Journal of Information Systems Engineering and Management
2025, 10(29s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ensemble Fusion of Classifiers with Kernel PCA for Breast Cancer Classification

Senthil Kumar T [1], Mardeni Roslee [2]*, Jayapradha [3], Chilakala Sudhamani [4], Abdullah Hadi Yahya Al-Quhali [5], Azlan bin Sulaiman [6]

[1, 2, 4, 5] Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
[1, 3] SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Tamil Nadu, India
[6] Telekom Malaysia Research & Development, Cyberjaya, Malaysia
** Corresponding Author's Email: mardeni.roslee@mmu.edu.my*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 19 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Breast cancer, predominantly affecting females, ranks as the most prevalent cancer among women globally, with potentially fatal consequences. Its invasive nature poses a significant health threat. Delayed diagnosis due to asymptomatic early stages hinders effective medical intervention. Early screenings prove pivotal in reducing breast cancer mortality. Beyond conventional diagnostic approaches, machine learning employs health data to predict breast cancer risk. This study employs Wisconsin breast cancer diagnosis data from the UCI machine learning repository. Class Imbalance is handled using AllKNN under-sampling technique and feature extraction using Kernel Principal Component Analysis (KPCA) were conducted. Logistic Regression, Support Vector Machine and Ensemble Learning (majority voting) are proposed for achieving high predictive accuracy.

Keywords: Breast Cancer, WBCD Dataset, Logistic Regression, SVM, Kernel PCA, Handling class imbalance.

INTRODUCTION

Breast cancer poses a significant threat to women's health, emphasizing the critical need for accurate and early diagnosis. Timely identification allows for effective treatment, ultimately reducing mortality rates associated with this disease. While traditional diagnostic methods involve pathology and imaging, the latter, particularly imaging diagnosis, has gained prominence for its non-invasive nature. However, relying solely on imaging may lead to delayed detection.

In this work, it is explored whether Fine Needle Aspiration biopsy (FNA), a minimally invasive pathological diagnostic method, facilitates improved accuracy and reduced number of false positives. In FNA, cells in the breast tumor are extracted and subjected to statistical analysis of cell attribute such as size, thickness, uniformity and smoothness. These data are then utilized to predict new cases, distinguishing between benign and malignant tumors.

LITERATURE SURVEY

One such prevalent approach in machine learning, which is the combination of Decision Tree and XGBooster and then a stacking classifier was taken by GK Kumar et al. [1] to enhance the prediction accuracy. In [2], Sohaib Asif et al. proposed a 5 1D Convolution Layers, 2 1D Max Pooling Layers, and 4 Fully Connected Layers with a simple 1D convolution neural network. Anyway the basis of their approach is the use of a hierarchical clustering technique on the decision trees in the Random Forest framework in work by DAQI CHEN et al. [3]. Specifically, Slamet Sudaryanto N et al.[4] employed ROS (random oversampling) and SMOTE (synthetic minority over sampling) algorithm of resampling and synthesis respectively. The constituting individual classification algorithms used are Random Forest, XGBoost, and AdaBoost. Second, classification results from the ensemble process were combined with majority voting, which selected the best result, and served as input into the XGBoost process to further improve the classification performance. Marion Olubunmi Adebiyi et al. [5] presents the procedure, that does use RF and SVM algorithms well known for their efficiency in classification tasks. In the other hand, the enhanced dataset is processed by LDA as a feature selection/mapping method targeting its discriminative features.

The methodology employed by Hua Chen et al. [6] focuses on leveraging machine learning algorithms, including Logistic Regression, XGBoost, Random Forest, and K-Nearest Neighbor, to establish various models for the classification and prediction of breast cancer. Feature selection through Pearson correlation and stratified sampling method is used for addressing imbalanced positive and negative samples. In Nurul Amirah Mashudi et al. [7] approach, Feature selection initiates the process, with the Gain Ratio method employed to assess and select attributes. The highest Gain Ratio values are identified, and attributes with ratios less than 0.1 are omitted, resulting in a reduced set of 23 attributes.

The study by Premkumar Duraisamy & associates In the paper [8], PCA is performed on incoming raw dataset to reduce its dimension, then with the reduced features achieved by applying PCA, different classification models like Logistic Regression,K-NN,SVM (Support Vector Machine), Decision Trees and Random Forest were tested out as well I used ensemble learning a.k.a Voting Classifier Praveen Sahu et al. [9] guest three numbers of feature selection methods, such as PCA, L1 Regularization (L1), Genetic Algorithm(GA) and Recursive Feature Elimination(RFE). They are able to generate different features dependency on their technique. The feature fitting is then applied to test Support Vector Classifier, Logistic Regression, Decision Tree & Random Forest models with the best class of features. The classification algorithms employed by Santad Promtan et al. [10] for model creation encompass the Decision Tree Classifier, Logistic Regression, KNeighbors Classifier and Naïve Bayes Classifier. Feature selection is done using the Random Forest Classifier. The research work by AHMED HAMZA OSMAN et al. [11] centers around the development of an integrated Radial Basis Function (RBF) neural network enhanced by ensemble features through the application of the boosting method.

