Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Optimizing Breast Cancer Prediction by Implementing Feature Selection with Principal Component Analysis

Swati L Nalawade[1], Dr. Suvarna M Patil[2*]

[1]*Research Scholar, Bharati Vidyapeeth Deemed to be University, Institute of Management and Rural Development Administration, Sangli, Maharashtra, India, swatinalawade14@gmail.com*
[2]*Assistant Professor, Bharati Vidyapeeth Deemed to be University, Institute of Management and Rural Development Administration, Sangli, Maharashtra, India, Suvarnampatil@gmail.com*

Breast cancer is a major global health concern, where early detection plays a pivotal role in improving patient outcomes. Advances in machine learning (ML) offer significant potential in enhancing the accuracy of cancer diagnosis, leading to more effective treatment strategies. This study explores the application of machine learning techniques for breast cancer prediction, utilizing a dataset collected from cancer hospitals in Pune, Maharashtra, India. The data includes clinical and diagnostic variables, such as lifestyle factors, hereditary background, and cancer stages.

The primary focus of this research work is to evaluate the impact of dimensionality reduction using Principal Component Analysis (PCA) on the performance of several machine learning classifiers. By reducing the dimensionality of the dataset, the study aims to improve model interpretability, computational efficiency, and predictive accuracy. The classifiers are evaluated both with and without PCA to determine the optimal approach for breast cancer classification.

The results indicate that PCA significantly enhances model performance in terms of accuracy, efficiency, and generalizability, offering a more streamlined approach to breast cancer prediction. This research work suggests that incorporating dimensionality reduction into machine learning workflows can provide valuable support for early diagnosis and personalized treatment in breast cancer care.

Keywords: Breast Cancer Prediction, Feature Selection Methods, Dimensionality Reduction in Oncology, Machine Learning, Principal Component Analysis (PCA).

1. Introduction

Breast cancer remains one of the most prevalent and life-threatening diseases worldwide, affecting millions of women and men each year. It is characterized by the uncontrolled growth of abnormal cells in the breast tissue, which can form tumors capable of invading nearby tissues and metastasizing to other parts of the body. Early detection and accurate diagnosis of breast cancer are crucial for effective treatment and improving patient survival rates. However, a significant challenge in clinical practice is the late diagnosis of the disease, which often results in limited treatment options and poorer prognoses [1].

In recent years, advancements in diagnostic technologies and data-driven approaches have opened new avenues for improving cancer detection and prognosis. Machine learning (ML) has emerged as a promising tool in this regard, offering the potential to enhance diagnostic accuracy and optimize treatment strategies. By leveraging large datasets and sophisticated algorithms, machine learning models can identify patterns in patient data that may be difficult for clinicians to discern through traditional methods.

This study focuses on the development of predictive models for breast cancer classification using machine learning techniques. Specifically, the research investigates the impact of dimensionality reduction through Principal Component Analysis (PCA) on the performance of various classifiers, including Random Forest, XGBoost, Support Vector Machine, Decision Tree, and Naive Bayes. The dataset, collected from cancer hospitals in Pune, Maharashtra, India, comprises a rich set of clinical and diagnostic information, including hereditary factors, lifestyle data, and cancer stages.

The primary objective of this research is to evaluate how dimensionality reduction enhances model efficiency and interpretability, with the ultimate goal of improving breast cancer prediction and facilitating early diagnosis.By comparing the performance of these classifiers with and without PCA, this study aims to identify the most effective approach for breast cancer prediction, contributing valuable insights to the ongoing development of AI-driven tools in healthcare.

2. Literature Review

Huseyin Yilmaza and Fatma Kuncanb [2]reported that by applying Principal Component Analysis (PCA) to reduce the dataset size, they improved the efficiency of breast cancer diagnosis. Among five machine learning algorithms tested, Logistic Regression demonstrated the highest accuracy at 98.8% after PCA was applied.

Chitra Desai [3] explores different methods for feature selection and dimensionality reduction, emphasizing the use of principal component analysis to condense breast cancer data. They also explore detection using machine learning algorithms, alongside feature selection through one of these algorithms.

The Mohammad Kaosain Akbar [4] applied Principal Component Analysis (PCA) to breast cancer datasets, demonstrating that while the attributes are correlated, they influence the principal components differently. They trained two machine learning models using Logistic Regression—one with the original data and the other with PCA-transformed data. Their findings revealed that the model trained on the PCA-transformed data outperformed the one trained on the original dataset in terms of accuracy, precision, F1-score, and recall. They plan to extend this study using datasets with more attributes for further comparison.

Zuhaira Muhammad Zain et al. [5] found that using PCA for feature extraction enhanced the performance of the Naïve Bayes (NB) and REPTree classifiers, achieving F-measure values of 76.1% and 72.8%, respectively, on the WPBC dataset. They recommend further investigation into alternative feature extraction methods to improve prediction accuracy and tackle the challenge of imbalanced data in prognostic breast cancer datasets.

Boluwaji A. Akinnuwesi et al. [6] created a breast cancer risk assessment and early diagnosis model by integrating Support Vector Machine with Principal Component Analysis following multi-stage preprocessing. This model showed significant improvement over prior studies, reaching 97.62% accuracy, 95.24% sensitivity, and 100% specificity. The BC-RAED model successfully classifies breast cancer risk and distinguishes cases as malignant or benign.

The Sara Ibrahim et al.[7] aimed to improve breast cancer classification by selecting significant features using correlation analysis and variance before applying classification methods. They evaluated classifiers on the WBCD dataset, using dimensionality reduction techniques like correlation and principal component analysis. After selecting the top seven classifiers, hyperparameter tuning was performed, and two voting techniques (hard and soft) were applied. The proposed approach achieved better results than previous methods, with 98.24% accuracy, 99.29% precision, and 95.89% recall.

The Subash Kumar [8] describe Principal Component Analysis (PCA) as an unsupervised technique that identifies the underlying structure among variables by determining orthogonal lines of best fit. PCA transforms the data to find which features explain the most variance, eliminating components that contribute less. In their study, they use 12 attributes, including 10

real-valued features from the WDBC dataset, to classify cancer stage as either malignant or benign.

