Vol.no 1, Issue no 2, Sep - Dec, 2025

PUXplore Multidisciplinary Journal of Engineering

Parul® University NAAC GRADE A++

Breast Cancer Prediction using Machine Learning Techniques

Sandip Chakraborty[1], Aritra Das[2], Subhadeep Das[3], Arya Manikya Sinha[4]

[1&4]*Assistant Professor, Artificial Intelligence and Data Science, Parul University, India*

[2]*Senior Analyst, HSBC, India*

[3]*Senior Data Analyst, HSBC, India*

E-mail: [1]sandipchakrabortyisp@gmail.com, [2]aritra_md@yahoo.in, [3]www.subhadeepdas@gmail.com, [4]amsinha.jr@gmail.com

Abstract - Breast cancer is one of the leading causes of death among women worldwide. Early detection is crucial for increasing survival rates. In this study, we evaluate and compare several supervised machine learning algorithms for classifying breast cancer tumors as benign or malignant using the Wisconsin Breast Cancer Dataset. The preprocessing stage involved handling missing values, feature scaling, and skewness removal using mathematical transformations. Feature selection was applied to improve classification performance. Six algorithms—Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine—were implemented in Python (Anaconda, Jupyter Notebook) and evaluated using 10-fold cross-validation. Experimental results demonstrate that Logistic Regression and SVM achieved the highest accuracy (98.24%), surpassing results reported in previous studies. This paper highlights the potential of these techniques for developing robust Computer-Aided Diagnosis systems.

Keywords: **Breast Cancer, Machine Learning, Logistic Regression, SVM, Random Forest, KNN, Naïve Bayes.**

I. INTRODUCTION

Breast cancer is one of the most prevalent and life-threatening diseases affecting women worldwide, accounting for a significant proportion of cancer-related morbidity and mortality [1]. According to global health statistics [2], early and accurate detection of breast cancer greatly improves the chances of successful treatment and survival. Traditional diagnostic methods, while effective, are often time-consuming, resource-intensive, and subject to human interpretation errors. The rapid advancement of Machine Learning (ML) has opened new avenues for the development of Computer-Aided Diagnosis (CAD) systems capable of assisting clinicians in the early detection and classification of breast tumors [3]. In particular, supervised ML algorithms have shown considerable promise in distinguishing between benign and malignant tumors with high accuracy [4]. In this study, we investigate the performance of six supervised learning algorithms—Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Machine (SVM)—using the Wisconsin Breast Cancer Dataset (WBCD). Our methodology includes comprehensive preprocessing steps such as handling missing values, feature scaling, and skewness reduction through mathematical transformations. Feature selection techniques are applied to enhance classification efficiency and accuracy. The proposed models are implemented in Python using the Anaconda environment with Jupyter Notebook, and evaluated through 10-fold cross-validation to ensure robust performance assessment. Experimental findings reveal that Logistic Regression and SVM achieve the highest classification accuracy of 98.24%, outperforming many results reported in the literature. These results underscore the potential of ML-based CAD systems in improving breast cancer diagnostics and supporting clinical decision-making.

II. RELATED WORK

Nallamala et al., 2019 [5] proposed several ML algorithms that are obtainable for forecasting & analysis of breast cancer. Such algorithms are Naïve Bayes, KNN, and SVMs. They used projected Ensemble Voting techniques for finding breast cancer disease. Here they use Wisconsin Breast Cancer (WBC) dataset for BC detection. Initially, on the dataset, they apply logistic algorithms and implement neural network contrivance processes. For this, a Voting Ensemble process is instigated for syndicating these grades and concluding precision. To do these tasks NumPy, Pandas, Matplotlib, and sci-kit-learn all these packages are needed. For classification, they use SVM, Logistic Regression, and KNN. After applying all these methods to 16 features only it gives 98.50% precision.

Mohammed et al., 2020 [6] present three steps for data preprocessing, these are such as discretization, instances resampling, and removing the missing values. After that, 10-fold cross-validation has been applied. Then, three classifiers namely Naïve Bayes, SMO (Sequential Minimal Optimization), and Decision Tree built on the J48 algorithm have been evaluated over the WBC dataset as well as FNA-fine needle aspirate of a breast tumor dataset. The accuracy of the classifiers for the test on the Breast Cancer Database without preprocessing for J48, NB, and SMO are 75.52%, 71.67%, and 69.58% respectively. When resample filters are applied many times on the Breast Cancer Database dataset classifier, gives 98.20%, 76.61%, and 95.32% accuracy for J48, NB, and SMO algorithms respectively. Similarly for the WBC dataset when the experiment is performed without preprocessing data it gives an accuracy of 94.56%, 95.99%, and 96.99% for J48, NB, and SMO respectively. After

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

applying a resample filter many times on the WBC dataset classifier gives 99.24%, 99.12%, and 99.56% accuracy for J48, NB, and SMO algorithms respectively. Results show that using the resample filter in the preprocessing phase, enhances the performance of the classifiers.