The study by Asikur Rahman et al. [12] attempts to discover what configurations are most effective for each classifier by using hyperparameter tuning and the grid search method. In [13], Atiqur Rehman et al., worked to find the best feature extraction algorithm by comparing one of the above methods, Principle Component Analysis (PCA), Sparse PCA, Kernel PCA and Incremental PCA. Together with machine learning models, these algorithms are used to improve prediction of heart failure. These experimental findings show that integration of Kernel PCA along with a linear discriminant analysis (KPCA and LDA) model as well as Sparse PCA along with Gaussian Naïve Bayes (SpPCA and GNB) model achieves 91.11 % accuracy in HF classification.

Somya Goyal [14] observe an innovative Neighborhood based Under Sampling (N-US) algorithm to tackle with class imbalance problem. The purpose of this study is to demonstrate the high accuracy in defective module prediction achieved by the proposed N-US approach. Thus, the N-US algorithm strategically under samples the dataset with the goal of increasing minority data visibility while maintaining majority data visibility while decreasing information loss.

In this study [15] by M. Shyamala Devi et al., various under sampling methods, including Cluster Centroids, Repeated ENN, AllKNN, Edited Nearest Neighbors, Condensed Nearest Neighbor, Instance Hardness Threshold, and Near Miss are applied to the Cardiotocographic Fetal Heart Rate dataset. The resulting under sampled datasets from these methods are then subjected to fitting all classifiers, and their performance is scrutinized both before and after feature scaling. In the final step, a thorough performance analysis is conducted, employing metrics such as Precision, Recall, F-score, Accuracy, and running time. The experimental findings demonstrate that the Decision Tree classifier maintains a 98% accuracy both before and after feature scaling, particularly when incorporating under sampling techniques like Edited Nearest Neighbors, Repeated ENN, and AllKNN. In this study [16], Al-Zadid Sultan Bin Habib et al. applied the ensemble learning-based Hard Voting (HV) technique combined with random under-sampling to WDBC dataset. After scaling these features using Robust Scaler, random under-sampling was employed to balance the classes (benign or malignant). Four prominent machine learning classifiers, namely Decision Tree Classifier (DTC), Random Forest (RF), k-Nearest Neighbor (KNN), and Support Vector Machine (SVM), were utilized to create an HV meta-classifier.

The survey [17] by Avneesh Atrey et al. employs predictive analysis with Support Vector Machines (SVM) to distinguish between malignant and benign cases. Various SVM kernels, including linear, polynomial, and RBF are applied, and their corresponding results are computed. The findings indicate that the RBF kernel yields superior results, achieving an accuracy of 93.9%, F1-score of 89.5%, precision of 91.5%, and recall of 89.3%. The survey [18] by Zohaib Mushtaq et al. utilizes chi-2 feature selection with Gaussian Naive Bayes, accompanied by Kernel-based PCA to reduce the dimensional space of six selected features. Various kernel functions such as sigmoid, poly, linear, cosine, and radial basis are applied. The successful classification of the WBC dataset involves employing the sigmoid function on K-PCA in tandem with chi-square feature selection.

In this study, our research emphasizes the application of classification algorithms to predict early breast cancer.

The principal contributions of this work are:

1. To handle the class imbalance problem, we have incorporated under-sampling using the AllKNN technique. The primary purpose of under-sampling in the context of imbalanced datasets is to address the issue of class imbalance by reducing the number of instances in the majority class.
2. Robust Scaling for normalization to deal with outliers in the dataset.
3. Pre-processing of data was performed followed by feature extraction of data set using Kernel Principal Component Analysis (KPCA).
4. A custom ensemble model (Voting Classifier) is instantiated with two classifiers (logistic regression and Support Vector Machine) and outputs the class label that receives the majority of the votes. Grid search cross-validation method has been used for best hyper-parameters selection

PROPOSED METHODOLOGY

A. Dataset Overview

Input dataset is collected from WBCD Library. The Wisconsin Diagnostic Breast Cancer dataset is a well-known dataset in the field of machine learning and is often used for classification tasks. The dataset is derived from fine-needle aspirate (FNA) biopsies and contains features computed from digitized images of breast cancer masses. The primary objective of using this dataset is typically to distinguish between benign and malignant tumors based on the provided features.

The dataset is publicly available and can be accessed from various machine learning repositories, such as the UCI Machine Learning Repository. There are 569 occurrences with 30 features (attributes) for each instance in the dataset.

B. Architecture

The proposed architecture involves data loading, preprocessing, model training, hyperparameter tuning, and ensemble techniques, ultimately providing a comprehensive approach to breast cancer classification. Fig. 1 shows our proposed architecture flow chart. The use of various algorithms and ensemble methods enhances the model's ability to accurately classify benign and malignant cases. The thorough evaluation and comparison contribute to understanding the strengths and limitations of each model in the context of breast cancer classification.