Nguyen et al. [9] assessed performance of breast cancer classification models of supervised and unsupervised for WBCD dataset. The study incorporated scaling and Principal Component Analysis for feature selection, with data divided into a 70:30 ratio for training and testing. The findings revealed that the ensemble voting method was effective in breast cancer prediction, with the voting classifier, AdaBoost, logistic regression, and support vector machine achieving approximately 90% accuracy.

In a separate study, Ahmed et al. [10] compared the performance of decision trees, artificial neural networks, and support vector machines. Their findings indicated that SVMs surpassed both DTs and MLPs. However, their study had limitations, including the loss of cases during follow-up, missing records, and the exclusion of important variables like S-phase fraction and DNA index, which may have impacted model performance.

Omondiagbe et al.[11]discussed the classification of different types of breast cancer from WBCD datasets using support vector machine, artificial neural network and naive Bayes algorithms. Researcher introduced a hybrid approach for diagnosis of breast cancer, utilizing LDA to reduce feature dimensionality, followed by applying the optimized feature set to an SVM. This method resulted in an accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07%, and an AUROC of 0.9994.

Researchers also implemented recursive feature elimination (RFE)[12], a wrapper approach that evaluated all feature subsets based on their accuracy scores, selecting those with top-ranking features. Their study focused on applying principal component analysis (PCA) to neural networks, utilizing both PCA and linear discriminant analysis for feature extraction and CFS and RFE for feature selection.

Jamal et al. [13] evaluated the performance of support vector machine and extreme gradient boosting machine learning algorithms and reduced the number of data attributes through feature extraction using clustering with k-means and PCA and accessed their performance. Their findings revealed that k-means, although not typically used for dimensionality reduction, outperformed PCA. They also applied different analysis and scaling techniques concluding that the highest accuracy on Wisconsin dataset was achieved by combining PCA with a back-propagation neural network.

Most of the existing literature evaluates classifier performance primarily based on accuracy, which increases when true positives and true negatives outweigh false positives and false negatives. However, it is equally crucial to assess performance using recall (false negatives), precision (false positives), and the F-measure, as missing a diagnosis could have severe implications for patient outcomes.

3. Research Design

In this study, researcher employed a Design and Creation Approach for data collection from various cancer hospitals in the Maharashtra region. Data on breast cancer patients were gathered from multiple cancer hospitals, incorporating features such as clinical data and histo-pathological findings. To enhance the accuracy of the model results and develop a robust predictive model, a significant dataset was utilized. A sufficient amount of the data was allocated for training, validation, and testing of the ML model. Then breast cancer dataset was divided into training and testing sets using a 70:30 ratio, accomplished through the train-test split method in Python’s Scikit-learn library.

3.1 Data Collection

The breast cancer dataset encompasses information related to a variety of factors, including lifestyle habits, dietary patterns, and hereditary predispositions. Additionally, the dataset contains detailed data regarding the stages of cancer at the time of diagnosis, ranging from early-stage to advanced-stage cancer, thus providing a comprehensive view of the patients’ diagnostic statuses. The dataset consists of 1,224 patient records collected from participating cancer hospitals in the Pune region, covering a wide age range of 25 to 85 years.

The dataset includes both clinical and diagnostic data, such as pathology reports. Guided by medical expertise, key attributes affecting breast cancer were selected for further analysis. To prepare the dataset for breast cancer research, the researcher initially worked with 30 attributes obtained from the collected data. Data preprocessing involved multiple steps, such as cleaning, transformation of data, and data normalization.

As shown in Figure 1, after the preprocessing phase, the researcher performed feature selection and dimensionality reduction by evaluating the mean and variance of the input features. The most relevant features were then used for classification algorithms such as Support Vector Machine, Naïve Bayes, Random Forest, XGBoost and Decision Tree.

The workflow of the model diagram illustrates the execution process. Initially, the researcher collects the breast cancer data and documented in .csv format, followed by performing all data pre-processing tasks. This pre-processed dataset is then used to develop a machine learning model using the Python platform. The model outputs whether the entered patient's record indicates No cancer, Benign stage, or Malignant stage of cancer.

The following sections provide detailed explanations of each step in the proposed methodology.

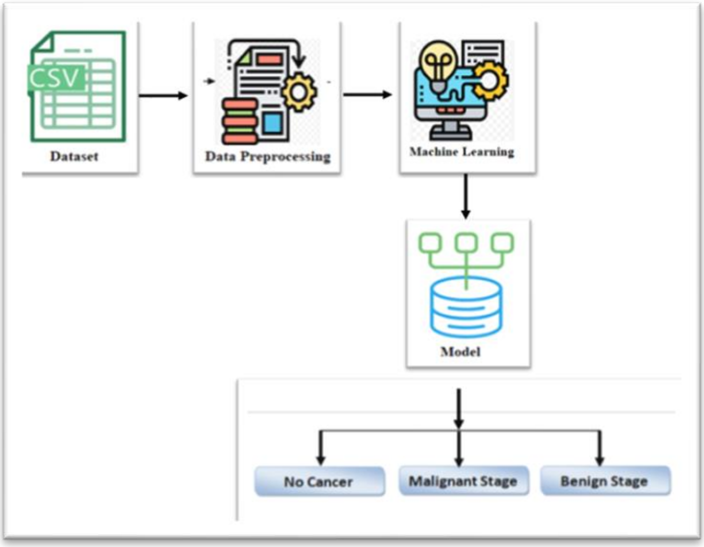


Figure1. Accurate breast cancer prediction proposed methodology

3.2 Data Pre-Processing

Data pre-processing was conducted to enhance quality of data and generate a precise dataset suitable for model building. Without effective pre-processing, various challenges can arise, including inconsistencies, errors, noise, missing values, and model overfitting.

While performing pre-processing on the collected dataset it is found that last 4 attributes contain maximum missing values and it directly impacts on the performance of model so these attributes are dropped and not considered for further analysis.

During the preprocessing phase, the researcher implemented feature reduction through Principal Component Analysis.

3.3 Dimensionality Reduction Using Principal Component Analysis

In high-dimensional datasets, like those often encountered in breast cancer prediction, redundant and irrelevant features can reduce model accuracy and increase computational complexity. Principal Component Analysis (PCA) is employed to address these challenges, serving as an effective technique for dimensionality reduction.

