UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Breast Cancer Prediction Using Support Vector Machine Ensemble with PCA Feature Selection Method

Nurul Hidayah Parman[1], Noor Hidayah Zakaria[2] & Rohayanti Hassan[3]
Faculty of Computing,
Universiti Teknologi Malaysia
Johor Bahru, Johor, Malaysia
Email: nurulhidayah.parman@graduate.utm.my[1]; noorhidayah.z@utm.my[2]; rohayanti@utm.my[3]

Submitted: 17/1/2024. Revised edition: 28/3/2024. Accepted: 12/5/2024. Published online: 31/5/2024
DOI: https://doi.org/10.11113/ijic.v14n1.461

***Abstract*—Breast cancer is the most prevalent cancer among women worldwide and ranks second in cancer-related mortality, comprising 11.6 percent of all cancer cases. Given that survival outcomes are largely contingent on the stage at which the disease is detected, early detection plays a pivotal role in securing the best prognosis for patients. Machine Learning algorithms are increasingly employed in breast cancer diagnosis due to their accuracy and capacity to anticipate the likelihood of recurrence. In this research, Support Vector Machine (SVM) was chosen as one of the classifiers, recognized for its precise predictive capabilities in cancer prediction and prognosis. To enhance model accuracy and mitigate variance, Principal Component Analysis (PCA) feature selection was incorporated into the study. Among the methods explored, the boosting ensemble method utilizing SVM as the base classifier demonstrated superior performance in breast cancer prediction. SVM as the base classifier with boosting ensemble method has outperformed the SVM models by increasing the accuracy value from 94% to 96% with a precision and recall score of 97%. Consequently, this research contributes to the advancement of patient diagnosis by implementing a classification algorithm tailored for breast cancer prediction.**

Keywords—Breast Cancer, Support Vector Machine, AdaBoost, Principal Component Analysis, Min-Max Scaling

I. INTRODUCTION

Breast cancer is a significant global health concern. In 2018 alone, it affected about 2.1 million new cases and led to approximately 626,679 deaths, making this disease as a leading cause of female mortality [1]. Shockingly, a new case was diagnosed every 18 seconds. Risk factors for breast cancer include age, early onset of menstruation (before 12 years old), late menopause, giving birth at an older age (above 30 years), lack of breastfeeding, a history of benign and malignant breast conditions, family history of breast cancer, extended use of hormone therapy or contraceptive pills, alcohol consumption, physical inactivity, radiation exposure, smoking, and genetic predisposition [2].To identify breast cancer in women, Machine Learning (ML) algorithms are increasingly being used because of their precision and ability to forecast the likelihood of recurrence [3]. Study by Lou *et al*. [4] finds that the comparison of AUROC values indicated that the ANN model is superior to other prediction models. Another study by [5] compared between five nonlinear machine learning algorithms viz Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Nave Bayes (NB) and Support Vector Machines (SVM) on the Wisconsin Breast Cancer Diagnostic (WBCD) dataset is conducted. This study found that the accuracy of MLP on the training data is 96.70% which is better than the other four algorithms. On the other hand, study by [6] has assessed both the effectiveness of a predictive model and the significant factors that impact the survival rates of breast cancer patients. These findings have practical implications, especially in an Asian healthcare context. The study employed decision trees and survival curves to validate the crucial variables affecting breast cancer survival rates. If breast cancer is detected early, it can be treated rapidly and with less risk, resulting in a 25% decrease in mortality rates [7]. According to the research by [8], SVM is one of the classifiers that is commonly utilized in cancer diagnosis or prognosis due to its accurate predictive performance. However, the disadvantages of SVM from prior research, which include the fact that it is time consuming for massive amounts of data, therefore approximate approaches have been employed to reduce computation time. However, it worsens classification performance. To cope with the drawback, this research proposed SVM ensembles.

To improve the accuracy of the classifier model, an ensemble method termed boosting ensemble method (AdaBoost) with SVM as the base classifier (Boosted SVM) has been introduced. Boosting is a sequential ensemble strategy in which each successive model attempts to fix the errors of the prior model. When a weak model's base classifier is misclassified, its weight is increased, and the next base learner classifies it more accurately [9]. The goal of boosting is to reduce bias, which is the difference between the model's predicted and expected values [10]. To minimize overfitting in the data set and remove outliers, a scaling method known as min-max scaling is used. Then, a feature selection technique known as Principal Component Analysis (PCA) is implemented to reduce the number of parameters that describe the dataset, generate significant amounts of information in the absence of some parameters, and increase model accuracy.