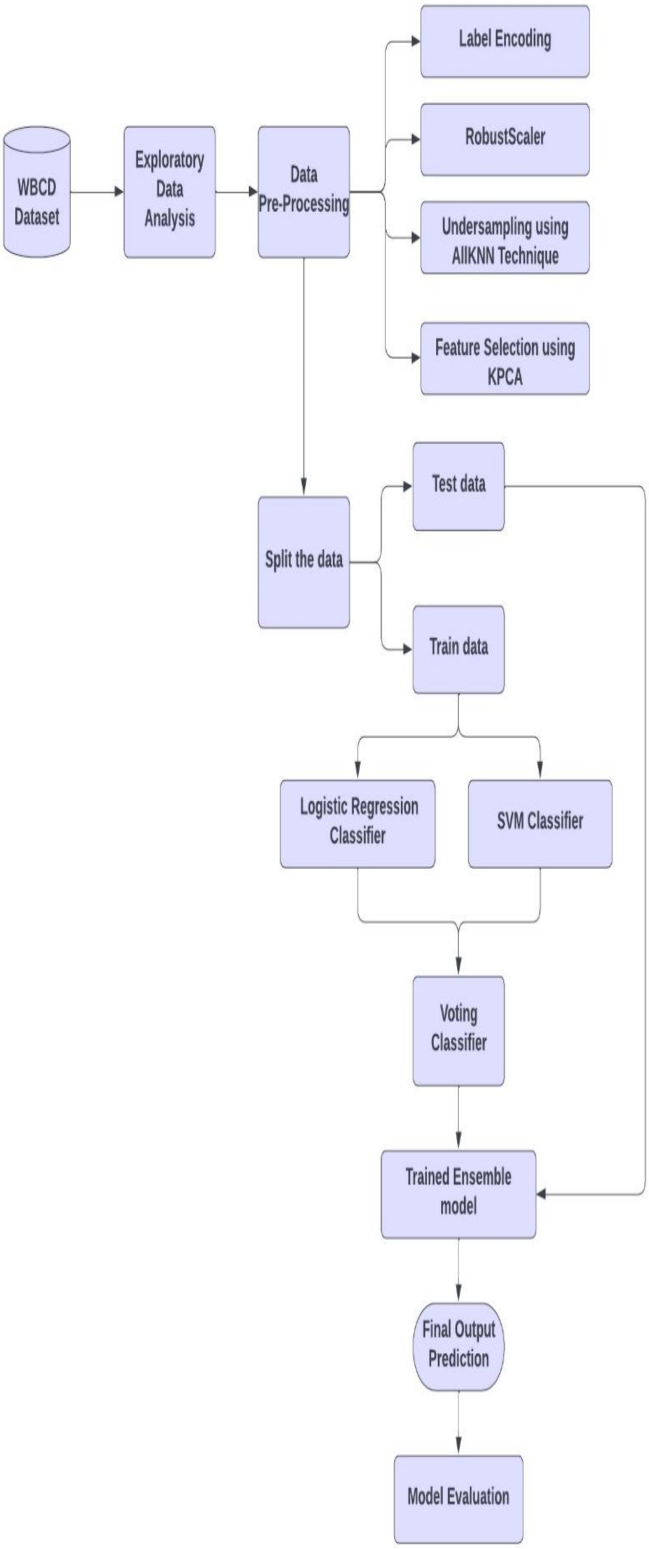


Fig. 1. Proposed Architecture Flow Chart

C. Exploratory Data Analysis

Exploratory analysis of the dataset including information on data types, summary statistics, and identification of missing values is performed. The graph in the Fig. 2 illustrates a class imbalance in the dataset, showcasing a predominant number of benign diagnoses with a count of 357, compared to 212 malignant diagnoses. This disparity underscores the higher proportion of benign instances within the sample.

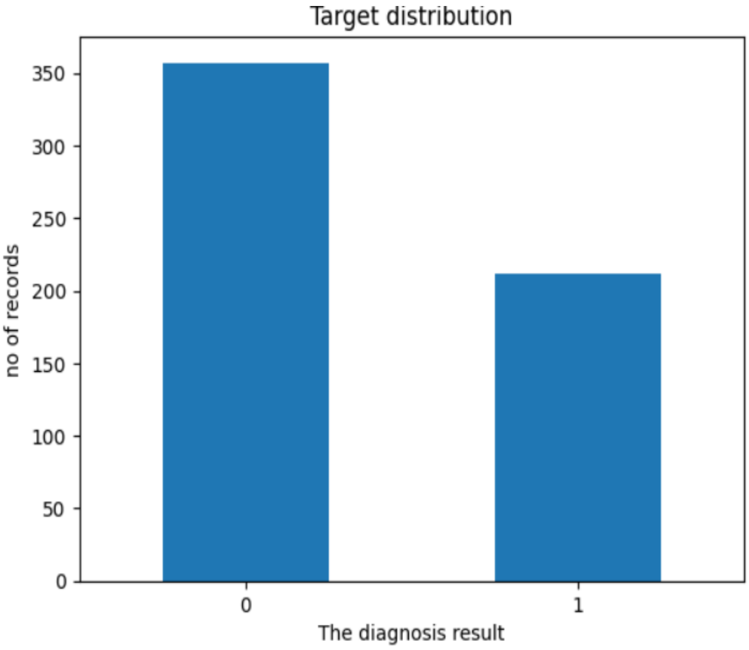


Fig. 2. Class Imbalance

Pearson Correlation Coefficient:

As a statistical measure, the Pearson correlation coefficient estimates the extend of the linear relationship between two variables.

