https://doi.org/10.48047/AFJBS.6.12.2024.3636-3647

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Advanced Machine Learning Techniques for Early Breast Cancer Detection through DNA Methylation Analysis

Dr. Velayutham Pavanasam
Professor, Department of Computer Science and Engineering,
Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research
Foundation (Deemed to be University), Paiyanoor, Chennai (Tamilnadu),
INDIA
Email: velayutham.avcs098@avit.ac.in

Deepak Chandra Uprety
Associate Professor and Research Coordinator, Department of
Cloud Computing, Noida Institute of Engineering and Technology,
Greater Noida (UP), INDIA Email: *deepak.glb.@gmail.com*

Dr. Rashmi Dwivedi
Assistant Professor / HOD, Department of
Economics, NIMS School of Humanities and
Liberal Arts,
NIMS University Rajasthan, Jaipur (Raj.), INDIA

Nelofar Bashir
Research Scholar, School of Computing and Artificial
Intelligence, NIMS University Rajasthan, Jaipur (Raj.),
INDIA
Email: *nelofarbashir111@gmail.com*

Balaji VS
Department of Artificial Intelligence and Machine
Learning Rajalakshmi Engineering College, Chennai, India
Email: *balajivsb1@gmail.com*

Amjed Khan Bhatti
Assistant Professor, Department of Computer Science and Information
Technology Government College for Women, Gandhinagar, Jammu (J&K),
INDIA
Email: *amjedbhatti07@gmail.com*

Article History
Volume 6, Issue 12, 2024
Received: 02 Jun 2024
Accepted: 25 Jun 2024
doi:
10.48047/AFJBS.6.12.2024.3636-3647

Abstract**:** Breast cancer remains a leading cause of mortality among women worldwide, highlighting the critical need for effective early detection methods. DNA methylation, involving variations in methylation levels, serves as a significant biomarker for identifying cancerous changes at an early stage. In 2018 alone, approximately 40,920 women lost their lives to breast cancer, emphasizing the urgency of improving diagnostic techniques. Recent advancements in technology have enabled the development of more precise and timely prediction models for such conditions. Among these advancements, machine learning has emerged as a transformative tool, capable of analyzing complex physical and behavioral data to predict diseases with high accuracy. In the context of early breast cancer detection, a cascaded approach utilizing advanced machine learning techniques has shown great promise. This study introduces a multi-step methodology that begins with the Standard Deviation Threshold based Differential Mean Feature Selection (DMFS) technique. By selecting the most informative features from the input data, this method optimizes prediction accuracy through a threshold set at the standard deviation of weight vectors. Following feature selection, Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA) are applied to further refine these features, enhancing the model's ability to distinguish between cancerous and non-cancerous states. This paper presents a comprehensive analysis of these advanced machine learning techniques and their application to DNA methylation data for early breast cancer detection. The findings demonstrate significant improvements in prediction accuracy, underscoring the potential of these methods to contribute to more effective and timely breast cancer diagnosis, ultimately aiming to reduce mortality rates associated with the disease.

Keywords**:** Breast Cancer Prediction, DMFS, PCA, Random Forest Classifier, DNA Methylation, Accuracy, F measure, RMSE

Introduction

Breast cancer remains a significant health challenge globally, being one of the leading causes of death among women. Historically, the battle against breast cancer has focused on improving detection and treatment methods to reduce mortality rates. Despite advances in medical research and technology, early diagnosis remains a critical factor in improving patient outcomes. Early detection often leads to more effective treatment options and a higher likelihood of survival. One promising avenue for early detection is the analysis of DNA methylation, which involves examining variations in methylation levels that can serve as early biomarkers for cancerous changes.

In previous years, research has established the importance of DNA methylation in cancer biology. Variations in methylation patterns have been identified as key indicators of the presence and progression of various cancers, including breast cancer. However, traditional methods of analyzing these patterns were often labor-intensive and limited in scope. The advent of high-throughput sequencing technologies revolutionized this field by enabling the comprehensive analysis of methylation patterns across the genome, providing a wealth of data for researchers to explore. Despite these advances, the challenge remained to accurately interpret this data and translate it into reliable diagnostic tools.