II. MATERIALS AND METHODS

In this research, there are three totals of phase involved included the first phase of this study focuses on a literature review and studies of existing research on classifying the cancerous and non-cancerous cell of breast cancer, and the second phases focus on designing, developing, and carrying out the experiment and the third phase focuses on the presentation, analysis, and discussion on the findings. Fig. 1 shows the framework of the research.

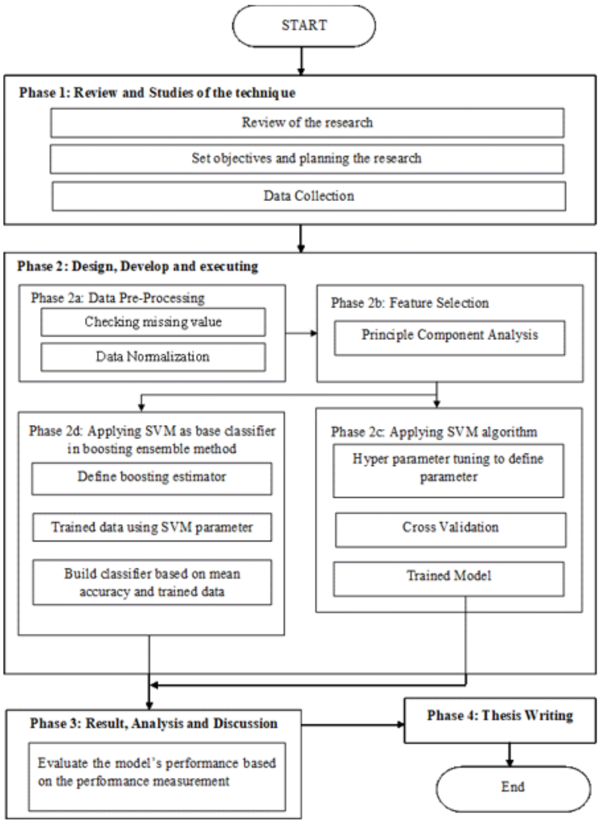

START
Phase 1: Review and Studies of the technique
Review of the research
Set objectives and planning the research
Data Collection
Phase 2: Design, Develop and executing
Phase 2a: Data Pre-Processing
Checking missing value
Data Normalization
Phase 2b: Feature Selection
Principle Component Analysis
Phase 2d: Applying SVM as base classifier in boosting ensemble method
Define boosting estimator
Trained data using SVM parameter
Build classifier based on mean accuracy and trained data
Phase 2c: Applying SVM algorithm
Hyper parameter tuning to define parameter
Cross Validation
Trained Model
Phase 3: Result, Analysis and Discussion
Evaluate the model's performance based on the performance measurement
Phase 4: Thesis Writing
End

Fig.1. Research Framework

A. Data

The dataset used for this paper has been taken from the Wisconsin Breast Cancer Data from the UCI Machine Learning Repository website [11]. This dataset has 569 instances and 32 attributes. There are over ten features that describe the nuclei in the cell whose digital image by fine needle aspirate is taken to classify the malignant and benign cells. The ten attributes that were calculated for the mean, standard error, and worst resulted in a total of 30 attributes in the dataset. The ten attributes are:

i. Area
ii. Compactness: (p*p/a-1), where p is the perimeter while a is the portions of the contour
iii. Concavity
iv. Fractal dimension; approximation of the coastline.
v. Perimeter
vi. Radius; mean of the distance between center and perimeter.
vii. Smoothness: local disparity in the length of the radius
viii. Symmetry
ix. Texture

B. Normalization

Normalization is used to overcome overfitting as well as to suppress outliers. It can be avoided by restricting the absolute value of the model's parameters. This can be accomplished by including a penalty term in the cost function depending on the magnitude of the model parameters. This scaling strategy limits the characteristics to a maximum and a minimum value, which is usually between zero and one.