It takes values between -1 and 1, where:

- Positive values near 1 mean a very strong positive linear dependency which means that one property directly depends on another.
- At the lowest level we get -1, which means there is a strong negative linear relationship meaning that, as one feature increases, the other tends to decrease.
- Values equaled or close to 0 indicate the absence or less than a direct relationship between the features.

Data Preprocessing

Handling missing values in the machine learning algorithm involve erasing nonrelevant features from the data set such as the 'Unnamed 32' and the 'ID'. Label encoding is used to transform categorical features into numeric ones The target variable shows 0 for benign tumors and 1 for malignant ones. Furthermore, the Robust Scaler is used for feature scaling where median and IQR are used to contain the impact of outliers and carefully balance the numerical features.

Handling Class Imbalance:

ALLKNN (All K-Nearest Neighbors):

Under-sampling technique is available in the form of AllKNN method and belongs to the imbalanced-learn (imblearn) package. How it does is by reducing the number of majorities using a similarity of the feature space where the majorities are at least almost close to the minorities. What is more, the parameter n_neighbors defines how many neighbors are taken into account while defining the majority class instances to be removed.

OPTIMAL VALUE OF K:

The value of k is very sensitive in the KNN algorithm on the number of neighbors it is going to honor in the algorithm. When analyzing the k-NN algorithm the k parameter has to be optimized depending on the input data. If the input data has more volumes of outliers or noise then, if has to prefer higher value of k. However, since in this classification the value of ties is not permissible, it would be better to choose an odd value for the variable k.

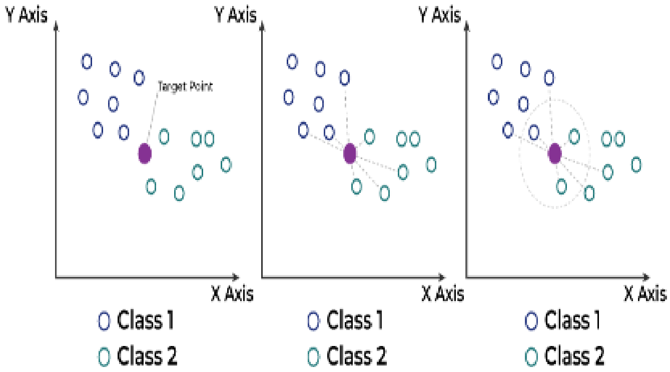


Fig. 3. Finding K-nearest neighbors

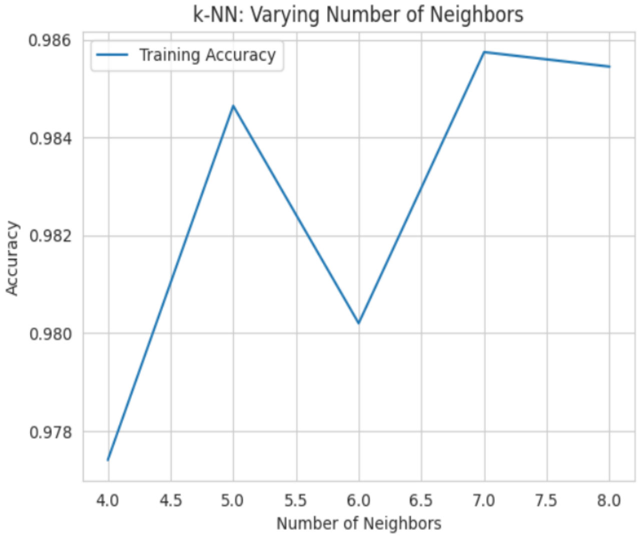


Fig. 3.1. Accuracy based on varying K value

How it Works:

- For each majority class instance, the algorithm calculates its distance to its k nearest neighbors as shown in Fig. 3.
- If a majority class instance has at least one neighbor from the minority class within its k nearest neighbors, it is considered as potentially contributing to the imbalance. Such instances are then removed to achieve a more balanced class distribution. Fig. 3.1. & Fig. 3.2. shows that optimal K-value is 7.

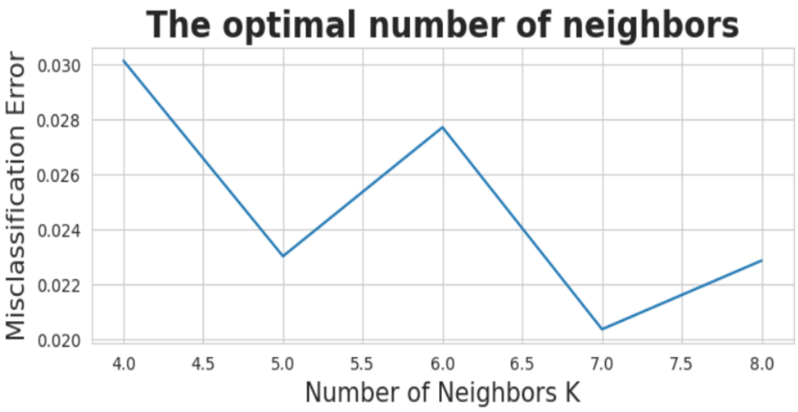


Fig. 3.2. MSE based on varying K value

Feature selection using Kernel PCA:

Kernel PCA – it is an improvement over traditional PCA in which kernel functions are used to transform the input data in higher dimensions as depicted by Fig 4, then the nonlinear relationship is learned. The beneficial aspect of using Kernel PCA is to diminish the number of features while looking at the interactions between features and the data set nonlinearity. Self-organization to a larger degree supports an efficient representation of data especially when it comes to the more complicated non-linear patterns.

