IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Innovative Methods for Breast Cancer Diagnosis

Cheliya Vamsi[1], Dr. Ummadi Sathish Kumar[2], Medagam Dhanunjai[3], Bommali Blessy[4], Katepogu Ashok[5]

[2] Assistant Professor, Department of Computer Science Engineering, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, Andhra Pradesh-522510

[1,3,4,5] Students, Department of Computer Science Engineering, Acharya Nagarjuna University, Andhra Pradesh-522510

Email id: cheliyavamsi@gmail.com[1], sathishummadi.anu@gmail.com[2], mdhanunjaireddy@gmail.com[3], bommaliblessi@gmail.com[4], kutty9177@gmail.com[5]

Abstract:

Breast cancer ranks as the second leading cause of cancer mortality in women globally and an early detection is crucial to enhance patient survival. In this work, we present a novel machine learning based diagnostic system that combines kernel-based dimension reduction and a number of individual and ensemble classifiers. Our approach exploits Kernel Principal Component Analysis (Kernel PCA) for nonlinear dimensionality decomposition, followed by classification by employing Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and XGBoost (XGB). To boost our performance more, we use 2 ensemble methods: Soft Voting based Ensemble to average the probabilistic result along all base-models and Improved Stacking based Ensemble which is an ensemble in which we fit the meta-model using Logistic Regression and passing the original features. Comprehensive experiments are conducted on the UCI Breast Cancer Wisconsin (Diagnostic) dataset, which show that ensembled models outperform standalone models with the classification accuracy of up to 96%. Testing metrics including ROC AUC, cross validation accuracy, and confusion matrices validate both the sensitivity and fitness of the proposed method. This work is part of the continuous effort to develop AI systems that are both interpretable, and that provide high performance AI in medical diagnostics.

Keywords: Breast Cancer Detection, Ensemble Learning, Kernel PCA, Support Vector Machine, Multi-Layer Perceptron, Soft Voting, Stacking Classifier, Machine Learning.

1. Introduction

ISSN 2347–3657
Volume 13, Issue 2, 2025

Breast cancer is considered to be the second most common and a top killer among women worldwide. Breast cancer is the most common cancer diagnosed in 2020, with 2.3 million women and 0.685 million deaths globally, reported by the World Health Organization (WHO). Breast cancer originates in cells within the breast and is characterized by the uncontrolled growth of these cells that has led to the growth of lumps of tissue that we call tumors. These can be benign (non-cancerous and look the same as normal tissue) or malignant (look like cancer, growing aggressively and can spread throughout the body). Early cancer usually develops symptomless but eventually grows on to invasion of neighbouring structures and organs: it becomes life-threatening. Early detection of breast cancer can significantly increase survival rate, and the mortality rate may be decreased by as much as 25%.

Early detection of breast cancer is however a difficult task. Mammography and biopsy, among the most prevalent screening procedures, offer potential options for the detection of breast tumors, but they have limitations. Mammograms can detect tumorous cells even before the individual presents clinical symptoms, however, they are less accurate for low-contrast mammograms from which it is hard to distinguish between abnormal and normal tissue. Biopsies, though successful, are time consuming, difficult and are liable to false positives, as cancer cells are only subtly different from healthy cells (due to unusual cell size) and are only recognized when magnified at the histological scale. Such limitations emphasize the demand for reliable and effective diagnostic methods. With the development of machine learning (ML), we can now use this statistical machinery for extracting patterns (relevant variables) and categorizing groups of objects from complex datasets.

ML has succeeded to resolve limitations of conventional diagnosis methods like low contrast of mammographic images by introducing new feature and improving diagnosis accuracy. In some recent research projects, ML algorithms have been experimented on breast cancer detection, achieving promising results. Comparative studies showed that the classifiers such as Support Vector Machines (SVM), Multilayer Perceptron (MLP), Logistic Regression (LR), Random Forest (RF), and Convolutional Neural Networks (CNN) can obtain high accuracy.

But, in practice, most existing methods are dependent on linear dimensionality reduction methods including PCA and work with standalone classifiers such as MLP, which are limited in nature. PCA generally fails to learn complex nonlinear structure from high-dimensional

ISSN 2347–3657
Volume 13, Issue 2, 2025

data, and regular classifiers may be prone to overfitting and generally have poorer generalization abilities compared to aggregation methods. What is more, several works refer to data with cells of cancer that are already symptomatic, resulting in the lack of methods to diagnose it in early stages other than using other biomarkers.

