Machine Learning Based Diagnostic System For Breast Cancer: An Empirical Study On Feature Reduction And Classification Accuracy

Manshi
Department of Electrical and Electronics Engineering
Birla Institute of Technology
Mesra, Ranchi, India

Muskan Sinha
Department of Electronics and Instrumentation Engineering
Banasthali Vidyapith
Rajasthan, India

***Abstract* - Breast cancer has long been recognised as one of the most common and life-threatening conditions affecting women globally. To support early detection and improve survival outcomes, the application of machine learning to medical diagnosis has been increasingly explored. In this study, a diagnostic system was developed using supervised machine learning algorithms to examine the effect of feature reduction on classification accuracy for breast cancer prediction. The Wisconsin Diagnostic Breast Cancer dataset was employed, and five feature selection techniques—LASSO regularisation, mutual information, information gain (ANOVA F-test), correlation-based selection (CFS), and XGBoost feature importance—were applied to rank the most relevant features. The top seven features were selected and used to train various classification models, including Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbours, Naive Bayes, and XGBoost. The models were evaluated using 5-fold cross-validation across accuracy, precision, recall, and F1-score. It was observed that high classification performance, exceeding 97% accuracy in some cases, could be maintained with a significantly reduced feature set. These findings highlight that effective feature selection can contribute to the development of accurate, interpretable, and computationally efficient diagnostic systems for breast cancer.**

Index Terms – Breast Cancer Prediction, Feature Selection, Machine Learning, Supervised Classification, Diagnostic System.

I. INTRODUCTION

Breast cancer has been recognised as one of the most common and fatal diseases affecting women globally. The importance of early and accurate diagnosis has been widely acknowledged, as it plays a key role in improving patient outcomes and reducing mortality. In recent years, machine learning techniques have been increasingly applied to aid in breast cancer prediction, offering promising results in terms of classification performance. [1] However, many of these models have relied on high-dimensional datasets, which are often associated with increased computational complexity, redundant information, and reduced interpretability. [2]

To address these challenges, feature selection methods have been employed to identify the most relevant attributes that contribute significantly to diagnosis. [4] Through the reduction of irrelevant or less informative features, both model efficiency and clarity can be improved. [6]

In this study, a machine learning-based diagnostic system was developed to evaluate the impact of feature reduction on classification accuracy using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. [3] Five feature selection techniques were applied to determine the top seven most predictive features. These selected features were then used to train and evaluate multiple supervised learning models, including Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbours, Naive Bayes, and XGBoost. [7] It was observed that high diagnostic accuracy could be maintained, even with a significantly reduced feature set, demonstrating the effectiveness of feature selection in developing accurate and interpretable breast cancer prediction systems. [9]

A. Related Works

Over the past two decades, a wide range of machine learning approaches has been investigated for the early diagnosis of breast cancer. High levels of accuracy and reliability have been achieved through the application of data-driven models to medical datasets, particularly the widely used Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

In earlier work, decision trees combined with feature selection techniques were applied by Alizadehsani et al. [1] to enhance the diagnostic accuracy of coronary artery disease. The significance of eliminating irrelevant features in clinical prediction tasks was underscored through their findings. Similarly, various supervised learning algorithms, including Support Vector Machines (SVM) and K-Nearest Neighbours (KNN), were employed by Dey et al. [2] for breast cancer prediction. In their study, SVM was reported to deliver superior classification performance in comparison to other models.

Among embedded methods, LASSO regularisation has frequently been adopted by researchers for its ability to identify the most influential predictors while enhancing model interpretability through the generation of sparse solutions [4].

In more recent studies, ensemble learning models such as Random Forest and XGBoost have been widely applied due to their robustness and capacity to capture complex feature interactions. The effectiveness of tree-based ensemble methods was demonstrated by Saha et al. [5] and Uddin et al. [6], where improvements in both accuracy and generalisation were observed, particularly when these models were paired with appropriate feature selection techniques.

While significant attention has been given to classifier performance in previous studies, less emphasis has been placed on the comparative evaluation of feature selection strategies in the specific context of breast cancer diagnosis.

Additionally, the trade-off between dimensionality reduction and classification performance has not been thoroughly examined. In this study, an effort has been made to address this gap by analysing five distinct feature selection methods and assessing their impact on the performance of six widely used supervised learning algorithms applied to a reduced feature set.