C. PCA Feature Selection

PCA is an unsupervised machine learning technique used to identify patterns and highlight similarities and differences in high-dimensional data. To prepare the data for PCA feature extraction, it's essential to normalize it to reduce variance since PCA heavily depends on high-variance variables. Failure to do so may introduce bias into the analysis. The first step involves computing the covariance matrix of the dataset, which quantifies how two variables change together. In the context of PCA, this matrix reveals relationships between the original features. Subsequently, eigenvalues and their corresponding eigenvectors are computed based on the covariance matrix. Eigenvalues are coefficients calculated on eigenvectors, influencing their length or magnitude. These eigenvalues are then sorted in descending order, retaining only the largest ones and discarding the rest. This step constitutes dimensionality reduction by eliminating values that explain minimal variance in the data. The final output is a new feature vector containing the principal components of the dataset. These components capture the maximum information from the original features while maintaining their independence. This transformed dataset is suitable for use in various machine learning models and classifiers.

D. Model

In this study, the Support Vector Machine (SVM) and Boosting ensemble method were employed as the primary models. SVM is a machine learning technique that leverages hyperplanes to classify datasets into two classes, namely benign and malignant. The implementation was carried out using the Python Scikit-Learn library. The SVM classifier was evaluated by testing various parameters, including:

i. C; the value after hyperparameter tuning is C=0.1

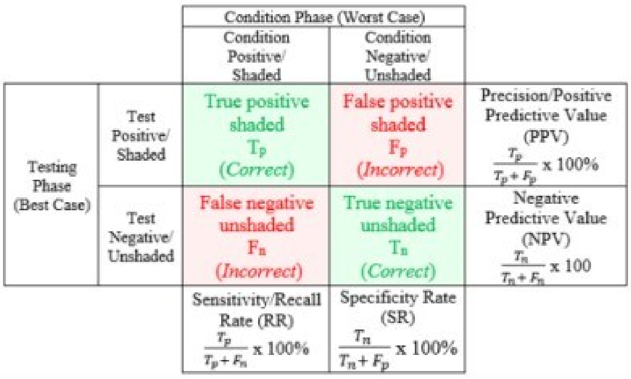
ii. Gamma; the coefficient of the kernel and the fixed value after hyperparameter tuning is 5.

iii. Kernel; it can be 'poly','rbf',' sigmoid', or 'linear' and after testing the skills of the kernel for the dataset, the linear kernel is the chosen kernel.

AdaBoost is a method for calculating output using many models and then averaging the results using a weighted average approach. By combining the benefits and drawbacks of these approaches and modifying them, a good prediction for a wide variety of input data can be produced [8].

E. Performance Measurement

This study assessed the predictive performance of the model using metrics such as accuracy, precision, and recall scores, which are derived from the confusion matrix generated after the classifier model's predictions. Accuracy measures the frequency of correct predictions, calculated by dividing the sum of true negatives (TN) and true positives (TP) by the total number of predictions. Precision evaluates the correctness of positive predictions, indicating how many of the positive predictions were accurate out of all positive predictions made. Precision is obtained by dividing TP by the sum of TP and false positives (FP). Recall measures the ratio of correctly classified positive instances to the total number of actual positive instances. It is calculated by dividing TP by the sum of TP and false negatives (FN).



		Condition Phase (Worst Case)		
		Condition Positive/ Shaded	Condition Negative/ Unshaded	
Testing Phase (Best Case)	Test Positive/ Shaded	True positive shaded T_p (Correct)	False positive shaded F_p (Incorrect)	Precision/Positive Predictive Value (PPV) $\frac{T_p}{T_p+F_p}$ x 100%
	Test Negative/ Unshaded	False negative unshaded F_n (Incorrect)	True negative unshaded T_n (Correct)	Negative Predictive Value (NPV) $\frac{T_n}{T_n+F_n}$ x 100
		Sensitivity/Recall Rate (RR) $\frac{T_p}{T_p+F_n}$ x 100%	Specificity Rate (SR) $\frac{T_n}{T_n+F_p}$ x 100%	

Fig. 2. Description of the confusion matrix (source: Gyamfi and Missah (2017))

III. RESULTS AND DISCUSSION

A. Principal Component Analysis

After applying the PCA in the dataset, the dimension of the data is reduced by creating two new attributes namely principal component 1 (legend M) and principal component 2 (legend B). The overview of the dataset is shown in Fig. 3.