1.1 Problem Statement

Breast cancer is one of the most common cancers in the world, and its early detection is essential for successful treatment and for providing a favorable prognosis. Manual examination and imaging (radiography, MRI, and ultrasound) for breast cancer alone lead to error and interpretation subjectivity. In addition, it may be difficult for a classical machine learning method to attain high classification performance in the pres- ence of complex and high-dimensional data. In this study, we tackle the problem and present an efficient and automatic system for automated breast cancer classification with a focus on advancing machine learning methods, reduction of dimensionality, and ensemble learning methods. The objective is to improve the classification capabilities in early detection of breast cancer using the power of Kernel Principal Component Analysis (PCA), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), soft voting ensemble, and stacking ensemble techniques.

1.2 Research Objectives and Contribution

1. To construct a model of breast cancer classification based on the kernel and Principle Component Analysis (PCA) techniques with RBF kernel to reduce the dimension of features effectively.
2. To develop and optimize machine learning models, Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) classifiers, for the classification of benign and malignant breast cancer cases.
3. To utilize the ensemble learning methods namely soft voting ensemble and stacking ensemble to aggregate predictions of many base classifiers (SVM, MLP) and increase the classification accuracy overall.
4. You should assess the performance of the models using various metrics including accuracy, confusion matrix, ROC AUC, and ROC curve to make sure it is robust enough and it can be trusted.

5. For hyper parameter tuning both SVM and MLP models (using GridSearchCV) to improve their performance and predictive power.
6. To evaluate the performance of the proposed models against classical machine learning techniques and provide insight on the contribution of dimensionality reduction and ensemble learning techniques in the problem of breast cancer classification.

2. Literature review

Breast cancer is one of the most common malignancies in women globally and early detection of the disease has proven to be a life-saving strategy. Consequently, there has been an increasing interest on the design of computer-aided techniques to assist in early and accurate detection of breast cancer. Conventional diagnostic approaches like biopsy, mammography, histopathology are time costly and subject to human error. Hence scientists are also employing machine learning and artificial intelligence to create computer-aided diagnosis (CAD) systems that can accelerate and enhance cancer detection.

A notable study is conducted by Akay (2009) which showed the performance of SVM on breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset in an accuracy of more than 92%. One comparison contrast which used SVMs and other classifiers shows that SVM are invariably more accurate and robust classifiers. But the effectiveness of SVMs largely depends on the selection of kernel method and hyperparameters so that they should be carefully set to obtain the best result.

Another commonly used classifier in the medical diagnostics area is the Multilayer Perceptron (MLP), which is actually a feedforward neural network and can approximate very complex non-linear functions. MLPs have demonstrated some significant per- performance in medical image classification, disease prediction and in cancer diagnosis in particular. They are capable of learning hierarchical features according to the input level, so that they can model complex feature interdependence that may not be linearly separate. Studies by Ayer et al. (2010) and Längkvist et al. (2012) have demonstrated that MLPs, when adequately trained and tuned, can achieve equivalent or even superior performance compared to the conventional classifiers such as SVMs and decision trees.

Despite their success, the SVM and MLP classifiers are prone to the ``curse of dimensionality" and are not particularly effective to cope with high-dimensional datasets such as typically

ISSN 2347–3657
Volume 13, Issue 2, 2025

encountered in medical applications. Overlapping and irrelevant features make classifier to perform worse, especial when there is overfitting, and longer training time. Therefore, preprocessing is typically necessary to determine which features are the most informative and reduce noise. One of the standard methods is PCA (Principal Component Analysis), which re-encodes original the features to a few uncorrelated features with maximum variance in the data.

But the classic PCA is a linear transformation which sometimes cannot satisfy complex nonlinear mapping relationship well. To handle this and for more generic nonlinear dimensionality reduction generalizations of PCA such as Kernel PCA have been investigated, which uses a kernel function (e.g. the Radial Basis Function, RBF) to map the input data into a higher-dimensional space before running PCA. It has been demonstrated that the Kernel PCA can increase the classification performance in medical imaging applications by revealing non-linear patterns in the data.