The following points outline its implementation and role in this study:

1. Purpose of PCA: PCA transforms the original variables into a set of uncorrelated principal components, capturing the highest possible variance in the data. This allows the model to focus on the most informative features, which minimizes the risk of overfitting while

enhancing interpretability and computational efficiency.

2. PCA in Feature Selection: By retaining components that capture the essential variance, PCA allows us to exclude redundant information without sacrificing model performance. This reduction in features contributes directly to more efficient and accurate predictive outcomes in breast cancer prediction.

3. Use of the Scree Plot: The scree plot is utilized to determine the optimal number of principal components to retain. By plotting each component’s explained variance, the "elbow point" in the plot highlights where additional components contribute minimally to the variance. Selecting components up to this point ensures that the model is both parsimonious and robust, capturing key data features without unnecessary complexity.

Through PCA and the scree plot, our approach leverages the most relevant features, improving the accuracy and stability of the breast cancer prediction model. The selected features were further evaluated for their variance using principal component analysis (PCA). To achieve dimensionality reduction, the researcher employed the widely recognized PCA algorithm. Specifically, the PCA function from the sklearn.decomposition module in the sklearn library was utilized. PCA applies linear dimensionality reduction by performing a singular value decomposition of the data, which projects it into a lower-dimensional space for feature selection.

Since PCA is sensitive to the scale of the features, it was necessary to standardize all features to ensure that those with higher variance did not disproportionately influence the results. For this purpose, the "StandardScaler" was applied to normalize the features before performing PCA for dimensionality reduction.

The resulting analysis demonstrates that most of the variance can be effectively captured using only the selected 18 features.

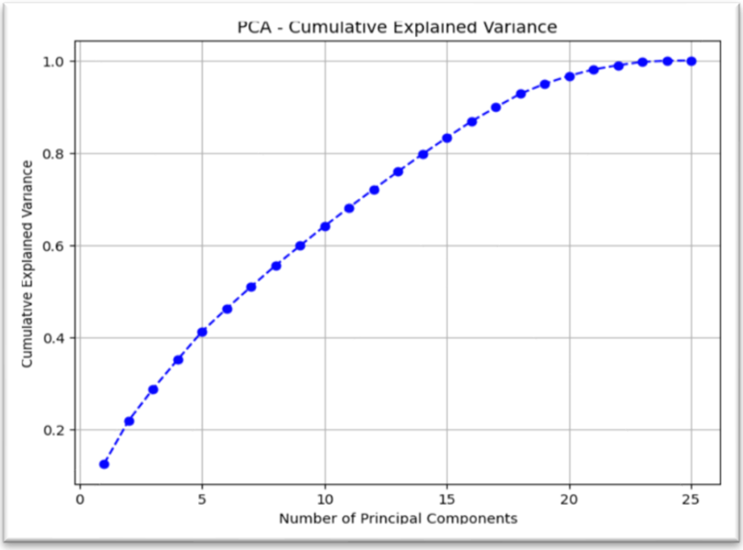


Figure2. Scree plot analysis in the Breast Cancer dataset.

Figure 2 illustrates the cumulative explained variance in relation to the number of principal components, as derived from the scree

plot analysis. Based on the scree plot analysis, the analysis suggests that the first 18 components are sufficient to explain approximately 90% of the variance, making them suitable for further analysis. This selection effectively balances model complexity with the preservation of information.