Presently, machine learning has emerged as a powerful tool in the analysis of complex biological data. Machine learning algorithms can process large datasets, identify patterns, and make predictions with high accuracy. In the realm of breast cancer detection, machine learning models have shown great promise in analyzing DNA methylation data to identify early signs of cancer. The current study employs a cascaded methodology that integrates several advanced machine learning techniques to enhance prediction accuracy. Initially, the Standard Deviation Threshold based Differential Mean Feature Selection (DMFS) technique is used to select the most informative features from the input data. By setting the threshold at the standard deviation of weight vectors, this method ensures that only the most relevant features are considered. Following this, Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA) are applied to further refine these features, optimizing the model's predictive capabilities.

The implications of this research are profound, offering a glimpse into the future of cancer detection and diagnosis. As technology continues to advance, the integration of more sophisticated machine learning algorithms and more comprehensive datasets will likely further improve the accuracy and reliability of early detection methods. Future perspectives include the development of more personalized diagnostic tools that can account for individual variations in DNA methylation patterns, potentially leading to tailored treatment plans that are more effective for each patient.

Machine Learning (ML), a subset of Artificial Intelligence (AI), enables machines to learn from data without heavy programming, facilitating experiences such as prediction, clustering, and decision-making [1]. Classification, essential in ML, helps identify future instances by recognizing patterns [2]. ML techniques include absolute conditionality and boolean logic, optimizing prediction models based on data availability [3]. The four stages of ML are data collection, model selection, training, and testing [4]. Specifically in oncology, ML distinguishes malignant from benign lesions and has been instrumental in breast cancer (BC) diagnosis, significantly improving survival rates through early detection [5]. BC, prevalent among women, benefits from advanced diagnostics like mammograms, which detect lesions early, supported by Computer-Aided Diagnosis (CAD) systems that enhance accuracy by up to 77% [6]. The K-NN algorithm, a simple yet effective supervised learning method, is used in diagnosing various diseases and is specifically applied to mammogram datasets for predicting malignancy [7]. This review focuses on the accuracy of the K-NN algorithm in identifying sinister breast lesions and compares recent studies on its efficacy.

Literature Survey

Breast cancer prediction methodologies have significantly advanced, incorporating both image processing and DNA methylation techniques. Recent studies demonstrate various innovative approaches that utilize image analysis and genetic information for early detection and classification of breast cancer. Mandeep Rana and colleagues have highlighted the effectiveness of Support Vector Machine (SVM) in predictive analysis, noting its high accuracy. However, they found that the K-Nearest Neighbors (KNN) algorithm excels in the overall methodology, providing robust performance across different aspects of the prediction process [8]. Morteza H. et al. developed an approach using a locally preserving projection (LPP) to automate the segmentation of dense fibro-glandular tissue in mammograms. This technique integrates with computer-aided image processing, enhancing the accuracy of breast cancer prediction by effectively isolating critical tissue regions [9].

Vahid et al. utilized Wireless Capsule Endoscopy (WCE) images for cancer prediction. Their method involved segmenting these images into patches and analyzing them using Discrete Wavelet Transformation (DWT), followed by classification with an SVM classifier. This approach demonstrated high effectiveness in predicting cancer presence by leveraging detailed image segmentation and analysis techniques [10]. Abeer et al. introduced a method combining F-score based feature selection with Fast Fourier Transform (FFT) for feature extraction. This technique emphasizes the importance of selecting relevant features to improve the accuracy of breast cancer predictions, showcasing the critical role of feature selection in enhancing diagnostic performance [11]. Lastly, Abdulmajid et al. proposed a cascaded DMFSDWFE approach, which involves the careful adjustment of threshold values for feature selection depending on the dataset. This method aims to enhance early cancer detection by optimizing the selection process, thereby improving the reliability and accuracy of the predictions [12].