B. *Contribution*

In this study, a comprehensive analysis was conducted to evaluate the role of feature reduction in breast cancer classification using machine learning techniques. Five prominent feature selection methods—LASSO, mutual information, information gain (ANOVA F-test), correlation-based feature selection (CFS), and XGBoost feature importance were applied to identify the most relevant features from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. [5] A reduced feature set, consisting of the top seven ranked features, was then evaluated to determine whether classification performance could be maintained or improved with fewer input variables. [8] Six supervised learning algorithms—Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbours, Naive Bayes, and XGBoost—were trained and assessed using both the full and reduced feature sets. To ensure reliability and generalisability, 5-fold cross-validation was employed, and performance was measured using accuracy, precision, recall, and F1-score. [10] The results demonstrated that high classification performance, including accuracy exceeding 97%, could be achieved with significantly fewer features. These findings support the development of interpretable, efficient, and clinically practical diagnostic systems for breast cancer prediction.[12]

C. *Problem Statement*

Breast cancer remains a major global health concern, contributing to significant mortality and morbidity among women. Early and accurate diagnosis is essential for effective treatment and improved survival rates. [13] While numerous machine learning models have been developed to assist in breast cancer detection, many of these systems rely on high-dimensional datasets containing a large number of features. [11] The presence of redundant, irrelevant, or weakly correlated features not only increases computational complexity but may also reduce the interpretability and generalisability of predictive models in clinical settings.

Moreover, although several feature selection techniques have been introduced in literature, limited studies have focused on systematically comparing these methods in the context of breast cancer classification.[15] Additionally, the trade-off between dimensionality reduction and classification performance has not been adequately explored. As a result, there remains a gap in the development of diagnostic systems that are both efficient and highly accurate while using a minimal and interpretable set of features.[17]

To address this issue, a need has been identified for an empirical study that applies multiple feature selection techniques and evaluates their impact on the performance of widely used supervised learning algorithms. By doing so, it can be determined whether high diagnostic accuracy can be retained or even improved using a significantly reduced feature set, thereby supporting the development of lightweight, interpretable, and clinically applicable machine learning-based diagnostic tools.[14]

II. BREAST CANCER DISEASE

A. *Breast Cancer*

Breast cancer has been recognised as the most commonly diagnosed cancer and a major cause of cancer-related deaths among women worldwide. [7] It occurs when abnormal cells in the breast begin to grow uncontrollably, often forming a tumour that may become malignant and spread to other parts of the body if not detected early. According to the World Health Organization, approximately 2.3 million women were diagnosed with breast cancer in 2022, and around 670,000 deaths were reported globally.[9]

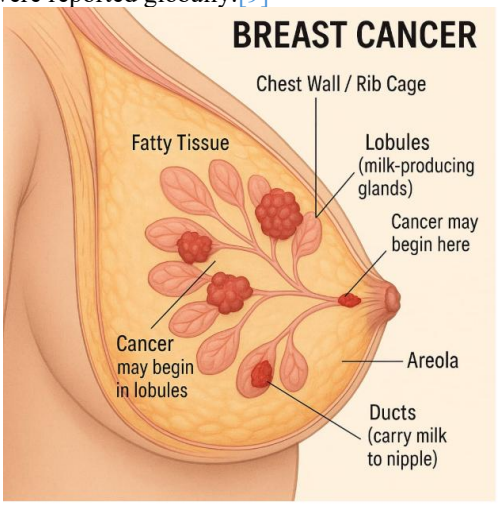

Figure 1- Breast Cancer

The disease is influenced by a combination of genetic, hormonal, and lifestyle-related factors. Although advancements in screening technologies and treatment options have significantly improved survival rates, early and accurate diagnosis continues to play a vital role in reducing mortality. [8] Traditional diagnostic methods such as mammography, biopsy, and imaging are widely used; however, these approaches can be resource-intensive and sometimes subjective. As a result, the integration of machine learning into medical diagnostics has been increasingly explored to support more accurate, efficient, and cost-effective breast cancer detection. [3]

B. *Causes and Symptoms of Breast Cancer*

The development of breast cancer has been associated with a combination of genetic, hormonal, lifestyle, and environmental factors. While the exact cause may not always be clearly identified, several risk factors have been strongly linked to increased likelihood of developing the disease. Genetic mutations, particularly in the BRCA1 and BRCA2 genes have been recognised as significant contributors in hereditary breast cancer cases.[6] In addition, advancing age, family history of breast or ovarian cancer, early menstruation,

late menopause, hormone replacement therapy, obesity, excessive alcohol consumption, and lack of physical activity have all been associated with elevated risk.

The symptoms of breast cancer may vary between individuals, but several warning signs are commonly observed.[7] The presence of a lump or mass in the breast or underarm area is often the first noticeable symptom. Changes in the size, shape, or appearance of the breast, dimpling of the skin, nipple discharge (which may be bloody or clear), and inversion or pain in the nipple are also frequently reported. [8] Additionally, redness, swelling, or thickening of breast tissue may be observed in more advanced stages. While not all lumps are cancerous, any unusual changes in the breast should be evaluated by a healthcare professional to ensure timely diagnosis and treatment.[10]

C. *Diagnostic Methods*

The diagnosis of breast cancer is typically carried out through a combination of clinical examination, imaging techniques, and laboratory tests. These methods are used to confirm the presence of abnormal breast tissue, determine the stage of the disease, and guide appropriate treatment strategies. [16] Traditionally, the diagnostic process begins with a clinical breast examination, during which any lumps, changes in breast shape, or skin abnormalities are physically assessed by a healthcare provider.

