Feature Reduction and Svm-Based Ensemble Machine Learning Techniques for Breast Cancer Prediction

Roselinevinnarasi.A[1], Hannah Inbarani H[2]
[1]*Research Scholar, Periyar University, Salem-11.*
[2]*Professor, Periyar University, Salem-11.*

***Abstract*—Breast cancer is leading health problem for the global community, and it is vital to screen breast cancer in an early stage. Machine learning (ML) is a powerful method that should be utilized in the diagnosis process since it finds complex trends in medical data. The Principal Component Analysis (PCA), K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) derived from soft voting have been postulated in the research paper. PCA method reduces the dimensions and redundancy. SVM assigns high decision boundaries compared to KNN which has a superior local classification. The team was tested using Wisconsin Breast Cancer Diagnostic (WBCD) data. It achieved 98% accuracy, 1.00 precision, 0.99 recall, and 0.99 F1-score, which is better than the scores of its individual classifiers. This method showed competitive performance with respect to Logistic Regression and reduced computation costs. This proves that it can be utilised to predict breast cancer and can be employed in medical diagnostics in general.**

***Index Terms*—KNN, Logistic Regression, Machine Learning, Multilayer Perceptron, PCA, Random Forest, SVM, WBCD.**

I. INTRODUCTION

One of the most predominant cancers that impact many women and is a leading cause of mortality is breast cancer [1]. Diagnosis should be accurate and timely. Machine learning (ML) techniques (ML) have become an effective tool in the medical diagnostics cycle because it can be applied to non-decomposable data and make reasonable clinical decisions. While Support Vector Machine (SVM), it works well on high dimensional datasets while K-Nearest Neighbor (KNN) is good at modelling local patterns but can't tolerate noise and irrelevant attributes.Other methods like Random Forest (RF), Logistic Regression (LR), Multilayer Perceptron (MLP) & XGBoost have been used as well, and each of those techniques has its unique advantages but also there is an inconsistency in the superiority of each approach from dataset to dataset [2]. To address these limitations, this paper suggests an ensemble of SVM and KNN with soft voting and aided by PCA. PCA is used for dimensionality reduction, SVM is used for global decision boundary and KNN is used for local classification. Experiments on the Wisconsin Breast Cancer Dataset (WBCD) shown that this integrated approach outperforms the individual classifiers while giving results nearly comparable to established classifiers such as RF, XGBoost and MLP [3].

A. Research Objectives

The primary "goal of the research is to achieve effective, accurate breast cancer prediction using a hybrid machine learning" approach. This study evaluates the performance of limitations for individual classifiers, such as SVM, KNN, LR, Random Forest, XGBoost as well as MLP on WBCD dataset, and also applies Principal Component Analysis for feature reduction. PCA is applied to reduce high dimensionality to preserve the data variance and a PCA-assisted soft-voting ensemble method, while combining SVM and KNN, is proposed. The SVM finds a global decision boundary for strength, and KNN finds a local neighbourhood for learning capability. Hyper parameter tuning is carried out in this work using 5-fold cross-validation to ensure robust model optimisation. An ensemble model is compared with traditional classifiers using "accuracy, precision, recall and F1-score. This research" contribute novel, computational approach with an effective hybrid model that improves diagnostic reliability and supports more accurate clinical

decision-making, distinguishing benign and malignant breast tumours.

B. Research Contribution

This research introduces a novel PCA-assisted soft voting ensemble model that integrates SVM and KNN to enhance the breast cancer prediction for accuracy. While applying PCA, the study achieved dimensionality reduction to improve efficacy and reduce noise in the dataset. The overall Experimental evaluation highlights how the ensemble model performs classical and deep-learning methods on structured medical data. This research study contributes a reliable and accurate model that supports and improves clinical decision-making in identifying benign as well as malignant breast cancer tumours.