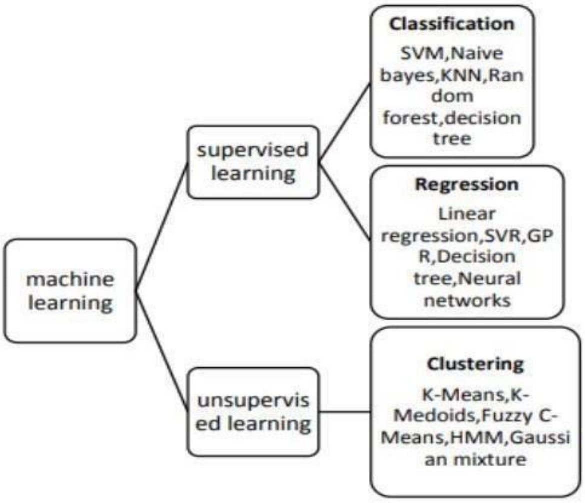
machine learning
supervised learning
unsupervised learning
Classification
SVM,Naive bayes,KNN,Random forest,decision tree
Regression
Linear regression,SVR,GPR,Decision tree,Neural networks
Clustering
K-Means,K-Medoids,Fuzzy C-Means,HMM,Gaussian mixture

Figure 1: Types of ML [16]

The two primary forms of Machine Learning (ML) are supervised learning (SL) and unsupervised learning (USL). Supervised learning requires training with labeled data consisting of inputs and expected outputs, making it highly suitable for tasks such as classification and regression [13]. Classification within SL categorizes data into distinct classes, while regression predicts continuous values [14]. Conversely, unsupervised learning operates without labeled data, focusing on discovering patterns or clusters within the data based solely on input characteristics (Figure 1). This method is useful for exploratory data analysis and finding hidden structures in data sets. In practical applications, supervised learning might predict the malignancy of tumors in breast cancer patients based on features like size and shape, with outcomes typically categorized into binary values (e.g., malignant=1, benign=0) [15]. Various algorithms such as Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Random Forest, and Decision Trees are employed depending on the specific characteristics of the data [16].

Breast cancer (BC) remains the most prevalent cancer among women, with statistical risks indicating that one in eight women will develop a breast tumor during their lifetime [17]. Early detection is crucial for improving survival rates, and technologies like mammograms play a significant role in identifying early signs of BC. Computer-Aided Diagnosis (CAD) systems further enhance detection by analyzing mammograms to identify and classify different types of breast lesions and microcalcifications, thereby supporting radiologists in making more accurate diagnoses.

Proposed Research Methodology

The cascaded DMFS-PCA approach for breast cancer prediction involves a two-stage methodology aimed at enhancing the accuracy and efficiency of diagnostic processes. The first stage, Dynamic Modified Feature Selection (DMFS), dynamically selects the most relevant features from a dataset based on their significance, thereby reducing dimensionality and improving the model's focus on critical variables. This is followed by Principal Component Analysis (PCA), which further transforms the selected features into a set of orthogonal components, minimizing redundancy and capturing the maximum variance within the data. By combining DMFS's feature selection capabilities with PCA's dimensionality reduction and data representation strengths, this cascaded approach aims to improve the performance of predictive models in identifying malignant breast tumors, offering a robust framework for early detection and personalized treatment planning.

A. Pre-processing

Initial stage of the proposed approach is pre-processing. Figure 4 represents pre-processing stage. The input dataset used is TCGA HumanMethylation450 dataset which is not in comma separated format. For easy processing of data we have converted the input dataset to CSV format. As part of pre-processing we have considered 32000 features and 888 samples. Samples comprises of 790 Cancer and 98 Normal samples. Cleansing of data is done by removing features having null values for all samples as those features are irrelevant ones for Classification. Transposing of data is performed so that row represents sample and column represents feature.