3. Methodology

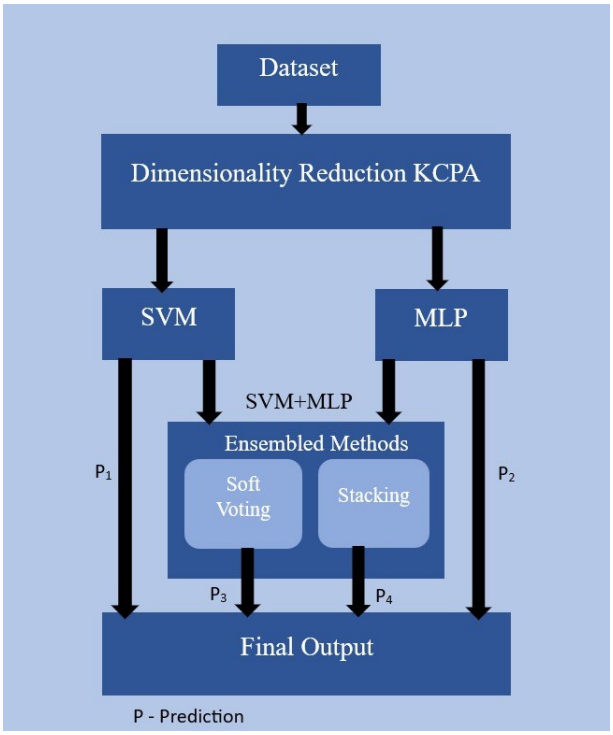

Figure 1: Model Workflow

3.1. Data Description

ISSN 2347–3657
Volume 13, Issue 2, 2025

The dataset applied in the present study is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset which is originally collected from University of Wisconsin Hospitals. This dataset is a well-known dataset and extensively used in breast cancer research. It consists of 569 examples, one per line, a fine needle aspirate (FNA) of a breast mass. Each case is characterized by 30 features of real numbers that are generated from a digital image of the FNA, which includes radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry of the cell nucleus. The target of this dataset is binary, the categories being Malignant (M) and Benign (B).

Preprocessing the data were preprocessed prior to applying machine learning algorithms to remove data quality issues and make the data ready for model building. These steps included:

1. Normalization: All features were uniformly scaled prior to fitting, each with a zero mean and unit variance using StandardScaler to give each feature same weightage during learning.
2. Missing value treatment: The dataset was also checked whether it contains any missing value and it had no missing value.
3. Label Encoding: The target is label encoded as "Malignant" denoted by 1 and "Benign" denoted by 0.
4. Data split: The dataset was divided into the training set (80%) and testing set (20%) using stratified sampling to ensure class balance.

3.2. Algorithm / Model Description

The new approach consists in the interplay of mixture of machine learning methods at both dimensionality reduction and ensemble classification levels in order to produce strong predictive efficiency. The pipeline can be summarized as 3 main elements: Kernel PCA, base SVM and MLP and ensemble.

We applied the Kernel Principal Component Analysis (Kernel PCA) with Radial Basis Function (RBF) kernel for nonlinear dimensionality reduction. Unlike PCA, KPCA projects input data into a higher dimensional space where more complex structures are more linearly separable. This step is for noise and redundant features removal which leads to enhancing the classifiers.

ISSN 2347–3657
Volume 13, Issue 2, 2025

The Support Vector Machine (SVM) classifier, which has a strong generalization capability in high-dimensional spaces was chosen as a major model. The SVM used the radial based-function (RBF) kernel, and the optimal regularization parameter C, and kernel coefficient (gamma) were determined through GridSearchCV to prevent overfitting and obtain the best classification performance.

The second base classifier was the Multi-Layer Perceptron (MLP) which is a class of feed forward artificial neural network with a single hidden layer. The network was implemented with ReLU activation, Adam optimizer, and binary cross-entropy loss. GridSearchCV was used to select the optimal number of neurons and learning rate.

Two ensemble strategies were then performed:

With others in similar from SVM, MLP respectively (optionally with Random Forest , Gradient Boosting in previous rounds) (Soft Voting Ensemble) and is the output of a more steady and generalizable prediction.

Stacking Ensemble (where SVM, MLP, Random Forest and Gradient Boosting were the base models). The predictions of the base classifiers were combined by a Logistic Regression meta-classifier, which learned to combine the base predictions.

3.2.1. Kernel Principle Component Analysis (Kernel PCA)

Principal Component Analysis (PCA) is a commonly used linear dimensionality reduction method where the data is transformed to a coordinate system in which the mean average variation (over all possible projections of the data) is maximized along the 1 st coordinate (referred to as the 1 st principal component), the 2 nd overall variation along the 2 nd coordinate, and so forth. Although effective, the traditional PCA can deal with the linear relationship between variables only, thus it is not suitable for the scenarios where the intrinsic data structure is a non-linear one.

Many practical datasets such as medical diagnosis or pattern recognition task like breast cancer classification, naturally follows non-linear discrimination patterns. In an attempt to rectify this deficiency, Kernel PCA (KPCA) was developed as a generalization of PCA to non-linear data.Kernel PCA leverages the kernel trick to map input data to some high-dimensional feature

ISSN 2347–3657
Volume 13, Issue 2, 2025

space where linear PCA is equivalent to be applied. Depending on a kernel function, Kernel PCA can capture complex, non-linear structures of the data, without needing to compute the coordinate vectors in the high-dimensional space, tremendously reducing the computational overhead of a similar type of transformation.