II. LITERATURE REVIEW

Recent studies have widely discovered the application of machine learning and ensemble methods for breast cancer classification and diagnosis. Jabber and Meera [1] proposed an ensemble-based classification method using Bayesian networks and radial basis function models, reaching an accuracy of 96.72%; however, the model suffered from limited Clarity and Model interpretability. Varsha Nemade and Vishal Fegade [2] showed a comparative study of multiple machine learning algorithms, with decision trees, k-nearest neighbors, support vector machines, random forests, and naïve Bayes classifiers, reporting an accuracy of 97%. Despite the high performance, the study highlighted challenges related to generalizability and the absence of external validation. Sam Khozama and Ali M. Mayya [3] presented a range-based breast cancer prediction model based on Bayes' theorem, collective with ensemble learning, employing classifiers such as SVM, KNN, decision trees, random forests, logistic regression, and naïve Bayes, but achieved a comparatively lower accuracy of 85.3%, with limitations in score interpretation and clinical actionability.

Additional improvements were demonstrated by Aqeel Ahmed Khan et al. [4], who combined dimensionality reduction techniques, such as principal component analysis and factor analysis, with machine learning models, with multilayer perceptron, SVM, logistic regression, random forests, and KNN, reaching a high accuracy of 98.64%. However, the inclusion of dimensionality reduction introduced challenges in model interpretation. Amreen Batool and Yung-Cheol Byun [5] proposed an adaptive voting ensemble learning algorithm that utilises SVM, naïve Bayes, KNN, and logistic regression, achieving 98.18% accuracy; however, the method suffers from increased computational complexity. Tayyaba Yasmeen et al. [6] presented a comparative study of advanced ensemble techniques, including random forest, XGBoost, and stacking methods, reporting an accuracy of 91.1%, while emphasizing issues related to feature selection quality and class imbalance. In addition, Disha H. Parekh and Vishal Dahiya [7] examined early breast cancer finding using a combination of ensemble and predictive machine learning models such as random forest, AdaBoost, naïve Bayes, XGBoost, KNN, and decision trees, achieving an accuracy of 89%, but with well-known concerns about overfitting, scalability, and deployment. Mohammed Amine Naji et al. proposed a common voting ensemble classifier including SVM, naïve Bayes, C4.5, KNN, simple logistic regression, and random forest algorithms, reaching an accuracy of 86.9%. Nevertheless, the study was limited by a limited dataset validation and the lack of real-time testing. Overall, although existing approaches demonstrate promising accuracy levels, challenges related to interpretability, generalizability, computational complexity, and real-world deployment remain unresolved, thereby motivating the need for more robust and clinically appropriate breast cancer prediction models [8].

III. METHODOLOGY

In this paper, the ensemble method is built upon SVM and KNN utilising BC data. The ensemble classifier is modelled utilising the SVM and KNN. These subdivisions will include two single classifiers and a proposed ensemble classifier. The dataset employed in the experiment is the WBCD [4]. Data consists of 569 samples, each characterised by 30 numerical variables. The variable of interest is to evaluate if cancer is benign or malignant.

1. *Total Samples: 569*
2. *Benign: 357*
3. *Malignant: 212*
4. *Features: 30*

Alternative Models for Evaluation

Besides SVM and KNN, various other models of ML were widely applied in the prediction as well as classification of breast cancer. These models encompass:

Logistic Regression (LR): A linear "model appropriate for binary classification, frequently serving as a reference point.

Random Forest (RF): An ensemble method consisting of decision trees, which enhances predictive accuracy and minimizes overfitting.

XGBoost: A gradient boosting technique recognized for its high precision and effectiveness with structured datasets.

Multilayer Perceptron (MLP): A neural network variant skilled at capturing complex, non-linear interactions within the data.

A. Svm

Both linear and non-linear datasets could be classified using supervised ML approaches such as SVM. It also determines the global boundary of the separating hyperplane between the various classes.

Hyperplane: In 2D space, a line separates data points into class *Margin:* It is the distance from the hyperplane to the closest support vectors.