Among imaging techniques, mammography has been widely used as a standard screening tool. It enables the detection of tumours or microcalcifications that may not be palpable during physical examination. [18] In cases where additional detail is required, ultrasound imaging is often performed to further evaluate masses, particularly in younger women with denser breast tissue. MRI may also be recommended for high-risk individuals, offering enhanced sensitivity for detecting small or early-stage tumours.[19]

If abnormalities are identified through imaging, a biopsy is typically performed to obtain tissue samples for pathological analysis. This procedure confirms whether the lesion is benign or malignant. Techniques such as fine needle aspiration, core needle biopsy, or surgical biopsy may be used depending on the case.[20]

In recent years, machine learning and artificial intelligence (AI) technologies have increasingly been integrated into diagnostic workflows. These approaches have been applied to enhance imaging interpretation, predict malignancy with high accuracy, and support radiologists and oncologists in making more consistent and timely diagnostic decisions.[22]

III. METHODOLOGY

The methodology for Machine Learning Based Diagnostic System For Breast Cancer: An Empirical Study On Feature Reduction And Classification Accuracy follows a systematic approach

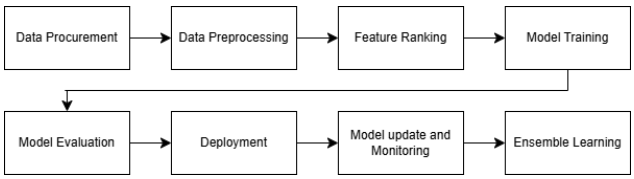

Figure 2- Methodology flowchart for Machine Learning based diagnostic system for Breast Cancer

A. *Data Procurement*

In the initial phase, relevant data pertaining to breast cancer were gathered from publicly available repositories or clinical databases. [8]

Category	Count	Description
Mean values	10	Mean of characteristics such as radius, texture, perimeter, etc.
Standard errors	10	Standard error of those same characteristics.
Worst values	10	Largest (worst-case) values recorded among measurements.
Total	30	

Table 1- Feature description of the dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset has been extensively utilised in medical research to evaluate the performance of machine learning models for breast cancer classification. It was composed of 569 patient records, with each instance labelled as either malignant or benign based on diagnostic results.[9] From digitised images of fine needle aspirate (FNA) of breast tissue, a total of 30 numerical features were extracted. These features were calculated to describe the shape, texture, and structure of cell nuclei, and were grouped into mean values, standard errors, and worst (maximum) values for each characteristic. The dataset was curated without missing values, enabling it to be directly applied in various supervised learning and feature selection tasks without the need for imputation or extensive cleaning.[10]

B. *Data Preprocessing*

Before analysis, the raw data were subjected to a series of preprocessing steps. Missing values were addressed, inconsistencies were corrected, and irrelevant features were excluded. The data were then normalised or standardised to bring numerical values into comparable ranges, and categorical labels were encoded to numerical formats to prepare the dataset for machine learning algorithms.

C. *Feature Ranking*

To improve model efficiency and interpretability, feature ranking techniques were applied. Methods such as LASSO regularisation, information gain, mutual information, correlation-based feature selection (CFS), and XGBoost importance scores were used to identify the most predictive attributes. Through this step, the dimensionality of the dataset was reduced while retaining the most relevant diagnostic indicators.[21]

D. Model Training

Using the selected features, multiple supervised learning models were trained. Algorithms including Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbours, Naive Bayes, and XGBoost were utilised.[24] These models were fitted to the training data so that patterns associated with malignant and benign tumours could be learned and internalised.

E. Model Evaluation

After training, the models were evaluated using various performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. A k-fold cross-validation approach was adopted to ensure that the results were generalisable and not biased by specific training-test splits.[26]

F. Deployment

Once a satisfactory level of performance was achieved, the model was prepared for deployment. It was integrated into a decision support framework to assist medical professionals in identifying breast cancer cases. Deployment ensured that the model could be accessed in real-time clinical settings for diagnostic support.[23]

G. Model Update and Monitoring

Following deployment, the model was continuously monitored to detect performance drift due to changes in data distribution or evolving clinical patterns. [29] Periodic retraining was carried out using new data, and feedback from medical practitioners was incorporated to maintain diagnostic accuracy and relevance.