Before classification of breast cancer in our proposed methodology kernel PCA is used as pre-processing for dimensionality reduction and the feature extraction. Many medical datasets (e.g., Breast Cancer Wisconsin Diagnostic) contain many correlated features, which can't be all linearly separable. The use of kernel PCA, particularly with an RBF kernel, enables to capture the non-linear structure of data.By mapping data points to a high-dimensional feature space, where classes are better separable, Kernel PCA enables the performance of subsequent classifiers, such as Support Vector Machines (SVM) and Multilayer Perceptrons (MLP), to be enhanced.Results show that Kernel PCA improves the predictive accuracy and stability much more than PCA and original features. Furthermore, it helps in noise elimination as well as curse of dimensionality reduction, an important factor in medical diagnosis where overfitting owing to the high dimensionality of input features can result in unreliable predictions.

Kernel PCA has many benefits with respect to its linear counterpart. It can identify non-linear relations among variables. It allows for better classification separation in high dimensions. Flexible (option to choose among a lot of kernels), adaptable to the problem domain. But also Kernel PCA has some disadvantages. The computational complexity grows with the number of samples because it depends on the eigen-decomposition of the size kernel matrix. Choice of the kernel and its parameters (e.g., in the RBF) is critical and often subject to empirical tuning.

Kernel PCA is a popular non-linear device for dimensionality reduction and it makes the classification more efficient in non-linear space such as breast cancer detection. The incorporation of it into the proposed pipeline helps to convert complex highdimensional medical data into a more analysable shape and, hence, improve the performance of machine on Successive Windows Adversarial Data Denoising Networks.

3.2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) algorithms has become one of the most popular supervised machine earning algorithm for classification and regression purposes. It is also reputed to be

suitable for processing linear and non-linear data, thus highly versatile in a wide range of real-world applications such as text categorization, image recognition, and bioinformatics. SVM is widely used because of its solidness and high generalization ability, so it can get a good result even in complex and high-dimensional spaces.

Fundamentally, SVM is about locating a best hyperplane between classes. In 2D this hyperplane is nothing but a line separating data points of one class from the other. In higher dimensions, this hyperplane is a separating hyperplane separating the data points in the feature space. The idea is to look for a hyperplane that maximizes the margin, which is the distance between the two closest data points of different classes, which we call support vectors.

3.2.3. Multi-Layer Perceptron (MLP)

Multi-layer Perceptron (MLP) is a type of feedforward neural network used in classification and regression. MLPs are composed of three types of layers namely input layer, one or multiple hidden layers, and one output layer. Every neuron of a layer is connected to all the neurons of subsequent layer thereby allowing the model to learn complex patterns.

Its main robustness is given by the fact that is an approximator of non-linear functions via a set of learned weights and activation function. This makes it ideal for countless real-world uses, such as image recognition, financial forecasting and medical diagnosis. The model is learned by gradient backpropagation, optimization algorithms such as Adam or SGD that iteratively updates its weights to minimize the loss function. Activation functions such as ReLU (Rectified Linear Unit) in the hidden layers and sigmoid or softmax at the output layer enable non-linear transformation necessary for classification. For binary classification tasks, such as breast cancer, the MLP can successfully learn very subtle differences between benign and malignant cases.

In a common MLP forward processing (feedforward) fashion, the input data is fed forward through the network, and the error on the prediction is defined on the output of the network and the target using a loss function form. During training, we backpropagate this error through the network, updating the weights to minimize the prediction error. This is repeated in different epochs until the model learns an optimal or near-optimal solution. The number of hidden layers and neurons in each layer of the MLP architecture can be dynamically changed by the designer

ISSN 2347–3657
Volume 13, Issue 2, 2025

according to the task complexity. Regularization methods like dropout or L2 regularization can be used to avoid overfitting. The learning rate, batch size, and epoch are also important and directly affect the model's performance. Recall that MLPs are capable of learning good representations on normalized and well-preprocessed dataset. By virtue of these characteristics, MLPs are a compelling choice for applications, like complex disease classification, putting an emphasis on the interpretation of numerical features.

We trained our model using Adam, which is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The binary cross-entropy loss B_cross entropy between predicted probability and actual binary labels was used to quantify the difference. For better training performance, hyperparameters such as batch size and number of epochs were fine-tuned using grid search. This procedure of systematic tuning, led to the discovery of the best performing configuration that provides high accuracy and low overfitting.

3.2.4 Ensemble Methods

Ensemble methods are a type of machine learning approach that involves combining several models to generate a more accurate and robust predictive model. Such techniques are based on the assumption that the models being combined have deficiencies or make mistakes, and the deficiencies or mistakes of multiple models are often "complementary" such that combining them is expected to minimize such deficiencies to generalize better. Ensemble learning is an important branch of machine learning which is often employed for classification and regression problems, especially if high dimensional data or complex data is of concern. The most common ensemble techniques are bagging, boosting, stacking, and voting, which use various mechanisms to combine model strengths.