Mathematical function

The training data with attributes $x_i$and labels y_i

The "hyperplane is described by the following equation

$$w^T x + b = 0 \quad (1)$$

we represents weight vector, which is perpendicular to the hyperplane.b represents bias (offset).The Decision function is defined as

$$f(x) = sign(w^T x + b) \quad (2)$$

1. Input: Train the dataset where $x_i \in R^n$ and $y_i \in \{0,1\}$
2. Choose the Kernel function k (x_i, y_i)
3. RBF $K(x,y) = e^{-\gamma||x-y||^2}$
4. Fits SVM with an RBF kernel.
5. Transform data implicitly using the Kernel.
6. Decision Function is $f(x) = \text{sign}(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b)$
7. Output: Predicted class 0(benign)or class 1(malignant)

Algorithm for SVM

B. Knn

The K-Nearest Neighbour is commonly used for real-time usage.This algorithm operates on principle of comparing labeled features of unknown sample with those of its neighbouring labeled samples, derived Euclidean distance or, in some cases, Manhattan distance. The Euclidean distance is square root of squared differences among corresponding points in sample space, and the following equation provides formula used to calculate Euclidean distance. Mathematical function,

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

where $x_i$and y_i is the position of" data samples.

1. Input**:** Training set D = $\{(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)\}$ Where a query point q, and the number of neighbors is K.
2. For each x_iIn D: compute distance d (q, x_i)=$\|q-x_i\|2$.
3. Sort each training sample by distance d(q, x_i).
4. Select K nearest neighbors.
5. Assign q the label y that occurs the most among the neighbors
6. Output: Predicted label for q(benign or malignant).

Algorithm for KNN

C. Ensemble Model

An Ensemble model is supervised ML technique. This model is used to create an accurate predictive model. This approach involves a diverse range of supervised learners to improve the model's predictive ability. There are different types of ensemble learning: Bagging, Boosting, Stacking, and Voting. Bagging is a technique that reduces variance. It can be merged with weak classifiers by taking the average of their predictions. This method is also known as parallel ensemble learning. It works well for complex models, especially those with high variance and low bias. The boosting aims to reduce variance and bias. Stacking is a method that trains data using multiple algorithms as a meta-modeling approach. Voting combines predictions from two models directly. There are 2 types of voting. It's a soft and hard vote. Soft voting takes into account the average predicted probabilities and chooses the one with the highest value, while Hard voting employs majority rules based on the predicted two classes.

D. Ensemble Method of Svm and Knn

To enhance prediction accuracy, a soft voting ensemble can be utilized by combining SVM and KNN classifiers. In soft voting, each underlying classifier provides class probabilities rather than hard classifications. The ensemble averages these probabilities as well as determines the class with the highest average probability.

Steps: Train Base Classifiers

Fit an SVM model on the training dataset (potentially using a suitable kernel such as RBF or polynomial).
Train a KNN model on the identical training dataset (select an optimal number of neighbors, k).
Predict Probabilities: Employ predict_proba (or a similar method) to obtain class probabilities from both the SVM and KNN.Combine predictions from SVM & KNN using:Soft Voting: Average predicted probabilities

$$y = \arg\max\left(P_{\text{SVM}}(y \mid x) + P_{\text{KNN}}(y \mid x)\right)/2 \qquad (4)$$

In this paper, we applied the ensemble method of soft voting to the Algorithm

ALGORITHM PCA-ASSISTED SVM-KNN ENSEMBLE
INPUT: BREAST CANCER DATA, NUMBER OF COMPONENTS _N, SVM, NUMBER OF NEIGHBORS K=5.
OUTPUT: PREDICTED LABEL Y.
STEP 1: LOAD THE BREAST CANCER DATASET.
STEP 2: APPLY PCA AND SCALE THE DATA.
STEP 3: TRAIN THE SVM AND KNN CLASSIFIERS ON PCA-TRANSFORMED DATA.
APPLY SUPPORT VECTOR MACHINE (SVM) AND K-NEAREST NEIGHBOR(KNN).
STEP 4: APPLY A SOFT VOTING CLASSIFIER.
OBTAIN PROBABILITIES BASED ON SVM AND KNN.
DETERMINE THE FINAL PROBABILITIES

$$p_{ensemble=}\frac{p_{svm}+p_{knn}}{2}$$

ASSIGN Y=ARGMAX($p_{ensemble}$)
STEP 5: RETURN PREDICTED LABELS Y.