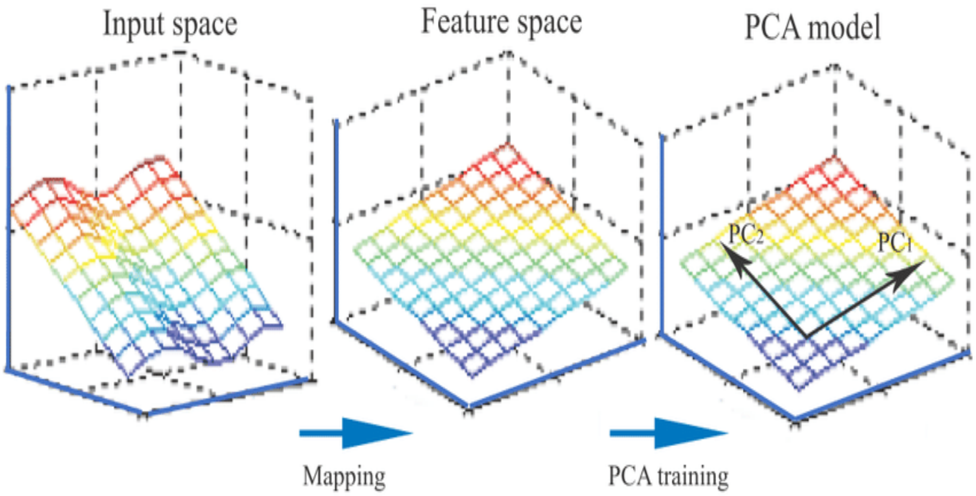


Fig. 4. Mapping Input data into Higher dimensional space

Steps in Feature Selection using Kernel PCA:

- Mapping to High-Dimensional Space:

Apply a chosen kernel function (e.g., RBF kernel) to map the original feature space into a higher-dimensional space.

- Principal Component Analysis:

Perform PCA in the high-dimensional space to identify principal components. These components capture the most significant variations in the data.

- Selecting Principal Components:

Based on the explained variance or other criteria, choose a subset of principal components that effectively represent the data.

- Inverse Transformation:

Apply the inverse transformation to map the selected principal components back to the original feature space.

D. Training and Testing

Training can be described as model construction using a dataset for enhancing and developing a machine learning model. Afterwards the model is tested to check the performance on data that the model has not seen before and whether the model is capable of providing correct prediction for unseen data. It is standard practice to use 80% of the data set to build the model and the rest 20% for testing the model.

E. Algorithm

Logistic Regression is one of the most used statistical methods of binary classification, where the outcome variable can take the value of only one of the two possible values. The main element of this technique is the logistic function (also called a sigmoid function) mapping a real-value input into a probability value of 0 to 1. The mathematical representation of the sigmoid function is:

$$\sigma(y) = \frac{1}{1+e^{-y}} \text{-----------} \qquad (1)$$

(y) here stands for linear combination of independent variables and their respective coefficients. The sigmoid function's output is taken as probability of the positive class (class one). One is to have a threshold value, typically 0.5; if the predicted probability is greater than this threshold value, the observation is then classified as class 1; otherwise it is classified as class 0.

Support Vector Machine: SVM, for short, Support Vector Machine, is also a popular supervised machine learning algorithm which is used for classification and regression models. It is categorized among discriminative models and it performs especially well in higher dimensions. A fundamental max aim of an SVM is the identification of an ideal hyperplane that will aid in the classification of the data points of distinct categories in the feature space. In the case of binary classification, this hyperplane brings the largest separation between the two classes.

Voting Classifier: Voting Classifier is an estimator which includes models for the various classifiers of classification techniques that are looked at as having individual confidence levels. The meta-classification approach of constructing the Voting classifier estimator from a group of various classification classifiers results in the stronger classifiers that compensate the others' disadvantageous results on the given dataset. Voting classifier uses majority voting depending on the applied weights on the class or the class probabilities and a record is classified based on the majority vote.

In our proposed methodology, Voting Classifier integrates Logistic Regression and Support Vector Machine into a soft-voting ensemble.