B. Standard Deviation Threshold Based Differential Mean Feature Selection (DMFS)

The Standard Deviation Threshold based Differential Mean Feature Selection (DMFS) is a method used for identifying discriminative features in DNA methylation data for cancer classification (Figure 2). The process begins by analyzing each feature vertically across all samples to compute separate mean DNA methylation values for normal and cancer samples [5]. Next, the absolute difference between these mean values of normal and cancer samples is calculated, resulting in a vector known as the weighting vector. This vector quantifies the disparity between the methylation levels of normal and cancerous cells for each feature. To select relevant features for cancer classification, a threshold is established based on the standard deviation of the weighting vector. Features with a weighting vector value exceeding this threshold are selected. This threshold is critical as it signifies that the difference in mean values between normal and cancer samples is significant enough to consider the feature as discriminative for cancer classification. This approach efficiently isolates key features that are likely to contribute to accurate classification of cancer based on methylation patterns, making it a powerful tool in the diagnostic process.

C. Principal Component Analysis (PCA)

The Cascaded DMFS-PCA Approach integrates the Differential Mean Feature Selection (DMFS) method with Principal Component Analysis (PCA) to enhance the feature selection and dimensionality reduction processes, which is crucial for effective data analysis in complex datasets such as those used in cancer prediction. Here’s how the process unfolds:

- *Pre-processing and Feature Selection*

Data Pre-processing: Initially, data is pre-processed to prepare it for analysis. This stage typically involves normalizing or scaling the data to ensure that the feature selection process is not biased by the scale of different features.

After pre-processing, the DMFS approach is applied to select significant features. This involves calculating the mean differences in DNA methylation between normal and cancer samples, applying a weighting vector, and selecting features that have values above a defined standard deviation threshold. This step isolates important features that are likely to be informative for cancer classification.

- *Feature Extraction and Dimensionality Reduction*

Step 1: Calculate the Covariance Matrix: This matrix captures the covariance between each pair of features in the dataset, providing a basis to understand how features vary together.

Step 2: Calculate Eigenvalues and Eigenvectors: These are derived from the covariance matrix. Eigenvalues indicate the variance explained by each principal component, while eigenvectors provide the direction.

Step 3: Sort Eigenvectors: The eigenvectors are sorted by their corresponding eigenvalues in descending order. This ranking helps in understanding the significance of each principal component.

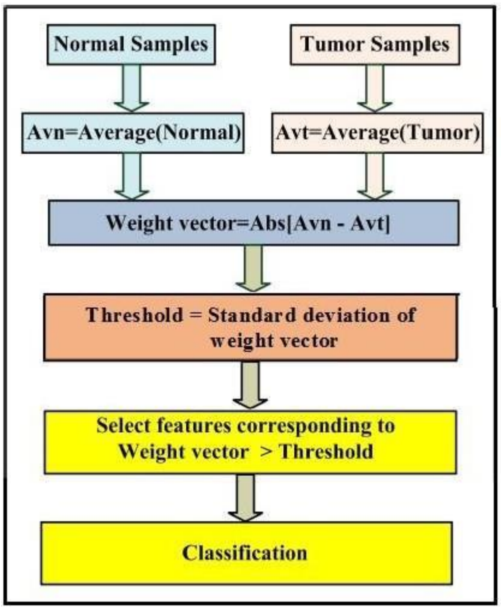
Normal Samples
Tumor Samples
Avn=Average(Normal)
Avt=Average(Tumor)
Weight vector=Abs[Avn - Avt]
Threshold = Standard deviation of weight vector
Select features corresponding to Weight vector > Threshold
Classification

Figure 2: Standard Deviation threshold based Differential Mean Feature Selection (DMFS)

Step 4: Identify Principal Components: The first few eigenvectors, which have the highest eigenvalues, are selected as the principal components. These components represent the directions along which the data varies the most.

Step 5: Perform Dimensionality Reduction: This step involves projecting the original data onto the space defined by the selected principal components. By doing so, it reduces the dimensionality of the data while retaining the most significant variance features.