Proposed Ensemble algorithm for SVM and KNN

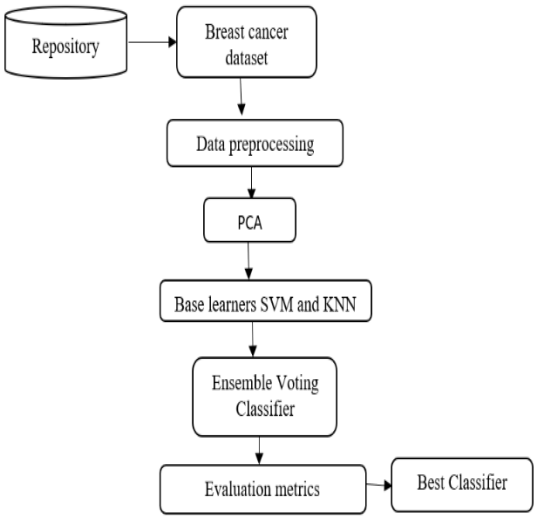

Fig I: Proposed Architecture for the ensemble method

The dataset is fed into preprocessing for scaling and feature reduction. The same processed data is given to both SVM and KNN. Each classifier makes its prediction. Predictions are combined in the Ensemble Layer using a Voting strategy. The final output is generated as the ensemble decision.We used PCA to reduce number of dimensions in dataset, keeping top 10 components. This approach retained over 90% of the variance, enabling efficient computation while preserving essential information.“For the K-Nearest Neighbor (KNN) classifier, we chose the value of k = 5 based on direct results from cross-validation. This choice helps balance bias and variance while improving classification accuracy”. This option balanced bias and variance, leading to the best classification performance on the breast cancer dataset.frameworks for better clinical decision support.

E. Hyperparameter Tuning

We used grid search combined with 5-fold cross-validation for hyperparameter adjustment in order to maximize model performance. For SVM, the regularization parameter C and kernel coefficient γ were varied over a predefined range. For KNN, multiple values of K (e.g., 3, 5, 7, 9) were tested. Combination of parameters that achieved highest cross-validation accuracy has been selected for the final ensemble model.

F. *Evaluation Strategy*

To ensure that the performance evaluation is robust and unbiased, we employed K-Fold Cross-Validation (k = 5). In each fold, 20% of data has been set aside for testing, and remaining 80% has been utilised for training (including PCA fitting and model learning). This process has been repeated across all folds, and the final performance metrics were reported as the average over all folds. To prevent data leakage, preprocessing steps—comprising scaling as well as PCA transformation—have been fitted exclusively on the training set of each fold and then applied to the corresponding test set. While WBCD dataset attended as primary benchmark in this study, we also recognize the importance of external validation. Future work will extend the evaluation to additional breast cancer datasets such as WDBC, METABRIC, and BreakHis to insure generalizability as well as robustness of proposed method across diverse patient populations**.**

IV. RESULT AND DISCUSSION

The predictive performance of various ML models SVM (RBF), KNN, Random Forest, LR, XGBoost, MLP, and the ensemble method (soft voting)—has been evaluated utilising precision, recall, and F1-score.

Discussion

Individual Classifiers

SVM (RBF) showed strong recall (0.97), meaning it was effective at correctly identifying malignant tumors.KNN (K=5) achieved moderate results, with recall slightly lower due to its sensitivity to noisy data and local patterns. Random Forest provided the best overall individual performance (F1-score 0.97), benefiting from ensemble decision trees. Logistic Regression achieved high precision (0.98) but slightly lower recall (0.90), indicating it was more conservative in labeling cases as malignant.XGBoost delivered balanced performance, though its recall (0.91) limited its F1-score. Multilayer Perceptron (MLP) captured non-linear feature relationships and achieved a strong F1-score (0.96), confirming that neural models can perform competitively on tabular medical data.