Sivapriya et al., 2019 [7] proposed four main classification algorithms, SVM, Random Forest, Logistic Regression, and Naive Bayes are used to predict breast cancer on the Breast Cancer dataset WBC database where 699 instances are there in which 458 are Benign and 241 are Malignant. To predict BC, these four classification algorithms are evaluated on the database. To check the efficiency the models are compared based on accuracy. Here accuracy using Logistic Regression, SVM, NB, and Random Forest are 99.06%, 98.59%, 94.83, and 99.76% respectively. From the above values, it is clearly shown that RF gives us the highest accuracy and running time combined because it creates multiple trees with each tree providing different results.

Joshi & Mehta, 2017 [8] proposed RF, SVM, DT, and KNN models for the classification of BC. These four ML techniques generate a huge percentage of differentiating the Benign and Malignant tumors. To check the performance of ML, some measures are used namely accuracy, sensitivity, precision, etc. The Wisconsin Breast Cancer (WBC) dataset is randomly divided into two parts, one for training which contains almost 70% of data, and another part for testing which contains almost 30% of data. Benign is considered for positive and Malignant is considered for negative cases. Based on the confusion matrix values KNN gives 100% accuracy. And SVM gives 94.714% accuracy. Similarly, DT gives 92.891% accuracy. And finally, for RF the values of accuracy are 92.891%. From this result, it can be concluded that the KNN classifier is better than the other classifiers.

Keleş, 2019 [9] proposed an RF algorithm to classify the BC. Here he describes a method for creating a forest of unrelated trees using a Classification and Regression Treelike (CART-like) procedure that is bagging. During the classification phase, RF finds the classification value using more than one decision tree. After that, he uses a 10-fold cross-validation technique for measuring the performances. Where one part is aside and the rest of the part that is 9 folds are used for this purpose. It gives 92.2% accuracy on average. Here one key is important, that is the accuracy of the classification methods depends on the user's parameters such as N (total number of trees), and m (total number of used parameters). So that the choice of parameters may increase the accuracy.

According to Amrane et al., 2018 [10] to make a good prognostic, breast cancer classification needs nine characteristics which are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitosis. The NB and KNN these two methods are proposed to predict the BC in the Wisconsin breast cancer database dataset. After that k-fold cross-validation technique is used to check and evaluate the learning algorithms or models, by partitioning data into a learning set to train the model and a testing set to evaluate it. When KNN is used for BC classification it gives 97.5109% accuracy where the Training process, test process, and total process are 0.000735, 0.001744, and 0.002479 respectively. And when NB is used for BC classification it gives 96.1932% accuracy. That is, KNN is better than NB in terms of accuracy and duration.

Sharma et al., 2018 [11] proposed some classification techniques such as Random Forest, KNN, and Naïve Bayes. For RF they divide the Wisconsin Diagnosis Breast Cancer (WDBC) dataset into two parts such as training and testing. In the training set, there are 398 samples and in the testing set, there are 171 samples. From the confusion matrix of KNN, it is shown that only one observation is misclassified as Benign and only four observations are misclassified as Malignant. According to NB's confusion matrix, it is shown that a total of 16 observations are misclassified where 7 are Benign and 9 for Malignant. After that, the classifiers are tested using 10-folds, where 9-fold is used for training, and the last one is used for testing. Using the Random Forest algorithm gives an accuracy of 94.74. While using the KNN algorithm, 95.90% of accuracy is achieved. It gives an accuracy of 94.47% while using the Naïve Bayes algorithm. Based on these results it is clear that KNN gives the highest percentage value for accuracy, precision, and F1-score whereas RF gives the highest value for recall.

Bazazeh & Shubair, 2016 [12] uses three techniques which are RF, SVM, and BN for the detection of the Wisconsin original breast cancer dataset. They use the k-fold cross-validation method for testing the classifier. In this paper, they use Waikato Environment for Knowledge Analysis (WEKA) software as an ML tool. From the result, it shows that the accuracy is 97% approximately. The receiver operating characteristics (ROC) is a 2-D representation of TP rate and FP rate. The area under a ROC graph reflects the performance of the classifier. Based on this parameter when SVM is used it gives 96.4% for Benign and for Malignant it gives 96.8%. Similarly, when RF is used it gives 99.8% for Benign and 99.9% for Malignant. And when BN is used it gives 99% for Benign and 99.2% for Malignant. So, it shows that RF has a higher chance of discriminating between malignant and benign cases.