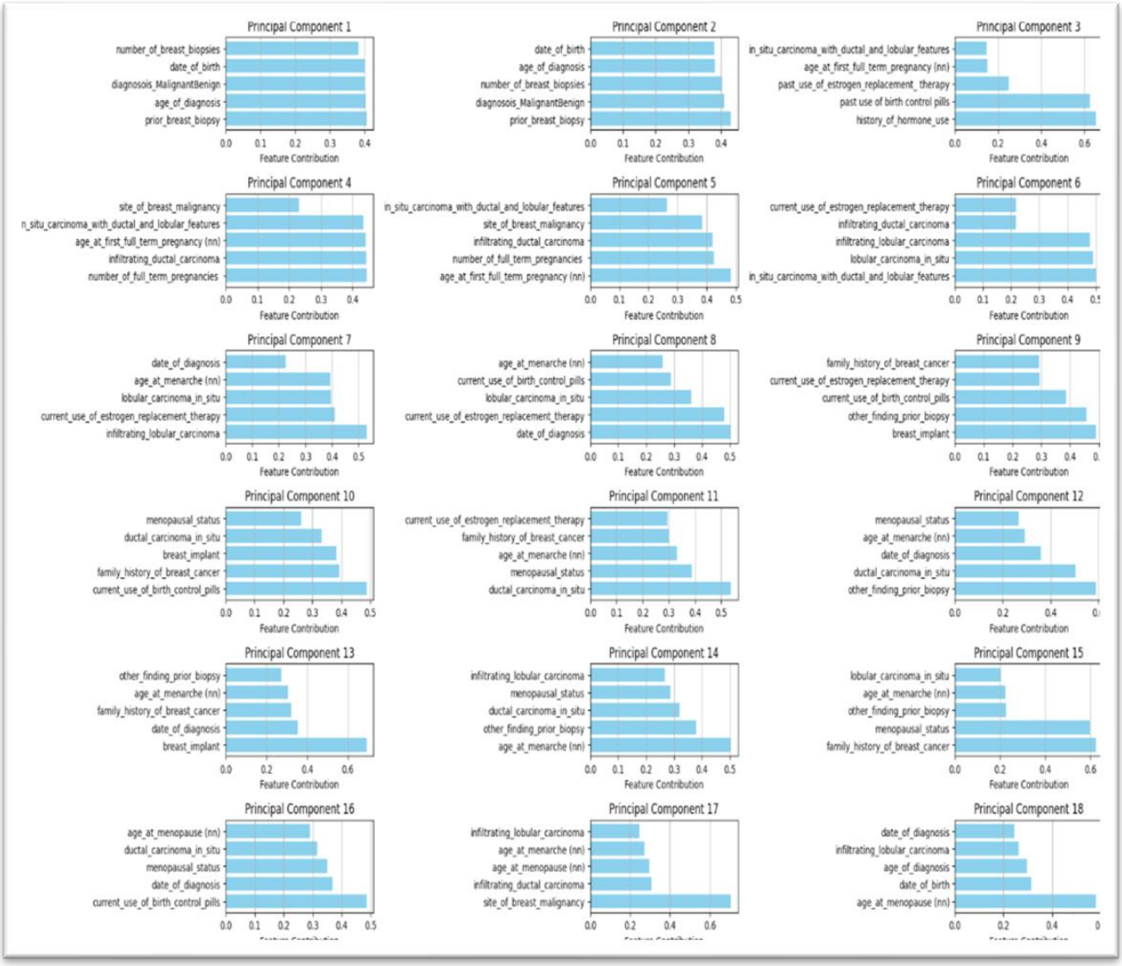


Figure3. Top Contributing Features for Each of the 18 Components in the Breast Cancer dataset.

From figure 3, the bar plot illustrates the contributions of the top features for each of the 18 principal components derived from the PCA applied to the breast cancer dataset.The bar plot effectively summarizes the contributions of the top features for each principal component, providing insights into the significance of various attributes in relation to the variance captured by the components. This visualization aids in feature interpretation and selection, enabling you to make informed decisions in subsequent analysis or predictive modelling efforts.

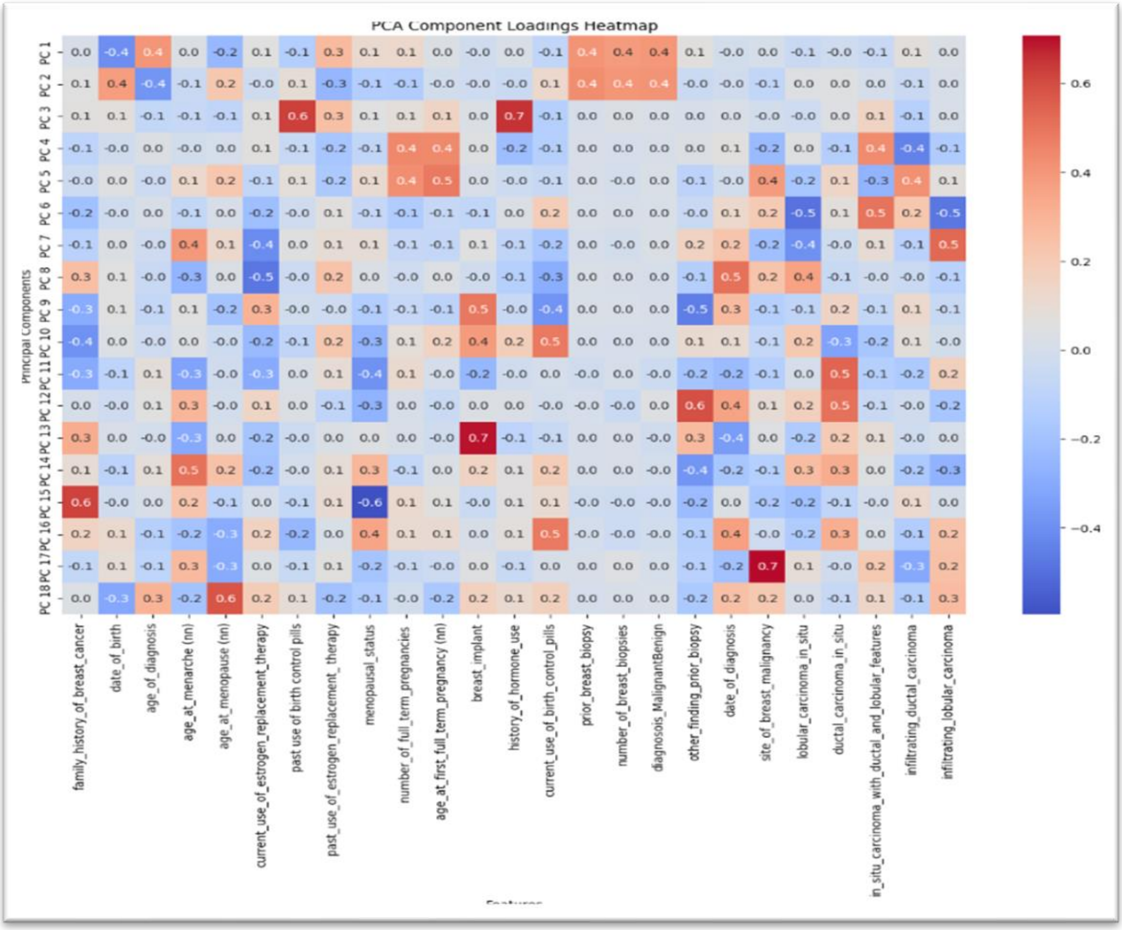


Figure4. Heatmap for Breast Cancer dataset.

Figure 4 illustrates a heatmap derived from the PCA analysis, visualizing the loadings of the principal components in relation to the original features of the breast cancer dataset. This heatmap succinctly summarizes how well these principal components capture the variance in the data while also identifying the most informative features. The visualization facilitates feature selection and interpretation, providing valuable insights for subsequent analyses or modeling efforts.

By examining the loading values, researcher can discern the importance and relationships of various features concerning the principal components. Such analysis is especially critical in breast cancer research, were pinpointing significant factors can improve diagnostic accuracy and treatment strategies, ultimately enhancing the effectiveness of breast cancer prediction.

4. Experimentation and Discussion

For analysis selected dataset consists of 1,224 patient records collected from participating cancer hospitals in the Pune region, covering a wide age range of 25 to 85 years. The 1224 instances with selected 12 characteristics in the database consist of three cases: 435 No cancer instances, 379 benign instances and 392malignant instances. A 70:30 split was used for training and testing to evaluate the algorithm, utilizing the train-test split method from Python's Scikit-learn library.

The evaluation and optimization of machine learning classifiers—such as Random Forest, XGBoost, Naive Bayes, SVM, and Decision Tree—are vital for ensuring their robustness and effectiveness in predicting breast cancer. Key performance metrics such as accuracy, precision, recall, and F1 score provide a thorough evaluation of model performance, with confusion matrices offering insights into classification patterns.