Ensemble Method (Soft Voting)

The soft voting ensemble outperformed all individualclassifiers with Precision=1.00, Recall=0.98, and F1-score=0.99.This highlights the effectiveness of combining classifiers, as the ensemble leveraged the strengths of both high-recall models (SVM, MLP) and high-precision models (Random Forest, Logistic Regression). The ensemble's robustness makes it a strong candidate for deployment in clinical decision support systems.

Table I: Performance Comparison of Different Models on the Breast Cancer Dataset

Model	Precision	Recall	F1-score
SVM (RBF)	0.95	0.97	0.94
KNN (k=5)	0.93	0.91	0.92
Random forest	0.98	0.95	0.97
Logistic regression	0.98	0.90	0.94
XGBoost	0.98	0.91	0.91
Multilayer perception	0.97	0.91	0.96
Ensemble (Soft Voting)	1.00	0.98	0.99"

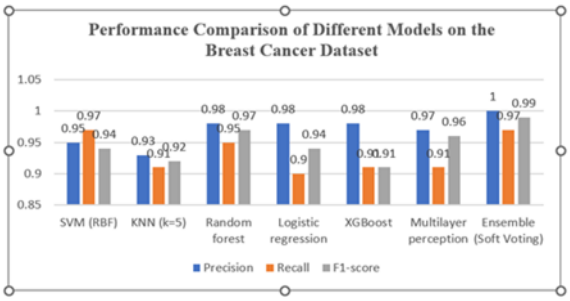

Fig V: Performance Comparison

Fig II: Performance comparison of Different Models on the Breast Cancer Dataset.

Key Insights

Models with higher recall (SVM, Ensemble) are particularly valuable in cancer detection since missing malignant cases (false negatives) can be critical. Precision also matters to avoid unnecessary biopsies; here, Random Forest, LR, and the Ensemble performed best. Overall, the ensemble method demonstrated the most reliable and accurate predictions, validating the benefit of integrating multiple ML algorithms.

Table II: Accuracy Comparison

"Model	Accuracy
SVM (RBF)	0.95
KNN (k=5)	0.96
Random Forest	0.97
Logistic Regression	0.95
XGBoost	0.95
Multilayer perception	0.97
Ensemble (Soft Voting)	0.98"

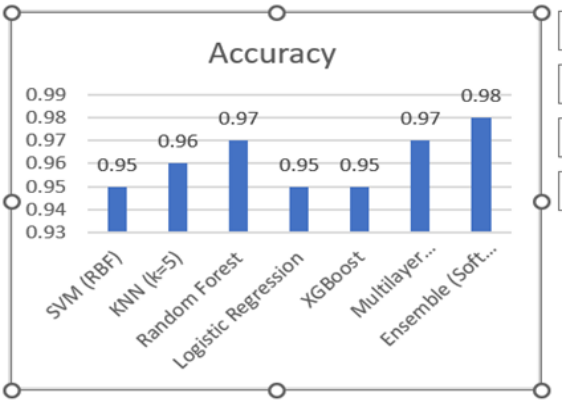

Fig III: Comparative analysis of accuracy

V. CONCLUSION

In this paper, several ML models including SVM, KNN, LR, Random Forest, XGBoost, MLP, and ensemble method were compared with the breast cancer dataset. The results revealed that although the classifiers performed well individually, soft voting ensemble model performed well among all the other models. It got precision (1.00), recall (0.98), as well as F1-score (0.99). Ensemble approach has been successful because it merged high recall of SVM and MLP with high precision of Random Forest and LR. This is especially so when diagnosis of breast cancer is done, where false negativity is an issue that should be reduced to achieve early diagnosis and early treatment. On the whole, the outcomes confirm that ensemble learning is an effective method in solving medical diagnosis problems. It is more accurate and reliable than individual classifiers. Future work can include extending this method to larger as well as more diverse datasets, refining hyperparameters, incorporating it into deep learning infrastructure.