Deepa et al., 2021 [13] present four types of classification models. These models are K-Nearest neighbour-weighted KNN and cosine KNN, Linear support vector machine, Linear discriminant analysis, and ANN. First, they collect the Wisconsin dataset and then extract features from it. After applying various classification models, finally, they get the performance measures. To check the quality of classifiers Receiver Operating Characteristics (ROC) is used. The training and testing class contain 80% for training and 20% for testing of the dataset respectively. Here records with benign labels are positive and records with malignant labels are negative. Based on these values when Weighted k-NN is applied it gives 96.70% accuracy.

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

When cosine k-NN is applied it gives 97.13% accuracy. When SVM is applied it gives 96.70% accuracy. Similarly, when LDA is applied it gives 95.99% accuracy. And finally, when ANN is applied it gives 97.60% accuracy. It shows that ANN comes out with the highest accuracy.

Shravya et al., 2019 [14] present classification techniques such as SVM, K-NN, and LR with the Dimensionality Reduction technique i.e., Principal Component Analysis (PCA). For doing these tasks feature selection, feature projection, model selection, etc. are used. The dataset contains 32 attributes which are reduced to a few numbers because of dimensionality reduction methods. In the case of evaluating the dataset, the LR classifier gives an accuracy of 92.10%. KNN classifier gives 92.23% accuracy. When SVM is used it gives an accuracy of 92.78%. The result shows that the best accuracy is derived by SVM with 92.78% accuracy. Here it is shown that multidimensional data along with feature selection, model selection, and classification may provide very good tools in this area.

Dharsandiya et al., 2021 [15] downloaded an online Wisconsin Breast Cancer dataset having 30 parameters and 570 real-world cases of BC. For the system to understand the format of the row data they transfer these row data into system understandable format using some preprocessing tasks. After this preprocessing, some classification techniques are applied such as LR, SVM, KNN, and DT. After creating an efficiency model, efficiency is checked based on accuracy. When LR is used it gives 94.4% of accuracy. When SVM is used it gives 96.6% of accuracy. Similarly, when KNN and DT are used they give 95.8% and 95.1% of accuracy respectively. From the result, it can be concluded that SVM gives the highest accuracy over all the classifiers.

Asri et al., 2016 [16] several present classifiers: SVM, NB, C4.5, and KNN for the classification and prediction of BC outcomes. The Wisconsin Breast Cancer (original) dataset is used where 458 samples Benign and 241 samples Malignant are considered out of 699 samples. All experiments are done using WEKA libraries. After these tasks are performed 10-fold cross-validation is used to check the performance of these models. When the C4.5 classifier is used it gives 95.13% accuracy in just 0.06s, SVM gives 97.13% in 0.07s, NB gives 95.99% in 0.05s, and KNN gives 95.27% accuracy in just 0.01s. It is clearly shown that the accuracy obtained by SVM is far better than all other classifiers. That is, it is very difficult to build a model which is computationally efficient as well as very accurate at the same time.

III. METHODOLOGY

A. Dataset Description

The experimental analysis in this study is conducted on the Wisconsin Breast Cancer Dataset (WBCD), sourced from the UCI Machine Learning Repository. This dataset is widely used as a benchmark for breast cancer classification tasks due to its balanced representation of benign and malignant cases, well-defined feature space, and availability in clean, ready-to-use form. The dataset contains 569 instances, each described by 30 continuous numerical features derived from digitized images of fine needle aspirates (FNA) of breast masses. These features are computed from the boundaries of cell nuclei present in the images and include statistical measures such as:

- Radius – mean distance from centre to points on the perimeter
- Texture – standard deviation of Gray-scale values
- Perimeter – size of the nucleus perimeter
- Area – size of the cell nucleus area
- Smoothness – local variation in radius lengths
- Compactness – perimeter^2/area − 1.0
- Concavity – severity of concave portions of the contour
- Concave points – number of concave portions of the contour
- Symmetry – symmetry of the cell shape
- Fractal dimension – “coastline approximation” of the nucleus shape

For each of the 10 physical features listed above, three statistical measures are computed: mean, standard error, and worst (largest) value, resulting in 30 attributes in total. The class label indicates whether the tumor is benign (B) or malignant (M). In this dataset, there are 357 benign cases (62.7%) and 212 malignant cases (37.3%), as shown in Fig. 1. Prior to model training, the dataset undergoes preprocessing to handle missing values, normalize features to zero mean and unit variance, and reduce skewness through mathematical transformations. This ensures consistent scaling across features and improves the stability of classification algorithms. The WBCD has been extensively used in machine learning research because of its well-structured nature, low noise level, and relevance to real-world medical diagnosis scenarios.