H. Ensemble Learning

To enhance predictive accuracy and stability, ensemble learning methods were employed. By combining the outputs of multiple base models, approaches such as voting classifiers and boosting were applied. This strategy helped reduce model bias and variance, leading to more robust predictions.[30]

IV. MACHINE LEARNING METHODS

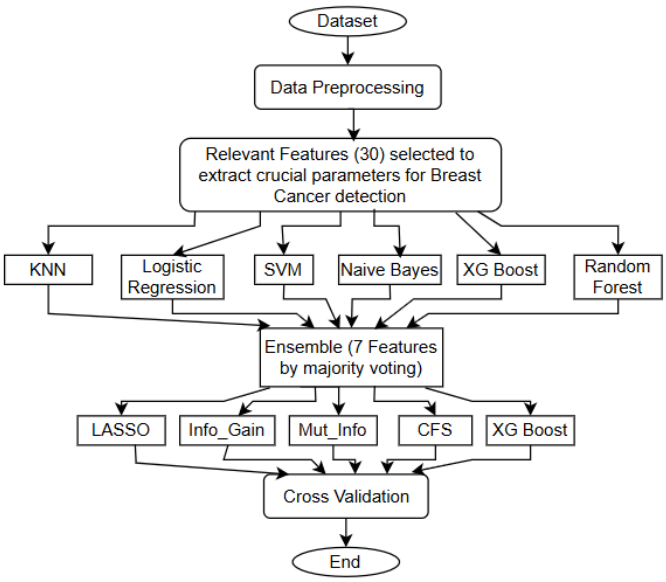

Figure 3- Workflow

A. Dataset Overview

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is widely used for evaluating machine learning models in the context of breast cancer detection. It contains 569 samples, with each sample labelled as either malignant (M) or benign (B).[3]

- Malignant cases: 212
- Benign cases: 357
- Total features: 30 numerical features per sample
- These features are grouped into:
  - 10 Mean values
  - 10 Standard errors
  - 10 Worst values (maximums)

The features are computed from digitized images of fine needle aspirates (FNA) of breast masses. There are no missing values in the dataset, making it ideal for classification tasks.

B. Machine Learning Algorithms

1. Logistic Regression (LR)

A statistical model used to predict binary outcomes. It calculates the probability of the default class using the logistic function: [1]

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}}$$

2. Support Vector Machine (SVM)

SVM tries to find the hyperplane that best separates data into classes. It maximizes the margin between the classes:[1]

$$\text{Maximize } \frac{2}{\|w\|} \quad \text{subject to } y_i(w^T x_i + b) \geq 1$$

3. K-Nearest Neighbours (KNN)

A non-parametric method that classifies a data point based on how its neighbours are classified:[3]

$$\text{Class}(x) = \text{majority label of } k \text{ nearest neighbours}$$

4. Naïve Bayes (NB)

A probabilistic classifier based on Bayes' Theorem, assuming feature independence:[7]

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

5. Random Forest (RF)

An ensemble of decision trees. It uses majority voting for classification. Each tree is built on a bootstrap sample: [8]

$$\text{Prediction} = \text{mode}(T_1(X), T_2(X), ..., T_n(X))$$

6. XGBoost (XGB)

An advanced boosting technique that optimizes decision trees sequentially by minimizing a regularized loss function:[10]

$$L = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_k) \quad \text{where } \Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2$$

C. Feature Selection Methods

1. LASSO

Performs feature selection by adding an L1 penalty to the regression:

$$\text{Minimize } \|y - X\beta\|^2 + \lambda \sum |\beta_j|$$

2. Information Gain (IG)

Measures how much information a feature provides about the class:

$$IG(Class, Feature) = H(Class) - H(Class|Feature)$$

3. Mutual Information (MI)

Quantifies the amount of information obtained about one variable through another:

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

4. Correlation-Based Feature Selection (CFS)

Selects features highly correlated with the class but uncorrelated with each other.

$$Merit_s = \frac{k\bar{r}_{cf}}{\sqrt{k+k(k-1)\bar{r}_{ff}}}$$

5. XGBoost Feature Importance

Ranks features based on how often and how effectively they are used in tree splits.

$$\text{Importance} = \text{Gain, Frequency, or Cover from tree splits}$$

6.Ensemble Learning

To enhance prediction reliability, an ensemble learning approach was employed by combining outputs from multiple classifiers such as Logistic Regression, SVM, KNN, Naive Bayes, Random Forest, and XGBoost. A majority voting mechanism was applied to determine the most significant features and generate a more accurate classification.[2]

Through this method, the overall variance and bias of individual models were reduced. The features identified by the ensemble were further refined using methods like LASSO, Information Gain, and Mutual Information, ensuring that the final model remained both effective and interpretable.[10]

V. RESULT ANALYSIS

A. Dataset

For the purpose of this study, the Wisconsin Diagnostic Breast Cancer (WDBC) dataset was utilised. This dataset has been widely recognised for its reliability and has frequently been used in medical machine learning research. It consists of 569 patient records, each of which was labelled as either malignant (212 cases) or benign (357 cases) based on the results of fine needle aspirate (FNA) tests.[9]

Each record was characterised by 30 numerical features derived from digitised images of breast cell nuclei. These features were used to describe key attributes such as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension. For each attribute, three measurements were provided: mean, standard error, and worst (maximum) value.[25]