REFERENCES

[1] Meerja Akhil Jabbar.,2020, Breast cancer data classification using ensemble machine learning**,** Engineering and Applied Science (EASR), pp 65-72.

[2] Yanxia Sun,ibomoiye domor mienye.,2022, A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects (IEEE), vol 10, pp 29-49.

[3] Sam Khozama, Ali M. Mayya,2022, A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning, Information Technology and Control,vol 51, pp 757-770.

[4] Varsha Nemade, Vishal Fegade2, 2023, Machine Learning Techniques for Breast Cancer Prediction, International Conference on Machine learning and Data Engineering,pp 1314–1320.

[5] Taarun Srinivas, Aditya Krishna Karigiri Madhusudhan, Joshuva Arockia Dhanraj, Rajasekaran Chandra Sekaran, Neda Mostafaeipour, Negar Mostafaeipour, and Ali Mostafaeipour,2022, Novel-Based Ensemble Machine Learning Classifiers for Detecting Breast Cancer, Hindawi Mathematical Problems in Engineering, pp 1-16.

[6] Tayyaba Yasmeen1, Muazzam Ali1, M U Hashmi1, M Adnan Hashmi, and Zeeshan Mehmood3, 2024, A Comparative Study of Advanced Machine Learning Ensemble Techniques for Classification of Breast Cancer, Journal of Computing & Biomedical Informatics vol 8, pp 1-13.

[7] Aqeel Ahmed Khan, Muhammad Abu Bakr,2024, Enhancing Breast Cancer Diagnosis with Integrated Dimensionality Reduction and Machine Learning Techniques, Journal of Computing & Biomedical Informatics, vol 7, pp 1-17.

[8] Disha H. Parekh, Vishal Dahiya,2023, Early Detection of Breast Cancer Using Machine Learning and Ensemble Techniques,vol 22, pp 231-237.

[9] Behrouz Zolfaghar, Leila Mirsadeghi, Khodakhastbibak, Kaveh Kavousi,2023, Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning, ACM Computing Surveys, Vol. 55, pp 1-34.

[10] Zakaria Senousy, Mohammed M. Abdelsamea, Mohamed Medhat Gaber, Moloud Abdar, U Rajendra Acharya, Abbas Khosravi, and Saeid Nahavandi,2021, IEEE Transactions on BiomedicalEngineering,pp 1-13.

[11] Omar El Alaoui, Hasnae Zerouaoui& Ali Idri, 2022, Deep Stacked Ensemble for Breast Cancer Diagnosis, Springer natural link, vol 468, pp 435-445.

[12] Bouchra El Ouassif, Ali Idri, and Mohamed Hosni,2021, Homogeneous Ensemble-based Support Vector Machine in Breast Cancer Diagnosis, conference paper in Research Gate, pp 350-360.

[13] Khan, M. F., Iftikhar, A., Anwar, H., & Ramay, S. A. (2024). Brain Tumor Segmentation and Classification using Optimized Deep Learning. Journal of Computing & Biomedical Informatics, 7(01), pp 632-640.

[14] Hasan, M., Abedin, M.Z., Hajek, P., Coussement, K., Sultan, M.N. and Lucey, B., 2024. A blending ensemble learning model for crude oil price forecasting. Annals of Operations Research, pp.1-31.

[15] Bukhari, O., Agarwal, P., Koundal, D. and Zafar, S., 2023. Anomaly detection using ensemble techniques for boosting the security of an intrusion detection system. Procedia Computer Science, 218, pp.1003-1013.

[16] Amirhossein Ahmadi, Mojtaba Nabipour, Behnam Mohammadi-Ivatloo, and Vahid Vahidinasab. 2021. Ensemble learning-based dynamic line rating forecasting under cyberattacks. *IEEE Trans. Pow. Deliv.* 37, 1 (2021), pp 230–238.

[17] Moloud Abdar, Mariam Zomorodi-Moghadam, Xujuan Zhou, Raj Gururajan, Xiaohui Tao, Prabal D. Barua, and Rashmi Gururajan. 2020. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recog.Lett.* 132 (2020), 123–1.