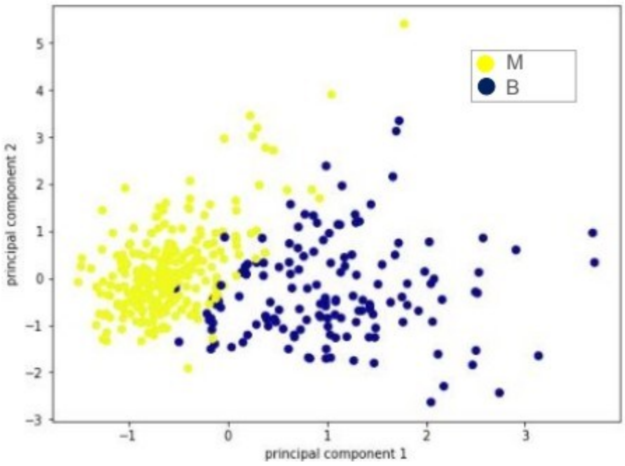


Fig. 3. The distribution of the data after PCA

From the figure above, the diagonal line between the two classes was clearly defined which means that the classifier would easily differentiate the malignant and benign classes. This proves that the dimension of the data has been reduced by the PCA method. To be precise, the data after the PCA process will be implemented for the next step which is boosting the ensemble method using SVM as the base classifier to know the efficiency of the ensemble method for this dataset.

B. Kernel Parameter Selection

During the kernel selection process, the default values of gamma and C were consistently used for each experiment. Notably, the sigmoid kernel exhibited a substantial change in performance before and after applying PCA feature selection, with the accuracy score of the classifier model increasing from 0.29 to 0.89. This significant improvement suggests that the Sigmoid kernel was ill-suited to the dataset's type and distribution without PCA feature selection. According to a study by Rimah *et al.* [12], the sigmoid kernel may not be positive or semi-definite for specific parameter values, potentially leading to incorrect results and negatively impacting classifier performance. This discrepancy in results could be attributed to the Sigmoid kernel's improper parameter selection, which may have occurred because the dataset without PCA feature selection had significantly higher dimensionality compared to the dataset with PCA feature selection. Consequently, further experimentation is warranted to assess the Sigmoid kernel's suitability under different conditions of the classifier model.

Additionally, it's worth noting that the accuracy of the model was higher before applying PCA feature selection for the RBF and polynomial kernels. This suggests that feature selection had an insignificant impact on these kernels, potentially contradicting the research objective. Consequently, the RBF and polynomial kernels will be excluded from the list of candidate kernels. As a result, the liner kernel will be chosen as the kernel for the following experiment because it shows a significant increment after implementing PCA feature selection. To be precise, the graph in Fig. 5 shows the difference between linear, RBF, and polynomial kernels.

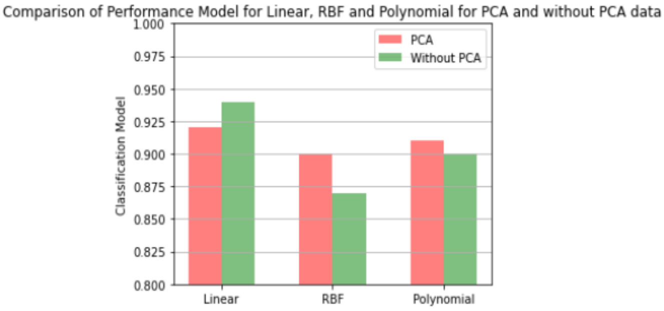


Fig. 5. Comparison of Linear, RBF, and Polynomial kernel with and without PCA

C. *Boosting Ensemble Method*

The comparison of the SVM and Boosted SVM using linear and sigmoid kernels was shown in Table I.

TABLE I. COMPARISON OF PERFORMANCE MODEL OF SVM CLASSIFIERS

Kernel	Model	Accuracy	Precision	Recall
Linear	SVM	0.94	0.96	0.96
	SVM with GridSearchCV	0.95	0.92	0.98
	SVM as base learner in AdaBoost	0.96	0.97	0.97
Sigmoid	SVM	0.93	0.95	0.93
	SVM with GridSearchCV	0.96	0.97	0.97
	SVM as base learner in AdaBoost	0.95	0.94	0.99

The results indicate that the Boosted SVM with a linear kernel model achieved higher accuracy, precision, and recall scores when compared to the SVM classifier model employing a sigmoid kernel. When examining the confusion matrices, it is evident that the Boosted SVM with a linear kernel had a higher total of true positives and false negatives (146) compared to the SVM classifier model (144). Furthermore, the total of false negatives and false positives for the linear kernel in the Boosted SVM was 6, while it was 8 for the SVM classifier model. In contrast, the confusion matrix for the Boosted SVM with a sigmoid kernel displayed a lower total of true positives and false positives (145) than the SVM classifier. This suggests that the SVM classifier predicts more accurately than the Boosted SVM using a sigmoid kernel. Overall, the linear kernel proved to be the best classifier model for this dataset. While the difference in the totals may not be substantial, it still indicates that the Boosted SVM with a linear kernel outperformed the other tested classifier models in this dataset.