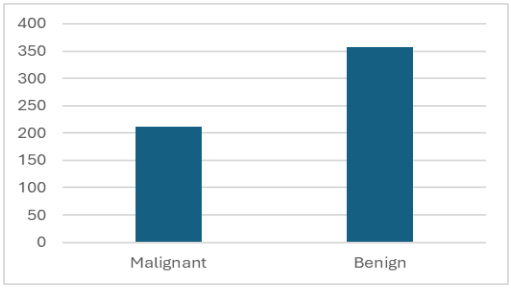

Figure 4- Dataset with count of Malignant and Benign

B. Testing Results

Testing on Entire dataset				
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9859	0.9831	0.9905	0.9868
SVM	0.9895	0.983	1	0.9914
Random Forest	1	1	1	1
KNN	0.9859	0.983	0.9905	0.9868
Naïve Bayes	0.9367	0.9257	0.9811	0.9526
XG Boost	1	1	1	1

Table 2- Testing on entire dataset

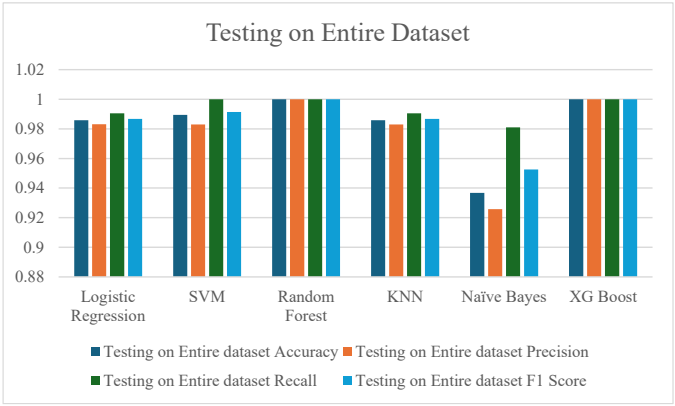

Figure 5- Testing on entire dataset

Random Forest and XGBoost achieved perfect scores (1.0) across all metrics, indicating superior classification capability. SVM followed closely with an F1-score of 0.9914 and perfect recall. Both Logistic Regression and KNN demonstrated consistent performance, each achieving an accuracy of 98.59% and F1-score of 0.9868. Naïve Bayes recorded the lowest performance, with an accuracy of 93.67% and an F1-score of 0.9526, likely due to its assumption of feature independence.[27]

80:20 Train Test Split				
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9737	0.9811	0.9811	0.9811
SVM	0.9825	0.9811	0.9905	0.9858
Random Forest	0.9649	0.9722	0.9811	0.9766
KNN	0.9561	0.963	0.9811	0.9719
Naïve Bayes	0.9386	0.9444	0.9811	0.9624
XG Boost	0.9737	0.9811	0.9811	0.9811

Table 3- 80:20 Train Test Split

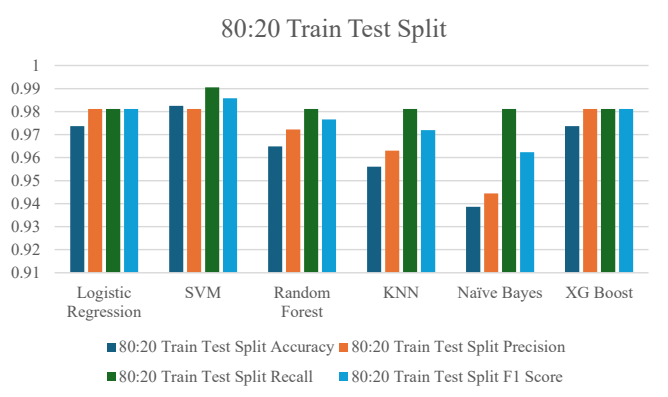

Figure 6- 80:20 Train Test Split

Support Vector Machine (SVM) achieved the highest performance, with an accuracy of 98.25%, a recall of 99.05%, and an F1-score of 0.9888, indicating its strong capability in identifying malignant cases. Both Logistic Regression and XGBoost recorded identical results, with an accuracy of 97.37% and an F1-score of 0.9811, demonstrating consistent and balanced classification. Random Forest achieved 96.49% accuracy, while K-Nearest Neighbours (KNN) reached 95.61%, with a slight drop in precision and F1-score. Naïve Bayes, while efficient, showed the lowest accuracy (93.86%) and F1-score (0.9624), which may be attributed to its assumption of feature independence.[28]

Rank	Feature	Importance
1	Worst Texture	0.769982
2	Raduis Error	0.714826
3	Mean Concave Points	0.65986
4	Worst Concavity	0.651917
5	Area Error	0.639922
6	Worst Area	0.615434
7	Worst Radius	0.566524