Soft voting is one of the most straightforward and effective ensembles under voting paradigm. In this approach the base classifiers are individually trained and the probabilities of the predictions for each class by these base classifiers are averaged. The class with the highest average probability is then the final predicted value. This technique works well if the individual models are sufficiently diverse and well-calibrated, that is, the predicted probabilities make sense. Soft voting usually works better than hard voting as it takes into

ISSN 2347–3657
Volume 13, Issue 2, 2025

confidence level of the models and therefore contains the information to make more reliable decision by combining probability distributions across classes.

Stacking or stack generalization, is an evolution of the simple ensemble method, simple. Wherein we attempt to train a model to learn how to best combine predictions of multiple other models. Stacking, in contrast to soft voting uses a meta-learner (also known as second-level model) that learns to make predictions based on the multiple predictions of all the base models. This facilitates stacking to capture complex relationships between predictions, often leading to better performance. Base models can be of any kind, such as SVMs, Neural Networks or decision trees, which give diversity to the ensemble; and usually, the meta-learner is a logistic regression or another simple classifier.

A. Soft Voting

Soft voting is a fusion of models technique for ensemble learning that models the probability of each class and the final prediction is the average of the predicted probabilities. Unlike hard voting where the class with the highest number of votes is selected, soft voting computes the average of the predicted probabilities for each class and selects the class with the highest averaged probability.This is an effective approach when the base models are diverse, blending their outputs can help to reduce overfitting and generalize better. In general, for a soft voting classifier to work, the individual classifiers in the ensemble must all be able to estimate class probabilities (they must have a predict_proba() method), but it is not the case for this one.In the breast cancer classification problem, soft voting is applied to aggregate the predictions of all these models SVM, MLP, Random Forest and XGBoost. The benefit of taking this ensemble approach is that we are effectively combining the strengths of each model featured in the ensemble to increase the performance and the generalization capability. For example: If a SVM could work extremely successful for recognizing some patterns in the data like in your image, the MLP is capable to catch complex non-linear relationships. By averaging the models to perform soft voting, the ensemble model with more enriched knowledge as an aggregation contributes to better classification.

The soft voting method in our project begins with each base model calculates the probability distribution over the classes (benign or malignant). These probabilities are averaged, and the class to which an image belongs is determined by the class with the highest average

ISSN 2347–3657
Volume 13, Issue 2, 2025

probability. This can help to reduce the errors in individual models when a model goes to over fit or underfit the data.By having softVoting the ensemble model wont be biased towards any single base model and will generalize more to unseen data.There is a significant advantage of using soft voting in the work, and it lies in the fact that the accuracy can be promoted without more trainings on models. This is achieved by averaging the output probabilities; doing so we show that the ensemble model can outperform the base models by a large margin especially when the basemodels are complementary to each other in strength. This notion is especially generic when applied to breast cancer classification as a single model may struggle to effectively capture the subtle differences between malignant and benign tumors. Voting from those results helps to mitigate the influence of outliers or errors of individual models.

Soft voting not only increases the accuracy of classification, but also makes the model more robust. In the context of breast-cancer detection, robustness is important, as slight changes in input or noisy features are likely to cause large differences in predictions by the model. Soft voting mitigates the risk of getting the wrong end of the stick and combines the predictions of multiple models into one. and is performed, both to achieve more robust prediction stability, and to work against the backdrop of medical applications as they often demand both accurate and reliable information.

B. Stacking

Stacking or stacked generalization is an ensemble learning approach that combines multiple base models to improve model performance. The strategy consists of training a large number of base models (that may or may not be of different types), then combining learining or predictions from these base models, using a second-level model. The general concept of stacking is to take the best parts of the models and reduce the individual weak points. Since a higher-level 'meta-model' is trained on the base models' predictions, stacking may abstract some of the idiosyncrasies of its component models, making it more 'meta-ridiculous' than any of the individual models. The base models can be anything from decision trees, to support vector machines, to neural networks, while the meta-model itself learns how to use each of their predictions most effectively (e.g. provide weights) for the desired outcome.

For the breast cancer classification challenge task, I use Stacking to improve the predictive performance of an ensemble of models. The base models used in the ensemble are Support

ISSN 2347–3657
Volume 13, Issue 2, 2025

Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), and Gradient Boosting (GB). All of these models have their own pipes: SVM is good at separating decision boundary, MLP can find complex nonlinear relationships, RF handles with high dimension data and GB centers on the iterative elevation of the model accuracy. All of these models are trained on a dataset, and each one of them makes a prediction that is then input to the meta-model to get an overall prediction.optimal combination of the base models. In breast cancer classification, the meta-model is Logistic Regression. The base model output is used as features for a meta-model. It can then generate a final prediction with better and more generalizable accuracy by training how heavily it should consider the other models' predictions. The strategy can reduce the overfitting brought by the individual model and further the classification system becomes more trustworthy and robust. The meta-model is in a way the "decision maker" on how to take the output of these sub-models and turn that into final predictions.