In conclusion, by using the PCA feature selection method, the new dataset extracted from the raw data that contains 32 attributes becomes 2 significant attributes that were named principal component 1 and principal component 2. The success of the dataset after PCA feature selection has been shown in an increment of the classifier model accuracy performance model. As shown by a comparison of the data before and after PCA feature selection using the SVM model, the SVM model performs better when using the dataset after PCA feature selection than it does when using the original dataset. Consequently, by implementing the dataset with PCA feature selection, the performance result based on their accuracy, recall, and precision was compared. Thus, SVM as the base classifier with boosting ensemble method has outperformed the SVM models by increasing the accuracy value from 94% to 96% with a precision and recall score of 97%.

IV. CONCLUSION

In this research, the following topics have been identified as recommendations for further work based on the flaws and limits of the dataset and methodology established:

a) The dataset used is imbalance data, which is 212 instances for Malignant data and 357 instances for Benign data. Thus, the data can be balanced using SMOTE technique. Even though the difference is not too wide, it may influence the performance model and increase the skill of the classifier model.

b) This technique can be applied to other machine learning classifiers such as Random Forest, KNN, Naïve Bayes, or Logistic Regression to find the better classifier model to predict breast cancer, to be compared with this research project.

c) The dataset used for this project can be replaced with another dataset that is relevant to the research that will be conducted further. This research can be done to compare the efficiency of the method for a different type of dataset.

d) By implementing this dataset, deep machine learning can be applied to know the method that will result in a better performance model.

ACKNOWLEDGMENT

This work was supported/funded by the Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2020/ICT02/UTM/03/1).

REFERENCES

[1] Breast cancer. (2019). *Nature Reviews Disease Primers*, *5*(1). https://doi.org/10.1038/s41572-019-0122-z.

[2] Arthur, R., Kirsh, V. A., Kreiger, N., & Rohan, T. (2018). A healthy lifestyle index and its association with risk of breast, endometrial, and ovarian cancer among Canadian women. *Cancer Causes & Control, 29*(6), 485-493. https://doi.org/10.1007/s10552-018-1032-1.

[3] Mehta, D., Mohite, A., Shinde, V., Khatri, R., & Dokare, I. (2022). Detection of breast cancer using machine learning algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4108758.

[4] Lou, S. J., Hou, M. F., Chang, H. T., Chiu, C. C., Lee, H. H., Yeh, S. C. J., & Shi, H. Y. (2020). Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: A prospective cohort study. *Cancers, 12*(12), 3817. https://doi.org/10.3390/cancers12123817.

[5] Bataineh, A. A. (2019). A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. *International Journal of Machine Learning and Computing, 9*(3), 248-254. https://doi.org/10.18178/ijmlc.2019.9.3.794.

[6] Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making, 19*(1). https://doi.org/10.1186/s12911-019-0801-4.

[7] Journal of Research in Medical and Dental Sciences, *6*(1), 365-368. https://doi.org/10.5455/jrmds.20186159.

[8] Yadav, A., Jamir, I., Jain, R. R., & Sohani, M. (2019). Breast cancer prediction using SVM with PCA feature selection method. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* 969-978. https://doi.org/10.32628/cseit1952277.

[9] Bhavsar, S., Arora, K., Koul, S., & Barapate, S. (2020, November 23). Understanding bagging & boosting in machine learning. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. (n.d.). UCI Machine Learning. Retrieved 2022, from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

[10] Ghosal, I., & Hooker, G. (2020). Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. *Journal of Computational and Graphical Statistics, 30*(2), 493-502. https://doi.org/10.1080/10618600.2020.1820345.

[11] UCI Machine Learning Repository Website. (n.d.). Retrieved September 30, 2023.

[12] Rimah, R. A., Dorra, D. B. A., & Noureddine, N. E. (2013). Practical selection of SVM supervised parameters with different feature representations for vowel recognition. *International Journal of Digital Content Technology and Its Applications (JDCTA), 7*(9). https://doi.org/10.4156/jdcta.vol7.issue9.50.