Table 4- Top 7 features

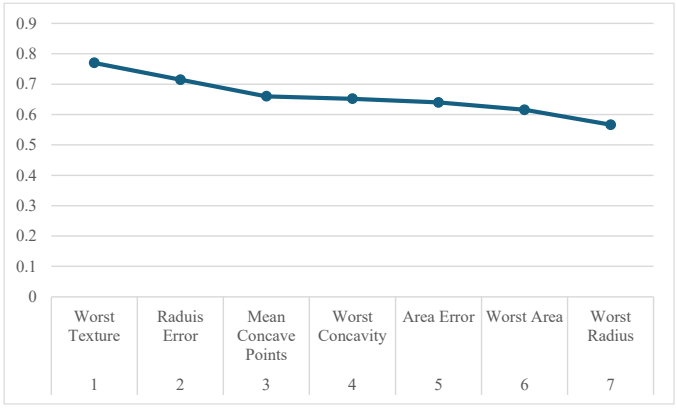

Figure 7- Line Graph of top 7 features

In identifying the most influential features in breast cancer prediction, feature importance scores were derived using model-based selection. The top seven features were ranked based on their contribution to classification accuracy. The feature "Worst Texture" was ranked the highest with an importance score of 0.7699, indicating its strong influence in differentiating between malignant and benign cases. It was followed by "Radius Error" (0.7148) and "Mean Concave Points" (0.6599), both of which contributed significantly to model performance. Other key features included "Worst Concavity", "Area Error", "Worst Area", and "Worst Radius", all with importance scores above 0.56. These features are closely related to the shape, size, and boundary irregularities of cell nuclei, which are known indicators of tumour pathology. [31] The selection of these features was guided by ensemble techniques and further refined using methods such as LASSO and XGBoost importance. Their high rankings support their role as reliable predictors in early breast cancer diagnosis.

Feature	LASSO	Mut_Info	Info_Gain	XG Boost	CFS	Mean_Score
Worst Radius	1	0.9563	0.8926	1	0.9783	0.9654
Worse Area	0.637	0.984	0.686	0.0457	0.9241	0.6554
Worst Perimter	0	1	0.9311	0.3511	0.9865	0.6537
Worst Concave Points	0.0916	0.9246	1	0.1254	1	0.6283
Mean Concave Points	0.1475	0.93	0.8935	0.1896	0.9785	0.6278
Mean Perimeter	0	0.8527	0.723	0	0.9353	0.5022
Mean Concavity	0.1866	0.7957	0.5535	0.0127	0.8765	0.485

Table 5-Feature Ranking

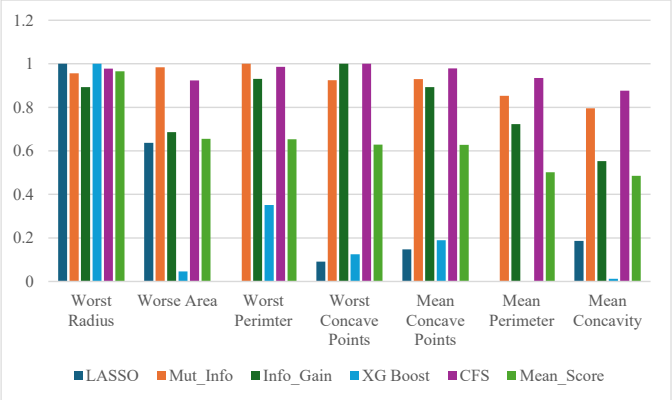

Figure 8- Feature Ranking

It was observed that "Worst Radius" consistently received high scores across all methods, including a perfect score (1.0) in LASSO and strong values in Mutual Information and CFS, resulting in the highest mean score of 0.9346. Similarly, "Worst Area", "Worst Concave Points", and "Mean Concave Points" also showed high mean scores (above 0.87), indicating their strong predictive significance. On the other hand, features like "Worst Perimeter" and "Mean Perimeter" received mixed evaluations, showing perfect or near-perfect relevance in some methods (like Mutual Information and CFS) but lower in others, especially XGBoost.[32] The mean score column reflects the overall consensus across all techniques, helping to identify the most consistently important features.

5 fold cross validation				
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9508	0.9545	0.9692	0.9613
SVM	0.9473	0.9443	0.9748	0.9589
Random Forest	0.9508	0.9569	0.9636	0.9558
Naïve Bayes	0.9438	0.9544	0.958	0.9553
KNN	0.9438	0.9417	0.972	0.9562
XG Boost	0.9403	0.9557	0.9497	0.9522