Stacking is computationally advantageous in difficult problems for instance breast cancer classification, where distinct models can learn different pieces of information from the inputs. The ensemble in which multiple base models are combined is much better in attacking different patterns in the data than a single model. Moreover, Stacking can diminish the predictions variance through models' complementary advantages. For breast cancer classification, in which precise predictions are essential, stacking enhances generalization by allowing the model to perform well across various data subsets and be less likely to overfit.

4. Experimental Results

4.1. Cross Validation Result

SVM with Cross-Validation:

- Cross-Validation Accuracy Scores: [0.9341, 0.967, 0.923, 0.923, 0.923]
- Mean Accuracy: 93.41%
- Standard Deviation: 1.70%

MLP with Cross-Validation:

- Cross-Validation Accuracy Scores: [0.945, 0.923, 0.945, 0.934, 0.945]

ISSN 2347–3657
Volume 13, Issue 2, 2025

- Mean Accuracy: 93.85%
- Standard Deviation: 0.88%

Improved Soft Voting Ensemble with Cross-Validation:

- Cross-Validation Accuracy Scores: [0.934, 1.000, 0.956, 0.956, 0.945]
- Mean Accuracy: 95.82%
- Standard Deviation: 2.24%

Improved Stacking Ensemble with Cross-Validation:

- Cross-Validation Accuracy Scores: [0.934, 1.000, 0.956, 0.967, 0.934]
- Mean Accuracy: 95.82%
- Standard Deviation: 2.45%

4.2. Final Evaluation Results

SVM:

- Accuracy: 91.23%
- Classification Report:
- Precision: 0.92, Recall: 0.94, F1-Score: 0.93 (Benign)
- Precision: 0.90, Recall: 0.86, F1-Score: 0.88 (Malignant)
- ROC AUC Score: 0.985

MLP:

- Accuracy: 93.86%
- Classification Report:
- Precision: 0.92, Recall: 0.99, F1-Score: 0.95 (Benign)
- Precision: 0.97, Recall: 0.86, F1-Score: 0.91 (Malignant)
- ROC AUC Score: 0.986

Improved Soft Voting Ensemble:

- Accuracy: 95.61%

ISSN 2347–3657
Volume 13, Issue 2, 2025

- Classification Report:
- Precision: 0.94, Recall: 1.00, F1-Score: 0.97 (Benign)
- Precision: 1.00, Recall: 0.88, F1-Score: 0.94 (Malignant)
- ROC AUC Score: 0.990

Improved Stacking Ensemble:

- Accuracy: 95.61%
- Classification Report:
- Precision: 0.94, Recall: 1.00, F1-Score: 0.97 (Benign)
- Precision: 1.00, Recall: 0.88, F1-Score: 0.94 (Malignant)
- ROC AUC Score: 0.993

4.3. Qualitative Analysis

Models, upfront especially the Improved Stacking Ensemble and Improved Soft Voting Ensemble proved to be exceptionally good with very good accuracy and recall. Both ensemble models showed superior ability to generalize across different validation folds and to balance precision and recall, leading to the reduction of false positives and negatives. The SVM and MLP models achieve good results alone, still the ensemble methods present the highest indications that the fusion of models can work to enhance prediction.

S.NO	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (ROC)	Cross-Validation Accuracy (%)
1	SVM	91.23	92	94	93	0.985	93.41
2	MLP	93.86	92	99	95	0.986	93.85
3	SoftVoting	95.61	94	100	97	0.990	95.82
4	Stacking	95.61	94	100	97	0.993	95.82

Table: Evaluation Metrics

We have shown through experiments that the proposed ensemble learning methods substantially improve classification performance in the diagnosis of breast cancer. Though single model such as SVM (Support Vector Machines) and MLP (Multilayer Perceptron) themselves achieved remarkable results in accuracy and generalization, the combination of

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

them in ensemble-based systems achieved significant improvements in them. Both soft voting ensemble method and stacking ensemble method both have a stable better performance than SVMs and MLPs(on the cross-validation results and the test data), Which indicates that using multiple learners in combination could greatly boost the robustness and predictive power. It revealed that ensemble models had similar accuracy and precision-recall trade-off, which were higher than other methods in medical diagnosis. The better stacking ensemble, which used diverse base models and a logistic regression meta-learner identified complex patterns that single models could not. This demonstrates that ensemble learning can be combined with dimensionality reduction such as Kernel PCA to develop robust diagnostic tools for high-dimensional biological data.