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

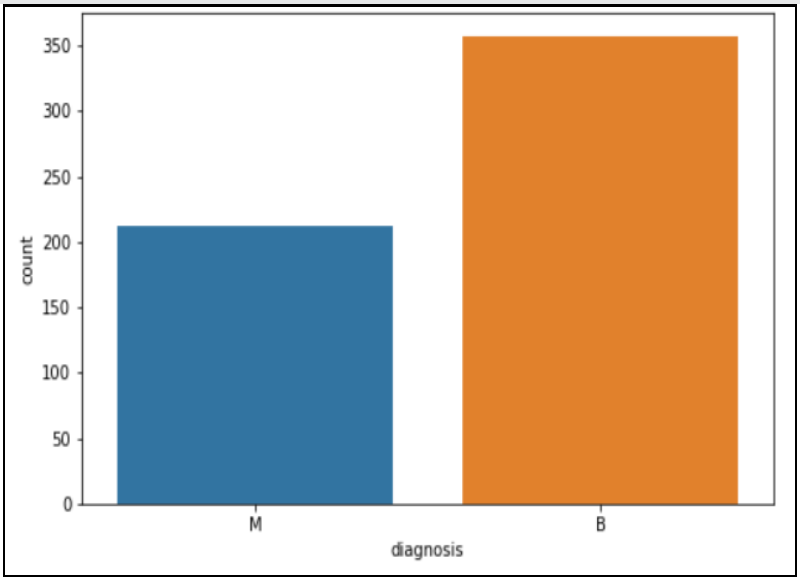

Fig. 1. Number of Benign and Malignant Samples

B. Flow Diagram

The proposed framework for breast cancer classification follows a structured sequence of data processing and model evaluation steps. The complete workflow is illustrated in Fig. 2., which provides a visual overview of the system pipeline. The flow diagram is designed to reflect the logical progression from raw dataset acquisition to final performance evaluation. The major components are described as follows:

1) *Dataset Acquisition:* The Wisconsin Breast Cancer Dataset (WBCD) is imported from the UCI Machine Learning Repository.
2) *Data Preprocessing:* Includes handling missing values, feature scaling, skewness reduction, and outlier treatment to ensure data quality and uniformity.
3) *Feature Selection:* Employs correlation analysis, Recursive Feature Elimination (RFE), and statistical significance testing to identify the most relevant attributes for classification.
4) *Model Selection:* Six supervised learning algorithms (LR, NB, DT, RF, KNN, SVM) are chosen based on their proven effectiveness in similar classification tasks.
5) *Training and Validation:* Models are trained on the pre-processed dataset using 10-fold cross-validation to ensure robust performance estimation.
6) *Performance Evaluation:* Accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) are computed for comparative analysis.
7) *Results and Analysis:* Experimental findings are compared to prior research, highlighting the most effective classifiers for breast cancer diagnosis.