Techniques like cross-validation and hyperparameter tuning, using methods such as GridSearchCV and RandomSearchCV, enhance model reliability and mitigate issues of overfitting or underfitting. Moreover, a focus on feature importance through systematic inclusion and exclusion testing allows researchers to identify and retain the most relevant attributes for classification. This comprehensive approach not only refines the models for optimal performance but also maintains their interpretability, leading to more reliable and insightful predictions in healthcare applications. Each classifier brings unique strengths, contributing to a cohesive framework for predictive modeling in breast cancer diagnosis, ultimately advancing the capabilities of machine learning in healthcare settings.

The following formulas are used to calculate and display the classification metrics in the result tables:

True Positive: Correctly positive (cancerous).

True Negative: Correctly negative (non-cancerous).

False Positive: Incorrectly positive.

False Negative: Incorrectly negative.

1. Accurately Classified Instances (TP + TN):

Accurately Classified Instances=True Positives (TP)+True Negatives (TN)

It calculates the total number of instances by encompassing both positive (cancerous) and negative (non-cancerous) cases.

2. Inaccurately Classified Instances (FP + FN):

This formula represents the total number of instances that the model misclassified, including both false positives and false negatives.

3. Kappa Statistic (Cohen's Kappa):

$$\boldsymbol{k = \frac{(Po - Pe)}{(1 - Pe)}}$$

Where: Po (Observed Agreement) is calculated as:

$$\boldsymbol{Po = \frac{(TP + TN)}{Total\ Instances}}$$

Pe (Expected Agreement) is calculated as:

$$\boldsymbol{Pe = \frac{(TP + FP) * (TP + FN)}{Total\ Instances^2} + \frac{(TN + FP) * (TN + FN)}{Total\ Instances^2}}$$

4.1 Model Evaluation without PCA:

Algorithm	TP	FN	TN	FP	Accuracy	F-1	Recall	Precision
Random Forest Classifier	170	14	30	29	84.16	0.84	0.84	0.85
Decision Tree	53	75	62	13	80.61	0.80	0.81	0.81
XGBoost classifier	163	20	137	21	80.69	0.81	0.81	0.81
Support Vector Machine	178	18	189	11	88.12	0.88	0.88	0.89
Naive Bayes Classifier	157	41	163	41	77.72	0.77	0.78	0.83

Table 1. Breast cancer prediction using ML algorithms 26 attributes

From Table1, the evaluation of various classifiers for breast cancer prediction revealed that the Support Vector Machine achieved the highest accuracy at 88.12%, closely followed by the Random Forest Classifier at 84.16%. These models demonstrated strong performance across all key metrics, including F1 score, recall, and precision, indicating their effectiveness in correctly classifying instances while minimizing false positives.

In contrast, the Decision Tree and XGBoost Classifier showed notably lower accuracy of 80.61% and 80.69%, respectively, suggesting that optimization did not significantly enhance the performance of the Decision Tree. The Naïve Bayes Classifier performed the weakest, with an accuracy of only 77.72%, highlighting its limitations in this context.

Algorithm	Accurately Classified Instances (TP+TN)	Inaccurately Classified Instances (FP+FN)	Kappa Statistics [(Po-Pe)/(1-Pe)]
Random Forest Classifier	200 (82.30)	43 (17.70)	0.307
Decision Tree	115 (56.65)	88 (43.35)	0.205
XGBoost Classifier	300 (87.98)	41 (12.02)	0.757
Support Vector Machine	367 (92.68)	29 (07.32)	0.905
Naive Bayes Classifier	320 (79.60)	82 (20.40)	0.558

Table 2. Summary statistic for ML models with 26 attributes

Table 2 lists the summary statistics for the fiveMachine Learning models with 26 attributes. The summary of Kappa statistic values indicates that all models achieve moderate to substantial agreement in classification. The Support Vector Machine classifier stands out with the highest Kappa value (0.905), reflecting better classification performance beyond chance. XGBoost (0.757) and Naive Bayes classifier (0.558) all perform similarly, while the Random Forest Classifier (0.307) and Decision Treeclassifier (0.205) shows weaker agreement, indicating that it may not be the most suitable option for this specific dataset.

4.2 Model Evaluation with PCA:

Algorithm	TP	FN	TN	FP	Accuracy	F-1	Recall	Precision
Random Forest Classifier	944	29	30	32	93.84	0.94	0.94	0.94
Decision Tree	914	40	41	50	90.85	0.91	0.91	0.91
XGBoost classifier	944	23	20	43	93.84	0.94	0.94	0.94
Support Vector Machine	845	51	33	135	84.01	0.85	0.84	0.84
Naive Bayes Classifier	777	89	96	164	77.24	0.77	0.77	0.77

Table 3. Breast cancer prediction using principal component analysis on 18 attributes

The table3 summarizes the performance metrics of five different machine learning algorithms on collected dataset as Random Forest, Decision Tree, XGBoost, Support Vector Machine, Naive Bayes Classifier. Key metrics such as True Positives, False Negatives, True Negatives, False Positives, Accuracy, F1 score, Recall, and Precision are used to assess the models.

- Random Forest and XGBoost both achieved the highest accuracy of 93.84% and strong performance across F1 score, recall, and precision (all 0.94), indicating excellent classification results.
- The Decision Tree follows with a slightly lower accuracy of 90.85% and consistent F1, recall, and precision values of 0.91.
- SVM displayed a notable drop in accuracy at 84.01% with an F1 score, recall, and precision of 0.85, reflecting a moderate level of misclassification, particularly with false positives.
- The Naive Bayes Classifier performed the weakest, with the lowest accuracy of 77.24%, and F1, recall, and precision values of 0.77, indicating that it struggles more with both false negatives and false positives.

The table illustrates how different classifiers perform in predicting breast cancer, with Random Forest and XGBoost providing the most reliable results, while Naive Bayes lags behind.