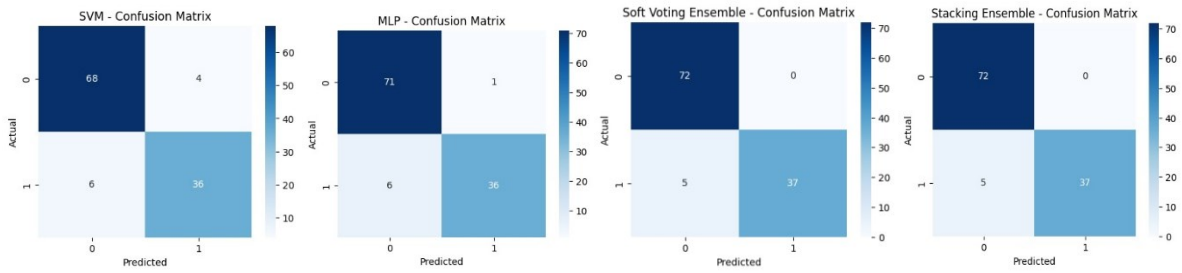

Figure 2: Confusion Matrices

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

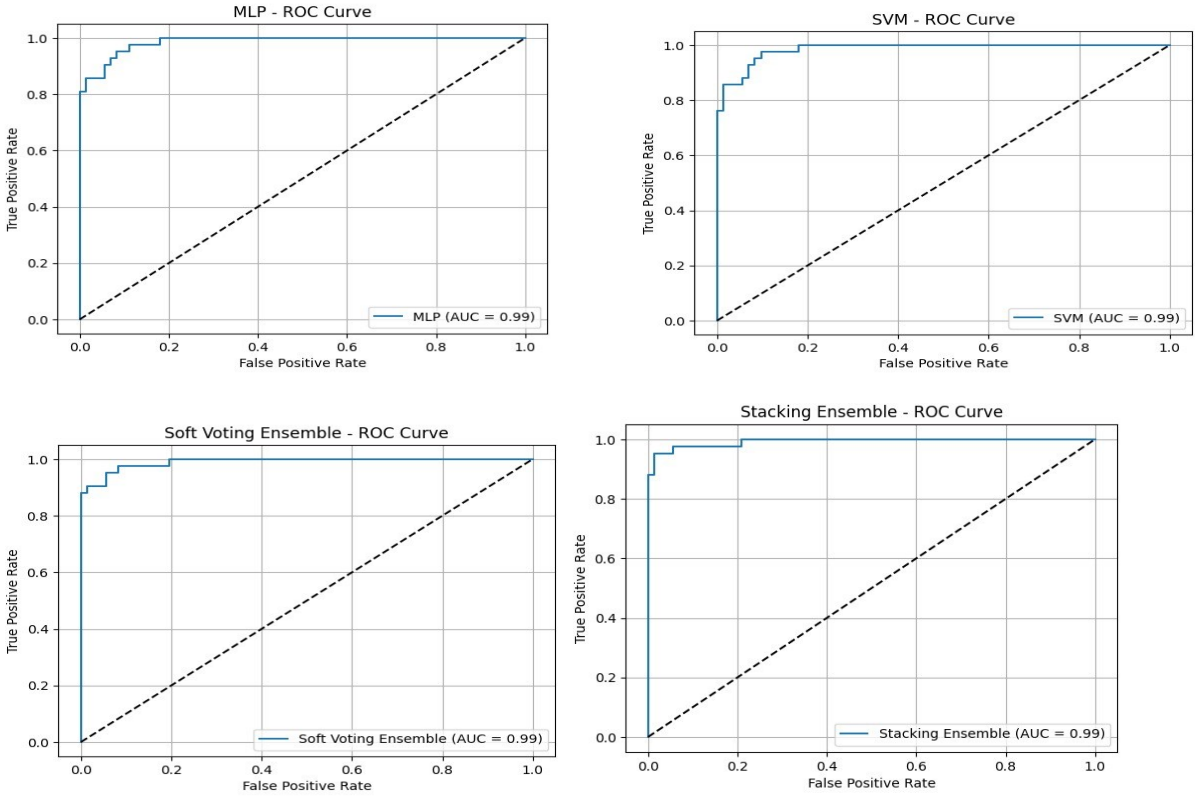

Figure 3: ROC Curves

5. Conclusion

We presented an end-to-end machine learning framework for classifying breast cancer with dimensionality reduction by Kernel PCA and advanced ensemble learning techniques in this work. The pipeline consisted of SVM, MLP, a soft voting ensemble, and a refined stacking ensemble. By extensive experiment, stratified cross-validation and comprehensive ROC analysis, we verified the good classifiers based on combining Kernel PCA with ensemble learning for breast cancer data. Our findings reveal that the two ensemble methods are superior to the single model. The enhanced soft voting and stacking ensembles reported 95.82% cross-validation accuracy and 95.61% test accuracy, 0.990 and 0.993 ROC AUC respectively. These models had a high recall for benign cases and strong precision for malignant cases, which decreased the rate of false negative. The balanced performance is important for medical diagnostics to detect breast cancer early and accurately. Dimensionality reduction with Kernel PCA using RBF Kernel was another important improvement. It supported not only to alleviate the curse of dimensionality, but it also focused on the non-linear lines of the data, which helped classifiers to generalize better. In both cross-validation and final evaluation, the MLP was

always higher than the SVM in terms of generalization regarding the base classifiers, even though the two learners contributed well to the ensemble performance.