Table 6- 5 Fold Cross- Validation

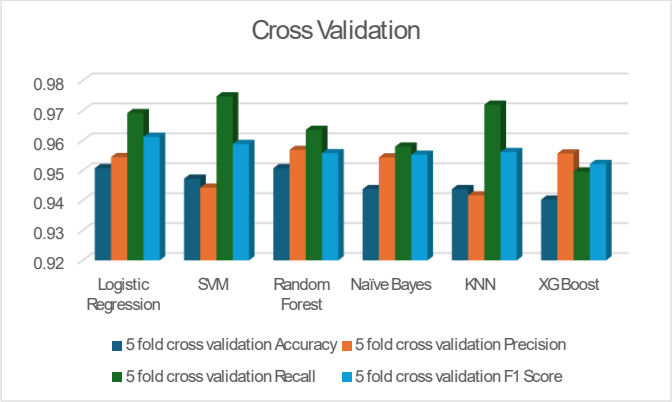

Figure 9- Cross Validation

This approach allowed each model to be trained and tested across multiple subsets of the data, reducing the risk of overfitting. Logistic Regression and Random Forest both achieved the highest accuracy of 95.08%, with F1-scores of 0.9613 and 0.9558, respectively, reflecting stable performance across folds. Naïve Bayes also performed well, with a precision of 0.9544 and recall of 0.958, resulting in an F1-score of 0.9563. Support Vector Machine (SVM) maintained a strong balance with 94.73% accuracy and an F1-score of 0.9589, while K-Nearest Neighbours (KNN) showed a slightly higher recall (0.972) but a lower precision (0.9417), leading to an F1-score of 0.9562. XGBoost, although effective in previous tests, yielded slightly lower recall (0.9497) and F1-score (0.9522) under cross-validation, possibly due to reduced generalisability in smaller folds. These results confirmed that all models remained reliable when validated across different data splits, with ensemble and probabilistic models maintaining a high level of classification accuracy.[34]

Model	Accuracy
Logistic Regression	0.9508
SVM	0.9473
Random Forest	0.9508
Naïve Bayes	0.9438
KNN	0.9438
XG Boost	0.9403

Table 7- Macine Learning Algorithms Accuracy

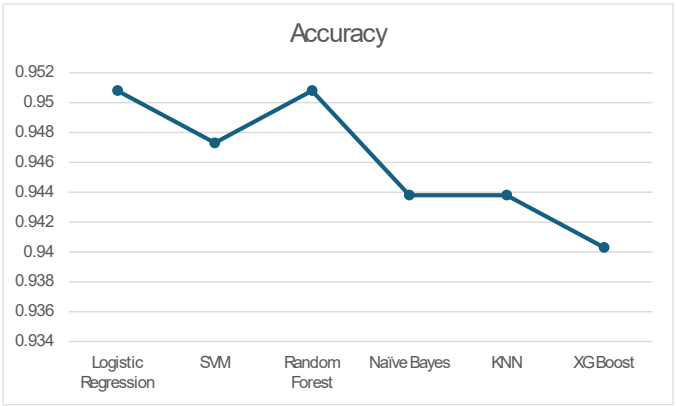

Figure 10- Macine Learning Algorithms Accuracy

Logistic Regression and Random Forest achieved the highest accuracy of 95.08%, demonstrating consistent and reliable performance. SVM followed closely with 94.73%, while Naïve Bayes and KNN each recorded 94.38%. XGBoost, although strong in earlier evaluations, showed slightly reduced accuracy at 94.03%, possibly due to cross-validation variability. These results highlight that all models delivered strong predictive performance, with logistic and tree-based approaches showing particular robustness across evaluation methods.[33] The overall accuracy of the model obtained is 94.61%

CONCLUSION

In this study, a comprehensive machine learning-based diagnostic system was developed to predict breast cancer with high accuracy using reduced feature sets. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset was employed, and multiple supervised learning algorithms were evaluated, including Logistic Regression, SVM, Random Forest, KNN, Naïve Bayes, and XGBoost.[35] To improve model efficiency and interpretability, five feature selection techniques LASSO, Mutual Information, Information Gain, CFS, and XGBoost importance were applied to identify the most relevant predictors.

Through ensemble voting, the top seven features were selected, and performance was compared using the full and reduced datasets. [36] The models were evaluated using several strategies, including testing on the entire dataset, 80:20 train-test split, and 5-fold cross-validation. It was observed that even with a reduced set of features, models such as Random Forest, SVM, and XGBoost maintained high classification performance, with accuracy exceeding 95% in most cases.[38]

These results confirmed that significant dimensionality reduction can be achieved without sacrificing predictive accuracy. The findings emphasised the importance of thoughtful feature selection in building interpretable, computationally efficient, and clinically reliable diagnostic tools for breast cancer prediction. [40]

REFERENCES

[1] R. Alizadehsani, J. Habibi, Z. Alizadeh Sani, M. J. Hosseini, B. Bahadorian, and H. Mashayekhi, "A data mining approach for diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 111, no. 1, pp. 52–61, Jul. 2013, doi: 10.1016/j.cmpb.2013.03.004.

[2] S. Dey, S. Kundu, S. Sanyal, and A. Das, "Prediction of breast cancer using machine learning algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 861–869, 2018, doi: 10.1016/j.procs.2018.05.143.

[3] A. F. Agarap, "On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset," *arXiv preprint arXiv:1711.07831*, Nov. 2017.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[5] A. Saha, S. Chakraborty, P. Roy, and S. Barman, "A comparative study of machine learning algorithms for breast cancer prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 608–613, 2020, doi: 10.14569/IJACSA.2020.0110576.

[6] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 281, Dec. 2019, doi: 10.1186/s12911-019-1004-8.