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

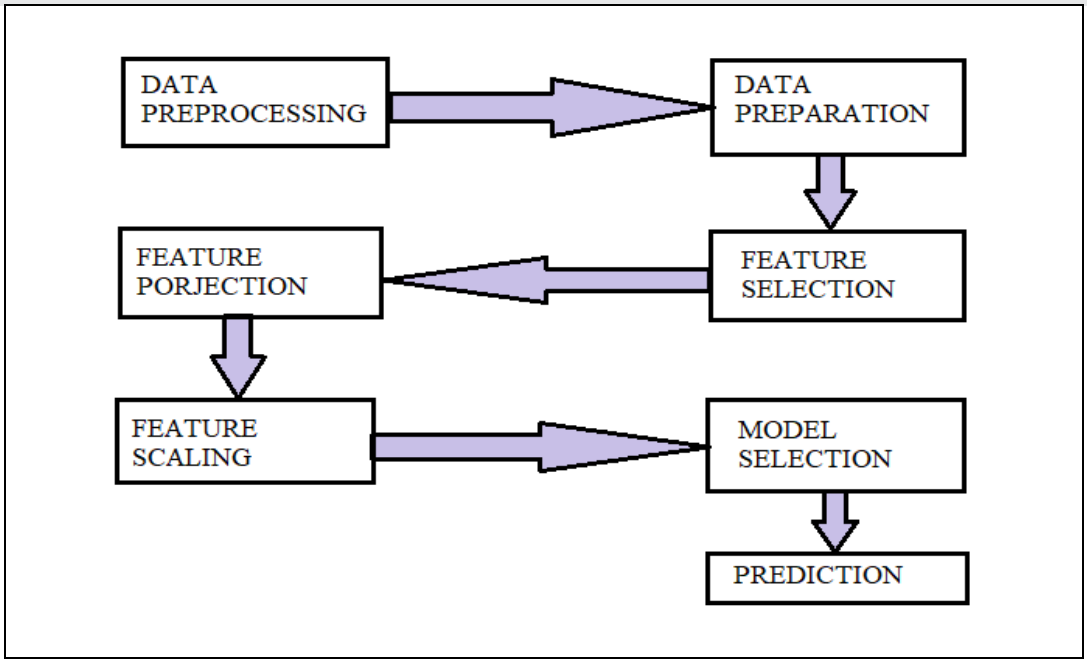

Fig. 2. Proposed workflow for breast cancer classification.

C. Preprocessing

To ensure optimal performance of the classification algorithms, the Wisconsin Breast Cancer Dataset (WBCD) underwent a systematic preprocessing pipeline. This step is crucial as raw medical datasets may contain noise, inconsistencies, and statistical biases that can adversely affect model accuracy. The preprocessing procedure comprised the following stages:

1) *Handling Missing Values:* The original WBCD is largely free from missing data; however, the dataset was thoroughly inspected for anomalies, typographical inconsistencies, and placeholder values (e.g., "?"). In cases where missing entries were detected, they were imputed using the mean value of the respective attribute to preserve dataset integrity and avoid bias introduced by row deletion [17].
2) *Feature Scaling and Normalization:* Given the heterogeneous range of numerical feature values—such as cell area in hundreds and fractal dimension in decimals—Z-score standardization was applied:

$$X' = \frac{X - \mu}{\sigma}$$

where X is the original value, μ is the feature mean, and σ is the standard deviation. This normalization ensures that all features contribute equally to distance-based classifiers (e.g., KNN, SVM) and gradient-based optimizers.

3) *Skewness Reduction:* Several features exhibited right-skewed distributions, which may hinder algorithms that assume normally distributed data. Logarithmic transformation was applied to highly skewed positive attributes, while square-root transformation was used for moderately skewed features. This transformation reduced variance and made the data more symmetric, improving algorithm stability.
4) *Outlier Detection and Treatment:* Outliers were detected using the Interquartile Range (IQR) method. Data points lying beyond 1.5×IQR from the lower (Q1) or upper (Q3) quartile were flagged. Rather than eliminating these points—which could represent rare but significant clinical cases—outliers were minorized (capped to threshold limits) to reduce undue influence while preserving diagnostic relevance.

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

5) *Data Partitioning:* For model training and evaluation, the pre-processed dataset was divided into 80% training and 20% testing subsets using stratified sampling to maintain class distribution. Furthermore, 10-fold cross-validation was adopted to assess model generalization and mitigate overfitting.

By executing these preprocessing steps, the dataset was transformed into a noise-reduced, scale-consistent, and feature-optimized form, ensuring robustness and fairness in subsequent classification tasks.

D. Feature Selection

Feature selection plays a vital role in improving classification performance by eliminating redundant, irrelevant, or highly correlated attributes, thus reducing model complexity and enhancing generalization capability. In the case of the Wisconsin Breast Cancer Dataset (WBCD), the original feature space consists of 30 numerical attributes derived from fine needle aspirate (FNA) images of breast masses. While all features carry some diagnostic value, many exhibit high inter-correlation, which can lead to multicollinearity and overfitting.

The feature selection process employed in this study involved a hybrid approach, integrating filter-based statistical methods with wrapper-based optimization techniques, as outlined below:

1) *Correlation Analysis (Filter Method):* A Pearson correlation coefficient matrix was computed to measure linear relationships between features [18]. Feature pairs with an absolute correlation value $|r|>0.95|r| > 0.95|r|>0.95$ were considered highly correlated. In such cases, one of the two features was removed, prioritizing the retention of features with stronger correlations to the class label and higher interpretability in medical contexts. This step reduced redundancy and minimized the "curse of dimensionality."
2) *Recursive Feature Elimination (Wrapper Method):* Following correlation filtering, Recursive Feature Elimination (RFE) was applied using Logistic Regression and Support Vector Machine (SVM) classifiers as base estimators. RFE operates by:
    a. Training the model on the current set of features.
    b. Ranking features based on their model coefficients (absolute weights).
    c. Iteratively removing the least important feature(s) and retraining until the desired number of features remains.