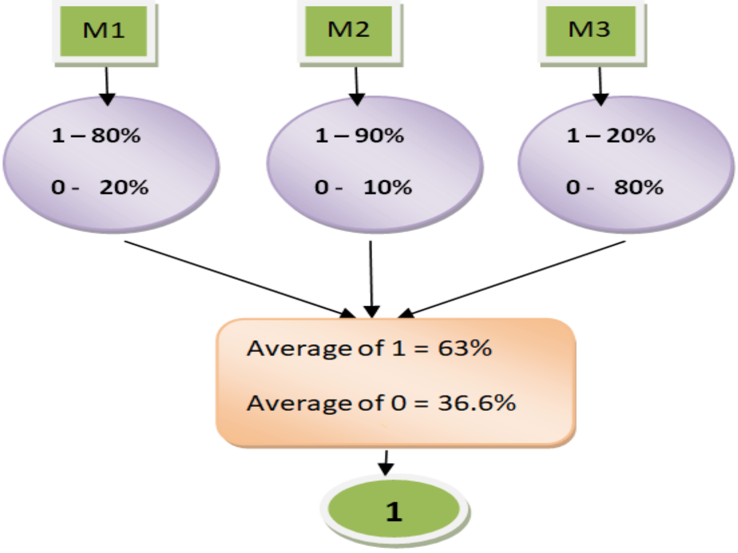


Fig. 5. Illustration of sample model prediction using Soft Voting Classifier

Soft voting classifier classifies input data based on the probability of all the classifiers as done by different classifiers as shown in Fig. 5.

EXPERIMENTAL RESULTS

In this section, we delve into the outcomes of our study and introduce the foundational models that serve as benchmarks for evaluating our proposed model. Our assessment of the proposed model's performance involves a comprehensive comparison with diverse machine learning classifiers and alternative methodologies.

A. Accuracy:

Accuracy is calculated as Total of True Positives plus True Negatives over the sum of True Positives plus True Negatives plus False Positives plus False Negatives.

Comparison of Test data Accuracy of various machine learning models "Without using any sampling techniques/Feature selection with default parameters" and "Using AllKNN under-sampling technique and KPCA Feature selection" is shown in Table 2.

Table 2. Comparison of Test data Accuracy

Comparison of Test Accuracy		
Model	**Without Under-Sampling Technique**	**With Under-Sampling Technique**
Logistic Regression	98.24	98.97
SVM	96.49	98.97
Voting Classifier (LR + SVM)	**97.36**	**100.00**
Gradient boost	95.61	97.95
K-Nearest Neighbors	96.49	94.89

The model put forward in this study demonstrated remarkable precision by achieving a 100% accuracy in predictions on the test set. This noteworthy accuracy level serves as a testament to the model's efficacy in accurately classifying the diagnoses within the dataset.

The Training data, Test data and All data Accuracy results of our comparative analysis, encompassing various machine learning classifiers using AllKNN under-sampling technique and KPCA Feature selection, are meticulously presented in below Table 3.

Comparing all the different machine learning models, we could see that our proposed model – Voting Classifier (Logistic Regression + Support Vector Machine) outperformed other models on using Training data, Test data and All data.

Table 3. Training, Test and All data Accuracy

After Under-Sampling using AllKNN and KPCA Feature selection			
Model	**Training Accuracy**	**Test Accuracy**	**All data Accuracy**
Logistic Regression	98.45	98.97	98.55
SVM	98.71	98.97	98.76
Voting Classifier (LR + SVM)	**98.45**	**100.00**	**98.76**
K-NN	98.45	97.95	98.35
Gradient boost	99.22	94.89	98.35

B. Confusion Matrix:

To assess the predictive performance of a machine learning (ML) model on a given dataset, a confusion matrix is used. This matrix organizes the model's outcomes into four categories: true positives, true negatives, false positives, and false negatives. These categories help illustrate the model's accuracy and errors in its predictions.

The Confusion Matrix for Logistic Regression model using AllKNN under-sampling technique and KPCA Feature selection with best hyper parameters using Grid Search is shown below in Fig. 6.

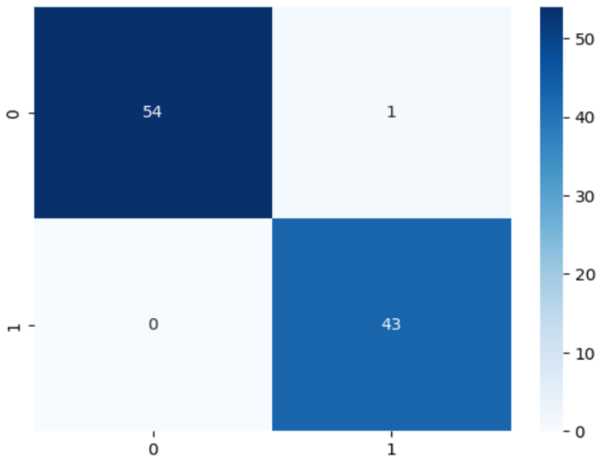


Fig. 6. Confusion Matrix for Logistic Regression

The Confusion Matrix for Support Vector Machine Classifier model using AllKNN under-sampling technique and KPCA Feature selection with best hyper parameters using Grid Search is shown below in Fig. 7.