Algorithm	Accurately Classified Instances (TP+TN)	Inaccurately Classified Instances (FP+FN)	Kappa Statistics [(Po-Pe)/(1-Pe)]
Random Forest Classifier	974 (94.11)	61 (5.89)	0.2925
Decision Tree	955 (91.39)	90 (8.61)	0.2930
XGBoost Classifier	964 (93.59)	66 (6.41)	0.2689
Support Vector Machine	878 (82.52)	186 (17.48)	0.3091
Naive Bayes Classifier	873 (77.53)	253 (22.47)	0.2546

Table 4. Summary statistic for ML models with principal component analysis

Table 4 presents the summary statistics for various machine learning models evaluated with principal component analysis in breast cancer classification. The table 4 provides a comparative analysis of five machine learning classifiers as Random Forest, Decision Tree, XGBoost, Support Vector Machine and Naive Bayes—based on their classification performance.

The Random Forest Classifier had the highest accuracy with 974 correctly classified instances and a Kappa statistic of 0.2925, indicating moderate agreement. The Decision Tree closely followed with 955 accurate classifications and a slightly better Kappa of 0.2930. The XGBoost Classifier achieved 964 accurate instances but had a lower Kappa value of 0.2689, suggesting weaker agreement. The Support Vector Machine recorded 878 accurate classifications and the highest Kappa of 0.3091, indicating better agreement despite lower accuracy. The Naive Bayes Classifier performed the least effectively, with 873 accurately classified instances and the lowest Kappa statistic of 0.2546. Overall, the table illustrates the relative strengths and weaknesses of each classifier in predicting breast cancer outcomes.

4.3 Comparative Analysis:

The comparative analysis assessed models with and without PCA to evaluate the impact of feature dimensionality reduction on performance. Figures 5 and 6 display the results of five classification models both with and without PCA. The findings reveal that PCA improves the performance of the Random Forest, Decision Tree, and XGBoost classifiers by increasing correctly classified instances and reducing errors, while no significant improvement was observed for SVM and Naïve Bayes.

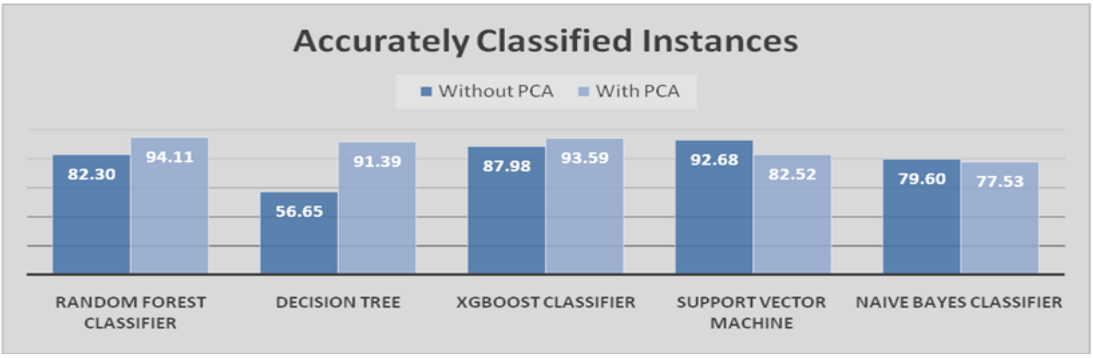


Figure 5. Accurately Classified Instances for five models without and with PCA on Breast Cancer dataset.

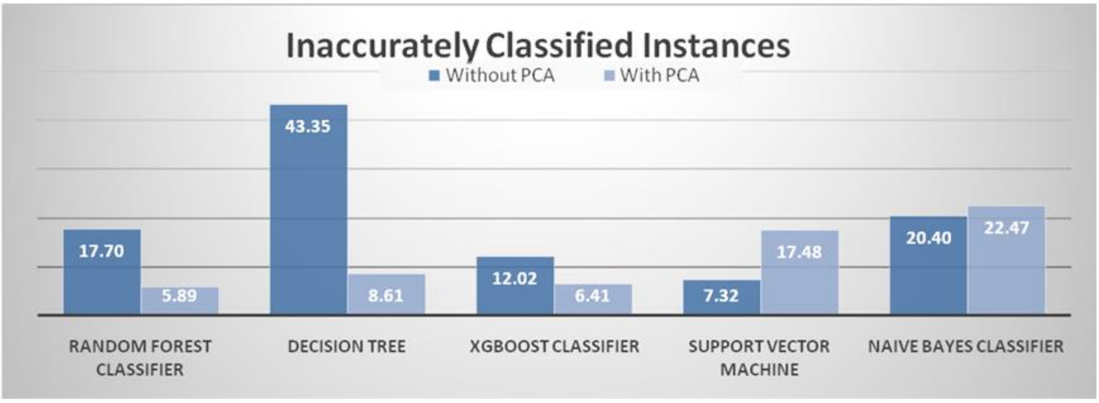


Figure 6. Inaccurately Classified Instances for five models without and with PCA on Breast

Cancer dataset.