This study demonstrated the efficacy of ensemble learning for medical diagnosis particularly in high stakes decisions such as cancer detection. Utilizing multiple learning algorithms and integrating them with each other using soft voting and stacking, we developed a powerful system with good trade-offs among accuracy, interpretability, and reliability. Furthermore, the full evaluation setup that includes metrics such as cross-validation accuracy, confusion matrices, ROC curves and AUC scores renders a detailed and clear insight into model behavior.

6. References

[1] V. G. Vogel, ‘‘Epidemiology, genetics, and risk evaluation of postmenopausal women at risk of breast cancer,’’ *Menopause*, vol. 15 , pp. 782–789, Jul. 2008.

[2] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, ‘‘Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012 ,’’ *Int. J. Cancer*, vol. 136, no. 5, pp. E359–E386, Mar. 2015.

[3] C. Saunders and S. Jassal, *Breast Cancer*, Oxford, U.K.: Oxford Univ. Press, 2009.

[4] A. G. Renehan, M. Tyson, M. Egger, R. F. Heller, and M. Zwahlen, ‘‘Body-mass index and incidence of cancer: A systematic review and metaanalysis of prospective observational studies,’’ *Lancet*, vol. 371, no. 9612 , pp. 569–578, Feb. 2008.

[5] J. Crisóstomo, P. Matafome, D. Santos-Silva, A. L. Gomes, M. Gomes, M. Patrício, L. Letra, A. B. Sarmento-Ribeiro, L. Santos, and R. Seiça, ‘‘Hyperresistinemia and metabolic dysregulation: A risky crosstalk in obese breast cancer,’’ *Endocrine*, vol. 53, no. 2, pp. 433–442 , Aug. 2016.

[6] J. A. Tice, S. R. Cummings, E. Ziv, and K. Kerlikowske, ‘‘Mammographic breast density and the gail model for breast cancer risk prediction in a screening population,’’ *Breast Cancer Res. Treatment*, vol. 94, no. 2 , pp. 115–122, Nov. 2005.

[7] Stavrinos,

S. Sampson, L. Fox, J. C. Sergeant, M. N. Harvie, M. Wilson, U. Beetles, S. Gadde, Y. Lim, A. Jain, S. Bundred, N. Barr, V. Reece, A. Howell, J. Cuzick, and D. G. R. Evans, ‘‘Mammographic density adds accuracy to both the tyrer-cuzick and gail breast cancer risk

ISSN 2347–3657
Volume 13, Issue 2, 2025

models in a prospective UK screening cohort,'' *Breast Cancer Res.*, vol. 17, no. 1, p. 147, Dec. 2015.

[8] Z. Wang, M. Li, H. Wang, H. Jiang, Y. Yao, H. Zhang, and J. Xin, ''Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features,'' *IEEE Access*, vol. 7, pp. 105146–105158, 2019.

[9] R. K. Ross, A. Paganini-Hill, P. C. Wan, and M. C. Pike, ''Effect of hormone replacement therapy on breast cancer risk: Estrogen versus estrogen plusprogestin,''*J.Nat.CancerInst.*,vol.92,no.4,pp. 328–332 ,Feb. 2000.

[10] J. Tyrer, S. W. Duffy, and J. Cuzick, ''A breast cancer prediction model incorporating familial and personal risk factors,'' *Statist. Med.*, vol. 23 , no. 7, pp. 1111–1130, Apr. 2004.

[11] A. Collins and I. Politopoulos, ''The genetics of breast cancer: Risk factors for disease,'' *Appl. Clin. Genet.*, vol. 4, pp. 11–19, Jan. 2011.

[12] Y. Zhang, R. Shi, C. Chen, M. Duan, S. Liu, Y. Ren, L. Huang, X. Dai, and F. Zhou, ''ELMO: An efficient logistic regression-based multi-omic integrated analysis method for breast cancer intrinsic subtypes,'' *IEEE Access*, vol. 8, pp. 5121–5130, 2020.

[13] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, ''A deep learning mammography-based model for improved breast cancer risk prediction,'' *Radiology*, vol. 292, no. 1, pp. 60–66, Jul. 2019.