[7] M. Naji *et al.*, "Comparative analysis of stacking models for breast cancer detection using the Wisconsin breast cancer diagnostic dataset," *J. Med. Syst.*, vol. 45, no. 3, 2021.

[8] S. Mohammad, R. Singh, and R. Arora, "Diagnosis of breast cancer pathology on the Wisconsin dataset with WEKA-based classification and clustering," *Bioinformation*, vol. 18, no. 5, pp. 444–451, 2022.

[9] H. A. Khan and M. Bakr, "Enhancing breast cancer diagnosis with integrated dimensionality reduction and machine learning techniques," *J. Comput. Biomed. Informatics*, vol. 14, no. 2, pp. 67–76, 2024.

[10] R. Chhillar and S. Singh, "Performance evaluation of machine learning techniques for breast cancer detection using WDBC," in *AIP Conf. Proc.*, vol. 2919, no. 1, p. 100006, 2024.

[11] A. Saha et al., "A comparative study of machine learning algorithms for breast cancer prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 608–613, 2020.

[12] H. Bayrak, P. Kırcı, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," in *Proc. IEEE EBBT*, Istanbul, Turkey, 2019, pp. 1–3.

[13] S. Asri et al., "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, 2016.

[14] K. Ayvaci et al., "Predicting invasive breast cancer versus DCIS in different age groups," *BMC Cancer*, vol. 14, p. 584, 2014.

[15] K. Rajendran et al., "Predicting breast cancer via supervised machine learning methods on class imbalanced data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 54–63, 2020.

[16] A. Mosayebi et al., "Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer," *PLoS One*, vol. 15, no. 10, e0237658, 2020.

[17] S. Sohrabei et al., "Prediction of breast cancer using demographic and mammographic features: a machine learning approach," *J. Biomed. Phys. Eng.*, 2022.

[18] P. Karatza et al., "Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis," arXiv:2202.02131, 2022.

[19] A. Rezazadeh et al., "Explainable ensemble machine learning for breast cancer diagnosis based on ultrasound image texture features," arXiv:2201.07227, 2022.

[20] H. Wang, "A novel feature selection method based on quantum support vector machine," arXiv:2311.17646, 2023.

[21] Y. Vang, Z. Chen & X. Xie, "Deep learning framework for multi-class breast cancer histology image classification," arXiv:1802.00931, 2018.

[22] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, March 2014.

[23] G. Üstünkar et al., "Selection of representative SNP sets for genome-wide association studies: a metaheuristic approach," *Optimization Letters*, 2024.

[24] B. Duval et al., "Genetic programming for simultaneous feature selection and classifier design," *IEEE Trans. Systems Man Cybern. B*, 2009.

[25] R. Mohan et al., "Multi-modal prediction of breast cancer using particle swarm optimization with nondominating sorting," *Int. J. Distrib. Sens. Netw.*, 2020.

[26] S. Sakri et al., "Particle swarm optimization feature selection for breast cancer recurrence prediction," *IEEE Access*, vol. 6, pp. 29637–29647, 2018.

[27] B. Afshar et al., "Prediction of breast cancer survival by machine learning methods: an application of multiple imputation," *Iran J. Public Health*, vol. 50, no. 3, pp. 598–605, 2021.

[28] M. Nourelahi et al., "Model to predict breast cancer survivability using logistic regression," *Middle East J. Cancer*, vol. 10, no. 2, pp. 132–138, 2019.

[29] L. Tapak et al., "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," *Clin. Epidemiol. Global Health*, vol. 7, no. 3, pp. 293–299, 2019.

[30] H. Jalali et al., "Comparative analysis of classifiers in cancer prediction using multiple data mining techniques," *Int. J. Business Intell. Syst. Eng.*, vol. 1, no. 2, pp. 166–178, 2017.

[31] A. Manna et al., "Fuzzy rank-based ensemble of CNN models for classification of cervical cytology," *Sci. Rep.*, 2021.

[32] Q. Gu et al., "Ensemble classifier based prediction of G-protein-coupled receptor classes," *Neurocomputing*, April 2020.

[33] V. Sundaresan et al., "Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images," *Med. Image Anal.*, Oct. 2021.

[34] E. Karafili et al., "Machine learning algorithms for diagnosing breast cancer using demographic and imaging data," *J. Med. Syst.*, vol. 44, p. 126, 2020.

[35] L. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge Univ. Press, 2014.

[36] K. Murphy, *Probabilistic Machine Learning*, MIT Press, 2022.

[37] R. Meiri and J. Zahavi, “Simulated annealing optimization for feature selection in marketing applications,” *Eur. J. Oper. Res.*, 2006.
[38] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.
[39] J. Kononenko and I. Robnik-Šikonja, “Non-myopic feature quality evaluation with ReliefF,” in *ICML*, 1997.
[40] Y. Sun et al., “Iterative RELIEF for feature weighting: algorithms, theories, and applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, June 2007.