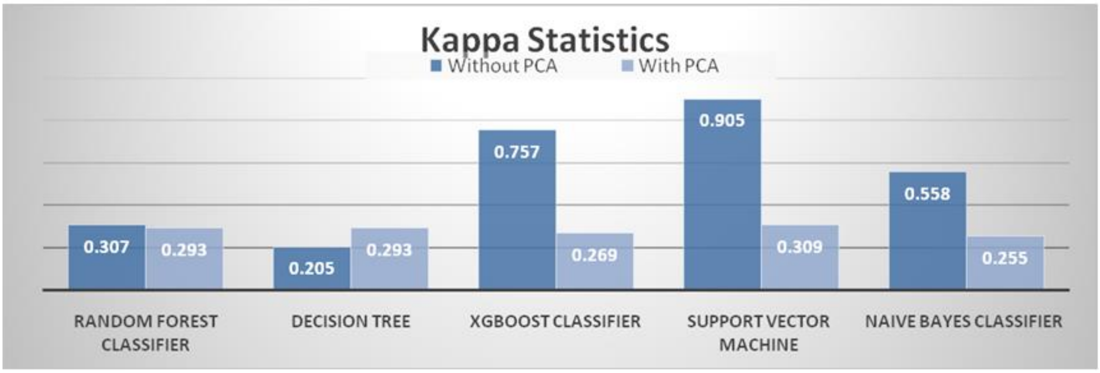
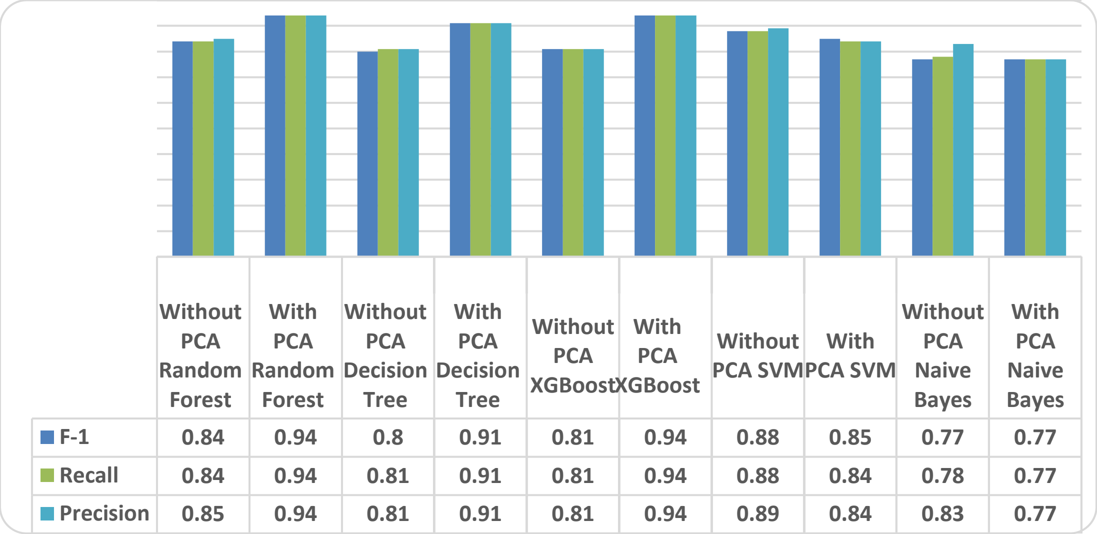


Figure 7. Kappa statistics for five models without and with PCA on Breast Cancer dataset.

Figure 7 shows, comparison of Kappa statistics after applying PCA shows that most machine learning models experience a reduction in consistency. XGBoost, SVM, and Naive Bayes show significant decreases in Kappa values, indicating a noticeable drop in performance agreement after dimensionality reduction. Random Forest sees only a slight decrease in Kappa, suggesting a minimal impact. Interestingly, Decision Tree is the only model that shows an improvement in Kappa, indicating enhanced consistency in its performance after applying PCA.



	Without PCA Random Forest	With PCA Random Forest	Without PCA Decision Tree	With PCA Decision Tree	Without PCA XGBoost	With PCA XGBoost	Without PCA SVM	With PCA SVM	Without PCA Naive Bayes	With PCA Naive Bayes
■ F-1	0.84	0.94	0.8	0.91	0.81	0.94	0.88	0.85	0.77	0.77
■ Recall	0.84	0.94	0.81	0.91	0.81	0.94	0.88	0.84	0.78	0.77
■ Precision	0.85	0.94	0.81	0.91	0.81	0.94	0.89	0.84	0.83	0.77

Figure 8. F-1 measures, Recall and Precision for five models without and with PCA on Breast Cancer dataset.

Figure 8 shows, PCA significantly improves the F1 score, recall, and precision for Random Forest, Decision Tree, and XGBoost, raising metrics from around 0.80-0.85 to 0.91-0.94. However, SVM experiences a slight decline in performance, and Naive Bayes shows minimal changes, with scores remaining around 0.77-0.83. Overall, PCA enhances performance for most models, except for SVM and Naive Bayes.

Algorithm	Original dataset with 26 attributes	PCA using n=18 component
Random Forest Classifier	84.16	93.84
Decision Tree	80.61	90.85
XGBoost Classifier	80.69	93.84
Support Vector Machine	88.12	84.01
Naive Bayes Classifier	77.72	77.24

Table 5. Comparison of accuracy of ML models for different cases

The results from Table 5 indicate that Random Forest Classifier and XGBoost Classifier significantly benefit from the dimensionality reduction achieved through PCA, demonstrating improved accuracy when using 18 components compared to the full 26 attributes. The Decision Tree classifier also shows enhanced performance with PCA. In contrast, the Support Vector Machine exhibits a decrease in accuracy when using PCA, indicating that it may be sensitive to the reduction in dimensionality. Lastly, the Naive Bayes Classifier shows consistent performance, with no improvement from the feature reduction.

Overall, the table illustrates how PCA can optimize the performance of certain classifiers while revealing sensitivity in others, underscoring the importance of feature selection in machine learning applications for breast cancer diagnosis.

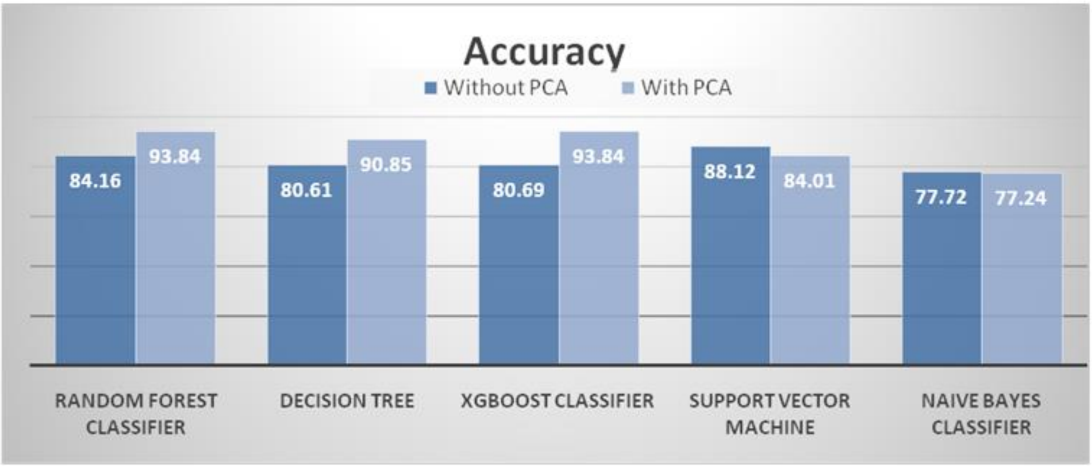


Figure 9. Comparison of Accuracy for five models without and with PCA on Breast Cancer dataset.