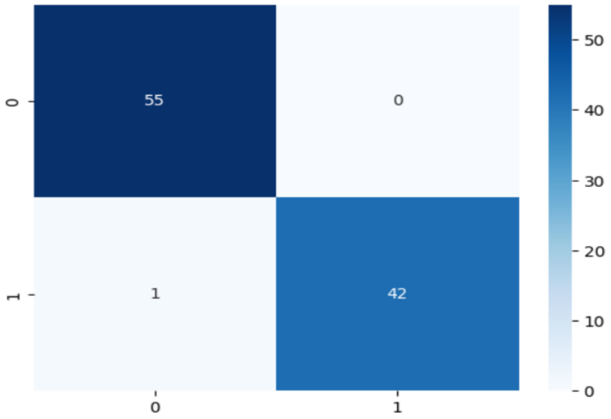


Fig. 7. Confusion Matrix for Support Vector Machine

The Confusion Matrix for the proposed Voting Classifier model is shown below in Fig. 8.

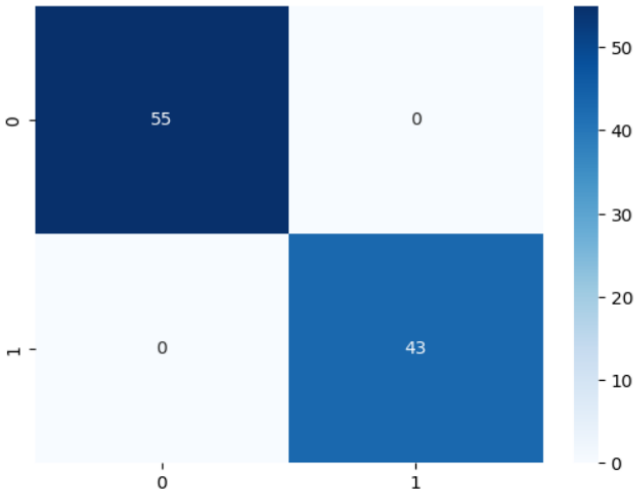


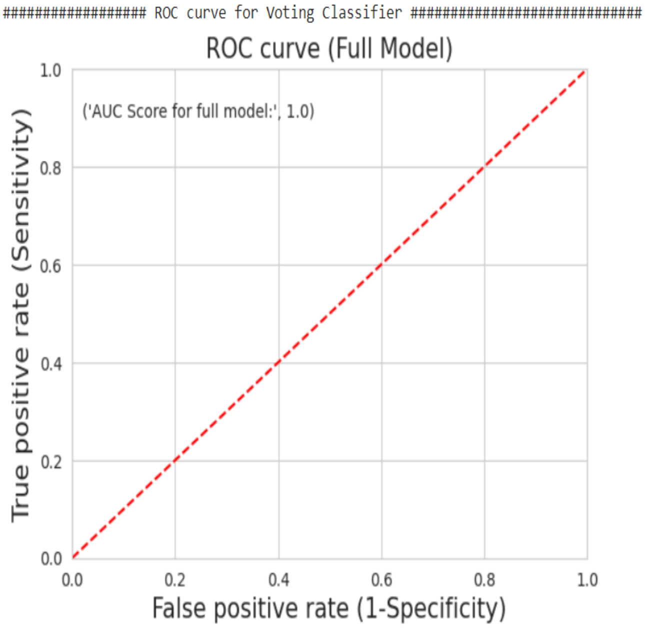
Fig. 8. Confusion Matrix for Voting Classifier

For our proposed Voting Classifier model, the confusion matrix highlights that the model successfully identified 55 true positives, affirming its adeptness in anticipating the malignant nature of the actual diagnoses. Additionally, the model correctly identified 43 true negatives, aligning with its expectations of benign diagnoses.

A. ROC Curve:

ROC curve is obtained by plotting of TPR against FPR for multiple threshold levels of the feature measurements. A model with the best possible assessment would be represented by an ROC plot that starts from the point of TPR = 1 and FPR = 0 and goes down as far as possible to the lower right corner of the plot.

The ROC for our proposed Voting classifier model is shown below.



```
################## ROC curve for Voting Classifier #############################
```

Fig. 9. ROC for Voting Classifier

CONCLUSION

Our primary focus involved a comparative analysis of accuracy (Acc) when employing classifier models—KNN, Gradient Boost, LR, SVM, and the voting ensemble model—with and without the application of AllKNN Under-sampling and KPCA Feature selection techniques. We observed an increase in performance when AllKNN under-sampling technique and KPCA Feature selection is used as compared to without sampling/feature selection technique. Key findings from our investigation include the following: Prior to integrating AllKNN Under-sampling and KPCA Feature selection techniques, LR exhibited a performance level of Acc = 98.24%. Upon implementing these techniques, overall performance improved significantly, resulting in LR achieving Acc = 98.97% with 12 principal components. Similarly, SVM demonstrated an accuracy of 96.49% before applying AllKNN Under-sampling and KPCA Feature selection, which then increased to Acc = 98.97% with the inclusion of 20 principal components.

Subsequently, employing our proposed model through the voting process and the ensemble model yielded a test accuracy of 100% and an overall data accuracy of 98.76%. The outcomes of our tests clearly demonstrated that utilizing the ensemble model with AllKNN Under-sampling and KPCA Feature selection techniques markedly enhanced the precision of each classifier.