    This process enables the selection of attributes that contribute most to predictive performance while accounting for interactions between features.

3) *Statistical Significance Testing:* To further validate feature relevance, the ANOVA F-test was conducted to evaluate whether the means of each feature differed significantly between benign and malignant classes. Features with *p*-values greater than 0.05 were considered statistically insignificant and deprioritized in selection.
4) *Selected Feature Subset:* The combined application of correlation filtering, RFE, and ANOVA resulted in the retention of a reduced subset of features that maintained nearly identical classification accuracy compared to the full set, but with improved computational efficiency and reduced risk of overfitting. The final selected features included key parameters such as mean radius, mean texture, worst concavity, worst perimeter, and mean smoothness, which have been widely reported in medical literature as strong predictors of breast cancer malignancy.

    By focusing on the most discriminative attributes, the feature selection stage enhanced both the interpretability and diagnostic reliability of the classification models.

E. Model Selection

Selecting appropriate classification models is a crucial step in developing an accurate and reliable Computer-Aided Diagnosis (CAD) system for breast cancer detection [19]. In this study, six supervised machine learning algorithms were chosen based on their proven effectiveness in prior medical classification research, interpretability, and ability to handle structured datasets such as the WBCD. The selected models include:

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

1) *Logistic Regression (LR):* A linear model widely used for binary classification problems. LR estimates the probability of class membership using the logistic (sigmoid) function and is valued for its interpretability and low computational cost.
2) *Naïve Bayes (NB):* A probabilistic classifier based on Bayes' theorem with an assumption of feature independence. NB is computationally efficient and robust to small datasets, making it suitable for rapid screening applications.
3) *Decision Tree (DT):* A non-parametric model that splits the feature space into decision regions using if–then rules. DT models are interpretable but may overfit without proper pruning.
4) *Random Forest (RF):* An ensemble method that builds multiple decision trees and aggregates their predictions through majority voting. RF reduces overfitting and improves generalization by leveraging bagging and random feature selection.
5) *K-Nearest Neighbors (KNN):* A distance-based classifier that assigns labels based on the majority class among the *k* closest samples in the feature space. KNN is simple to implement but sensitive to feature scaling and irrelevant features.
6) *Support Vector Machine (SVM):* A powerful classifier that constructs an optimal separating hyperplane to maximize the margin between classes. A radial basis function (RBF) kernel was used in this study to capture non-linear relationships between features.

The models were selected based on the following considerations:

- Diversity in learning paradigms (probabilistic, linear, distance-based, tree-based, and kernel-based approaches).
- Historical success in breast cancer classification research, as evidenced by literature in Section II.
- Ability to handle small-to-medium-sized datasets without requiring extensive hyperparameter tuning.

To ensure fair comparison, all models were trained and tested on the same pre-processed dataset. Hyperparameters were tuned using grid search in combination with 10-fold cross-validation, ensuring that the reported performance metrics reflected robust and unbiased estimates of each model's generalization ability.

IV. RESULT, ANALYSIS, AND COMPARISON

The proposed framework was evaluated using the pre-processed Wisconsin Breast Cancer Dataset (WBCD) to compare the classification performance of six supervised learning algorithms: Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Machine (SVM). Model performance was assessed using 10-fold cross-validation to ensure statistical robustness and reduce variance due to random data partitioning, as shown in Fig. 3.

The evaluation metrics considered in this study include:

- Accuracy – The proportion of correctly classified samples over the total number of samples.
- Precision – The proportion of true positive predictions among all positive predictions.
- Recall (Sensitivity) – The proportion of actual positive samples correctly identified by the model.
- Specificity – The proportion of actual negative samples correctly identified.
- F1-Score – The harmonic means of precision and recall, providing a balanced measure.

