataset WDBC was used to analyze findings using accuracy, F1 score, precision, recall, ROC AUC, and MCC. Not utilizing all dataset attributes for prediction and enhancing modeling yields good model performance. Thus, feature selection approaches were used to choose characteristics that affect model performance and forecast cancer patient status. Tables 5 show the number of features picked using the best feature selection techniques based on assessment metrics.

Results show LR and NN are optimal models. SVM, RF, and KNN are trustworthy alternatives but not as good as LR and NN. However, DT and NB tend to be less accurate in capturing data complexity in this issue area.

To increase breast cancer prediction and healthcare machine learning model efficacy, future research should examine new feature selection approaches, deep learning models, hyperparameter tuning, and different data kinds. Future study must also suggest new hybrid algorithms from ML models based on this paper's results. Integrating several ML models to improve breast cancer prediction in the healthcare industry and using diverse feature selection approaches to extract the best features from breast cancer datasets can achieve this.

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/healthcare 8020111
- 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
- Guido, Rosita, Ferrisi, Stefania, Lofaro, Danilo, and Conforti, Domenico. 2024. "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review". Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/info15040235
- Hoque, Rahmanul, Das, Suman G., Hoque, Mahmudul, and Hoque, Mahmudul. 2024. "Breast Cancer Classification using XGBoost". GSC Online Press. https://doi.org/10.30574/wjarr.2024.21.2.0625
- 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, T R, Thakur, Arastu, Gupta, Muskan, Sinha, Deepak K., Mishra, Kritika Kumari, Venkatesan, Vinoth Kumar, and Guluwadi, Suresh. 2024. "Transformative Breast Cancer Diagnosis using CNNs with Optimized ReduceLROnPlateau and Early Stopping Enhancements". Springer Nature. https://doi.org/10.1007/s44196-023-00397-1
- 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