Acknowledgement:

This work was supported and funded by a Telekom Malaysia Research & Development (TMR&D) grant, RDTC/241138, MMUE/240095.

REFERENCES

[1] KBN Tara, N Vadlamudi, GK Kumar. "A Stacked Ensemble-Based model for the prediction of breast cancer using Decision Tree and XGBoost." World Conference on Communication & Computing (WCONF), 2023.

[2] Sohaib Asif, Yi Wenhui, Si Jinhai, Yi Tao, Zafran Waheed, Kamran Amjad. "Novel One-Dimensional Convolutional Neural Network for Breast Cancer Classification". 2021 7th International Conference on Computer and Communications.

[3] Zexian Huang And Daqi Chen. "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm." IEEE Access, 2021. Digital Object Identifier 10.1109/ ACCESS.2021.3139595

[4] Slamet Sudaryanto N, Mauridhi Hery Purnomo, Diana Purwitasari, Eko Mulyanto Yuniarno. "Synthesis Ensemble Oversampling and Ensemble Tree-Based Machine Learning for Class Imbalance Problem in Breast Cancer Diagnosis." 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)

[5] Marion Olubunmi Adebiyi, Micheal Olaolu Arowolo, Moses Damilola Mshelia and Oludayo O. Olugbara. "A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis." Appl. Sci. 2022, 12, 11455. https://doi.org/10.3390/ app122211455

[6] Hua Chen, Nan Wang, Xueping Du, Kehui Mei, Yuan Zhou, and Guangxing Cai."Classification Prediction of Breast Cancer Based on Machine Learning." Computational Intelligence and Neuroscience, 2023

[7] Nurul Amirah Mashudi, Syaidathul Amaleena Rossli, Norulhusna Ahmad, Norliza Mohd Noor."Comparison on Some Machine Learning Techniques in Breast Cancer Classification." 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)

[8] Premkumar Duraisamy, Yuvaraj Natarajan, Ebin N L, Jawahar Raja P."A Comprehensive Comparison of Machine Learning Algorithms for Breast Cancer Prediction." 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)

[9] Praveen Sahu, Pragatheiswar Giri, Raja Sunkara, Raji Sundararajan."Extraction of Key Features and Enhanced Prediction Framework of Breast Cancer Occurrence." Sixth International Conference on Trends in Electronics and Informatics (ICOEI 2022)

[10] Santad Promtan, Phungern Khongthong, Chidchanok Choksuchat."Breast Cancer Prediction of Benign and Malignant Tumors by Classification Algorithms." 2023 4th International Conference on Big Data Analytics and Practices (IBDAP)

[11] Ahmed Hamza Osman And Hani Moetque Abdullah Aljahdali."An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model." IEEE Access Digital Object Identifier 10.1109/ACCESS.2020.2976149

[12] Md. Mijanur Rahman, Asikur Rahman, Swarnali Akter, Sumiea Akter Pinky."Hyperparameter Tuning Based Machine Learning Classifier for Breast Cancer Prediction." Journal of Computer and Communications, 2023, 11, 149-165

[13] Atiqur Rehman, Aurangzeb Khan, Muhammad Akhtar Ali, Muhammad Umair Khan, Shafqat Ullah Khan, Liaqat Ali, "Performance Analysis of PCA, Sparse PCA, Kernel PCA and Incremental PCA Algorithms for Heart Failure Prediction", Proc. of the 2nd International Conference on Electrical, Communication and Computer Engineering (ICECCE) 12-13 June 2020, Istanbul, Turkey.

[14] Somya Goyal, "Handling Class-Imbalance with KNN (Neighbourhood) Under-Sampling for Software Defect Prediction", Artifcial Intelligence Review (2022) 55:2023–2064. https://doi.org/10.1007/s10462-021-10044-w

[15] M. Shyamala Devi, S. Sridevi, D.Umanandhini, A. Peter Soosai Anandaraj, Sudheer Kumar Gupta, Bhumireddy Sidhartha, "Undersampling Aware Learning based Fetal Health Prediction using Cardiotocographic Data", Turkish Online Journal of Qualitative Inquiry (TOJQI) Volume 12, Issue 6, July, 2021: 7730-7749

[16] Al-Zadid Sultan Bin Habib, Kazi Tanvir Islam, Md Munimul Hasan Pranto, Mohammad Nooruddin, "Breast Cancer Classification Using Ensemble Hard Voting with Random Under-Sampling", 2020 11th International Conference on Electrical and Computer Engineering (ICECE)

[17] Avneesh Atrey, Neetu Narayan, Surbhi Vijh, Sumit Kumar, "Analysis of Breast Cancer using Machine Learning Methods", 12th International Conference on Cloud Computing, Data Science and Engineering (Confluence 2022)

[18] Zohaib Mushtaq, Muhammad Farrukh Qureshi, Muhammad Jamshed Abbass and Sadeq Mohammed Qaid Al-Fakih, "Effective kernel-principal component analysis based approach for wisconsin breast cancer diagnosis", ELECTRONICS LETTERS January 2023 Vol. 59 No. 2 wileyonlinelibrary.com/iet-el