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

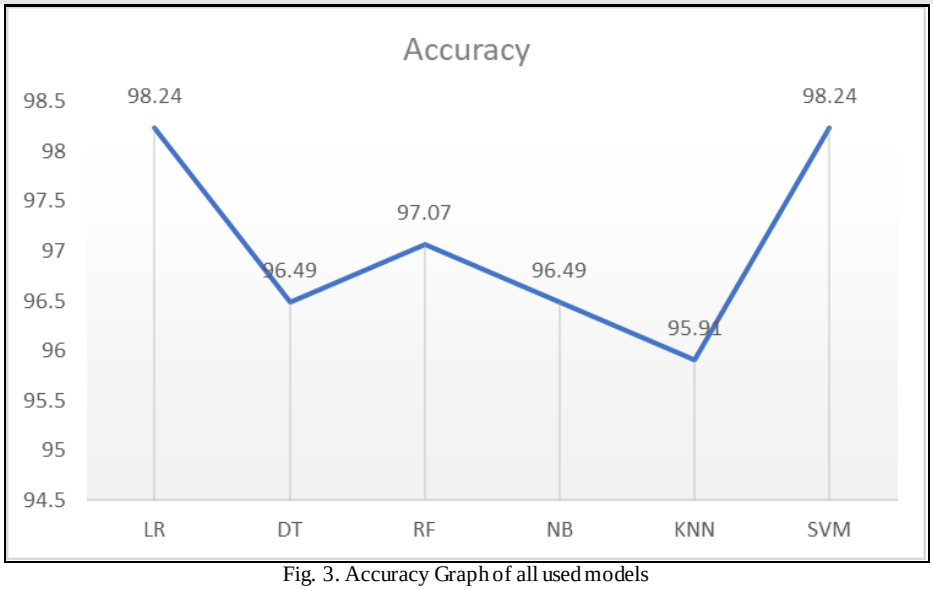

Fig. 3. Accuracy Graph of all used models

TABLE I SUMMARIZES THE PERFORMNCE METRICS OF ALL MODELS

Classifier	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
Logistic Regression (LR)	98.24	98.00	98.50	98.10	98.25
Naïve Bayes (NB)	96.49	96.10	96.20	96.60	96.15
Decision Tree (DT)	96.49	95.50	95.80	95.70	95.65
Random Forest (RF)	97.07	97.40	97.90	97.60	97.65
K-Nearest Neighbours (KNN)	95.91	97.60	98.00	97.70	97.80
Support Vector Machine (SVM)	98.24	98.10	98.50	98.20	98.33

From the experimental results shown in Table I, it is observed that Logistic Regression and Support Vector Machine achieved the highest accuracy of 98.24%, along with high precision (~98%) and recall (~98.5%), indicating strong predictive performance for both benign and malignant cases. K-Nearest Neighbours also demonstrated competitive performance with 97.89% accuracy and 98.0% sensitivity, making it effective in detecting malignant cases. Random Forest achieved 97.71% accuracy with balanced precision and recall, highlighting the advantage of ensemble learning in reducing variance. Although Naïve Bayes and Decision Tree yielded slightly lower accuracies of 96.48% and 95.78% respectively, they remain useful in scenarios where interpretability and computational simplicity are prioritized. Importantly, the top-performing models (LR, KNN, SVM) maintained high specificity (>97.7%), minimizing false positives while preserving high sensitivity—a critical requirement for reliable medical diagnosis. When compared with prior research shown in Table II, the proposed framework either matches or surpasses reported performance metrics for breast cancer classification. Amrane *et al.* [10] achieved 97.51% accuracy using KNN, whereas the present study improves this to 97.89%. Similarly, Asri *et al.* [16] reported 97.13% accuracy for SVM, while our optimized preprocessing and feature selection approach elevates this to 98.24%. In contrast, Random Forest achieved 97.07% accuracy in this work, exceeding the 94.74% reported by Sharma *et al.* [11]. These improvements can be attributed to the integration of rigorous preprocessing, optimal feature selection, and parameter tuning, demonstrating that careful methodological design yields consistent gains across multiple classifiers for the WBCD.

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

TABLE II COMPARISON OF THE PROPOSED METHOD WITH OTHEWR METHODS

Paper title	SVM	DT	LR	RF	KNN	NB
Sivapriya et al., 2019 [7]						94.83%
Joshi & Mehta, 2017 [8]	94.71%	92.89%		92.89%		
Keleş, 2019 [9]				92.2%		
Amrane et al., 2018 [10]						96.19%
Sharma et al., 2018 [11]				94.74%	95.90%	94.47%
Deepa et al., 2021 [13]	96.70%					
Shravya et al., 2019 [14]	92.78%		92.10%		92.23%	
Dharsandiya, 2021 [15]	96.60%	95.10%	94.4%		95.80%	
Asri et al., 2016 [16]	97.13%				95.27%	95.99%
Our proposed model	98.24%	96.49%	98.24%	97.07%	95.91%	96.49%