Figure 9 shows that, comparative analysis of model performance with and without PCA revealed that Random Forest and XGBoost benefited significantly from dimensionality reduction, with accuracies improving to 93.84%. In contrast, the Support Vector Machine showed a decrease in accuracy to 84.01% following PCA application, indicating potential drawbacks of dimensionality reduction for certain models.

5. Conclusion

In this study, the researchers investigated the application of machine learning techniques for predicting breast cancer using a dataset collected from cancer hospitals in Pune, Maharashtra. Principal Component Analysis (PCA) was employed to reduce dimensionality and enhance model performance.The results demonstrate that machine learning classifiers, especially Random Forest and XGBoost, show strong predictive capabilities in distinguishing between benign and malignant breast tumors. The integration of PCA further enhances the efficiency and accuracy of these models by reducing complexity and computational overhead without compromising predictive performance.

Among the evaluated classifiers—Random Forest, XGBoost, Support Vector Machine, Decision Tree, and Naive Bayes, the Random Forest and XGBoost yielded the highest accuracy (93.84%) both with and without PCA. In contrast, the Naive Bayes Classifier consistently lagged, highlighting its limitations in this application. The analysis also underscored the importance of performance metrics such as accuracy, F1 score, recall, and precision, alongside confusion matrices and Kappa statistics, to provide a comprehensive evaluation of each model's effectiveness.Overall, the findings illustrate the potential of machine learning algorithms in enhancing breast cancer diagnosis accuracy, emphasizing the need for careful feature selection and model choice to improve predictive capabilities in healthcare settings.The use of AI-driven models in healthcare offers a promising avenue for personalized medicine and optimized cancer management. Further research could explore additional feature selection techniques and larger datasets to enhance generalizability and model robustness.

References

1. Coccia,M., “The increasing risk of mortality in breast cancer: A socio-economic analysis between countries”. J.Soc.Adm.Sci.2019,6 218–230.
2. Yilmaz, H., Kuncan, F., "Analysis of Different Machine Learning Techniques with PCA in the Diagnosis of Breast Cancer". Journal of Engineering Technology and Applied Sciences 7 (3) 2022: 195-205.
3. “Analysis of Impact of Principal Component Analysis and Feature Selection for Detection of Breast Cancer Using Machine Learning Algorithms”, Chitra Desai, Department of Computer Science, National Defence Academy, Pune.Journal of Information and Computational Science, Volume 13 Issue 1 – 2023, ISSN: 1548-7741.
4. “Breast Cancer Prediction using Principal Component Analysis with Logistic Regression”, Mohammad Kaosain Akbar,Concordia Institute for Information Systems Engineering (CIISE). Concordia University, Montreal, Quebec, Canada. International Journal of Advances in Engineering and Management (IJAEM), Volume 4, Issue 10 Oct. 2022, pp: 1189-1196 www.ijaem.net ISSN: 2395-5252.
5. “Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis”. Zuhaira Muhammad Zain a,1, *, Mona Alshenaifi a,2, Abeer Aljaloud a,3, Tamadhur Albednah a,4, Reham Alghanim a,5, Alanoud Alqifari a,6, Amal Alqahtani a,7. International Journal of Advances in Intelligent Informatics ISSN 2442-6571 Vol. 6, No. 3, November 2020, pp. 313-327. https://doi.org/10.26555/ijain.v6i3.462http://ijain.orgijain@uad.ac.id

6. “Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques”, Boluwaji A. Akinnuwesia, *, Babafemi O. Macaulay b, Benjamin S. Aribisala b a Department of Computer Science, Faculty of Science and Engineering, University of Eswatini, Kwaluseni M201, Swaziland b Department of Computer Sciences, Faculty of Science, Lagos State University, Lagos, Nigeria. Informatics in Medicine Unlocked 21 (2020) 100459. Journal homepage: http://www.elsevier.com/locate/imu .
7. Ibrahim, S.; Nazir, S.; Velastin, S.A. “Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis”. 2021, 7, 225. https://doi.org/10.3390/jimaging7110225 Academic Editors: Leonardo Rundo, Carmelo Militello, Vincenzo Conti, Fulvio Zaccagna and Changhee Han.
8. Subash Kumar, “Principal Component Analysis on the Breast Cancer – Python”, International Journal for Modern Trends in Science and Technology, 6(10): 134-136, 2020. ISSN: 2455-3778 online DOI: https://doi.org/10.46501/IJMTST061024 Available online at: http://www.ijmtst.com/vol6issue10.html.
9. Nguyen,Q.H.; Do,T.T.; Wang,Y.; Heng,S.S.; Chen,K.; Ang,W.H.M.; Philip,C.E.; Singh,M.; Pham,H.N.; Nguyen,B.P.; et al. “Breast Cancer Prediction using Feature Selection and Ensemble Voting”. In Proceedings of the 2019. International Conference on System Science and Engineering (ICSSE), Dong Hoi City, Vietnam, 19–21 July 2019; pp. 250–254.
10. Ahmad,L.G.; Eshlaghy,A.; Poorebrahimi,A.; Ebrahimi,M.; Razavi,A.; “Using three machine learning techniques for predicting breast cancer recurrence”. J. HealthMed.Inform.2013,4,3.
11. Omondiagbe, D.A.; Veeramani, S.; Sidhu, A.S. “Machine Learning Classification Techniques for Breast Cancer Diagnosis”. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Kazimierz Dolny, Poland,21–23 November2019; Volume 495, p. 012033.
12. Chen, X.W.; Jeong, J.C. “Enhanced recursive feature elimination”. In Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, USA, 13–15 December 2007; pp. 429–435.
13. Jamal,A.; Handayani,A.; Septiandri,A.; Ripmiatin,E.; Effendi,Y., “Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction”. Lontar Komput. J. Ilm. Teknol. Inf. 2018, 9, 192–201.