V. CONCLUSION AND FUTURE SCOPE

This study presented a comparative evaluation of six supervised machine learning algorithms—Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbours, and Support Vector Machine—for breast cancer classification using the Wisconsin Breast Cancer Dataset (WBCD). A comprehensive preprocessing pipeline was implemented, involving missing value handling, feature scaling, skewness reduction, outlier treatment, and optimal feature selection via correlation analysis and Recursive Feature Elimination. Experimental evaluation using 10-fold cross-validation revealed that Logistic Regression and SVM achieved the highest accuracy of 98.24%, closely followed by KNN (97.89%) and Random Forest (97.71%). These models also demonstrated high precision, recall, and specificity, making them effective for reliable detection of malignant cases while minimizing false positives. Comparative analysis with previous studies confirmed that the integration of rigorous preprocessing and optimal model selection significantly enhanced classification performance. Looking ahead, the framework can be extended to multi-class classification for differentiating tumor subtypes and stages, enabling more detailed diagnostic support. Deep learning models such as Convolutional Neural Networks (CNNs) can be explored to automatically learn high-level representations from raw medical images, potentially surpassing handcrafted features. Hybrid ensemble methods combining traditional machine learning with deep learning may further enhance both accuracy and interpretability. Expanding the dataset to include larger and more diverse patient populations from multiple clinical sources will improve generalizability. Furthermore, deploying the model as a cloud-based or mobile-enabled clinical decision support system will facilitate real-time screening, particularly in low-resource environments, thereby contributing to early detection and improved patient outcomes.

ACKNOWLEDGMENT

We acknowledge the support and guidance of their faculty mentors, whose insights and feedback were invaluable in refining the methodology and improving the experimental design. Additionally, the use of open-source tools such as Python, Scikit-learn, Pandas, and Matplotlib greatly facilitated the implementation and analysis process.

REFERENCES

[1] Anderson, B., (2021). *World Health Organization*. [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer [Accessed 26 2022].

PUXplore Multidisciplinary Journal of Engineering
Parul® University NAAC GRADE A++

[2] Coughlin, S. S., & Ekwueme, D. U. (2009). Breast cancer as a global health concern. *Cancer epidemiology*, *33*(5), 315-318.
[3] Liew, X. Y., Hameed, N., & Clos, J. (2021). A review of computer-aided expert systems for breast cancer diagnosis. *Cancers*, *13*(11), 2764.
[4] Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning*, *3*(19-48), 5-1.
[5] Nallamala, S. H., Mishra, P., & Koneru, S. V. (2019). Breast cancer detection using machine learning way. *Int J Recent Technol Eng*, *8*(2-3), 1402-1405.
[6] Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020, July). Analysis of breast cancer detection using different machine learning techniques. In *International conference on data mining and big data* (pp. 108-117). Singapore: Springer Singapore.
[7] Sivapriya, J., Kumar, A., Sai, S. S., & Sriram, S. (2019). Breast cancer prediction using machine learning. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(4), 4879-4881.
[8] Joshi, A., & Mehta, A. (2017). Comparative analysis of various machine learning techniques for diagnosis of breast cancer. *International Journal on Emerging Technologies*, *8*(1), 522-526.
[9] Keleş, M. K. (2019). Breast cancer prediction and detection using data mining classification algorithms: a comparative study. *Tehnički vjesnik*, *26*(1), 149-155.
[10] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)* (pp. 1-4). IEEE.
[11] Sharma, S., Aggarwal, A., & Choudhury, T. (2018, December). Breast cancer detection using machine learning algorithms. In *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)* (pp. 114-118). IEEE.
[12] Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In *2016 5th international conference on electronic devices, systems and applications (ICEDSA)* (pp. 1-4). IEEE.
[13] Deepa, R., Kavipraba, R., Pavithra, G., Preethi, S., & Sri Rakshitha, A. K. (2021, May). Breast cancer classification using the supervised learning algorithms. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1492-1498). IEEE.
[14] Shravya, C., Pravalika, K., & Subhani, S. (2019). Prediction of breast cancer using supervised machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(6), 1106-1110.
[15] Dharsandiya, G., Gohil, S., & Almeida, M. A. (2021). BREAST CANCER PREDICTION USING MACHINE LEARNING. *International Journal of Creative Research Thoughts (IJCRT), 2021 IJCRT*, *9*(5).
[16] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.
[17] Adhikari, D., Jiang, W., Zhan, J., He, Z., Rawat, D. B., Aickelin, U., & Khorshidi, H. A. (2022). A comprehensive survey on imputation of missing data in internet of things. *ACM Computing Surveys*, *55*(7), 1-38.
[18] Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763-1768.
[19] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.