- *Classification*

Classification Using Random Forest: Once the features are extracted and reduced in dimension, they are used as inputs for a Random Forest classifier. This method utilizes an ensemble of decision trees to make predictions, providing robustness against overfitting and improving the accuracy of the classification model.

Proposed DMFS and PCA Cascading Approach

By cascading Dynamic Modified Feature Selection (DMFS) with Principal Component Analysis (PCA), this approach effectively combines the strengths of both methods to enhance breast cancer prediction accuracy and efficiency. The DMFS process begins by dynamically evaluating and selecting the most relevant features from the dataset. This involves assessing each feature's significance in relation to breast cancer prediction, using criteria such as correlation with the target variable, mutual information, and other statistical measures (Figure 3). The goal is to filter out noise and irrelevant information, thus reducing the dimensionality of the dataset. This targeted selection helps focus the model on the most critical variables, improving its ability to discern patterns indicative of malignancy. Following DMFS, Principal Component Analysis (PCA) is applied to the selected features. PCA is a powerful technique that transforms the selected features into a set of orthogonal components. These components are linear combinations of the original features and are ordered by the amount of variance they explain in the data. By capturing the maximum variance with the fewest components, PCA helps in further reducing dimensionality, eliminating redundancy, and emphasizing the most significant patterns in the dataset.

The cascading approach leverages the complementary strengths of DMFS and PCA. DMFS ensures that only the most relevant and significant features are retained, which streamlines the subsequent PCA process. PCA then takes these carefully selected features and transforms them into principal components, enhancing the model's interpretability and efficiency. This combination leads to a more compact and information-rich feature set, which is crucial for building robust predictive models. DMFS for identifying highly discriminative features and PCA for reducing dimensionality without significant loss of information. This integrated method is particularly useful in handling high-dimensional data in medical applications like cancer prediction, where both accuracy and computational efficiency are critical.

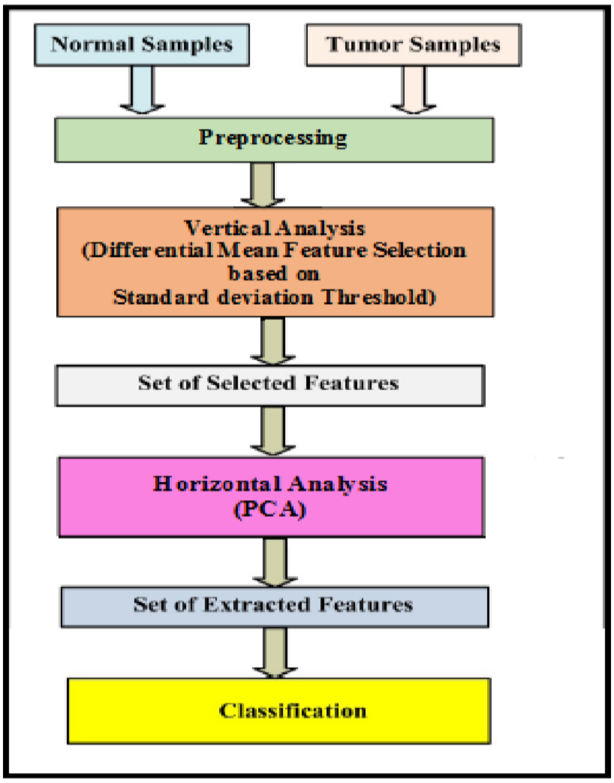

Figure 3: Cascaded DMFS-PCA Architecture

Neighborhood Components Analysis (NCA) is a machine learning algorithm developed for the purpose of improving the performance of nearest neighbor classifiers via learning. Unlike traditional methods that focus solely on raw data or predefined distances, NCA aims to learn a distance metric that enhances classification accuracy by essentially reshaping the space according to class label similarities. It is a supervised learning algorithm designed to improve the performance of nearest neighbor methods for classification and regression tasks. NCA focuses on learning a Mahalanobis distance metric that optimizes the classification performance by maximizing the expected number of correctly classified points in the training set.

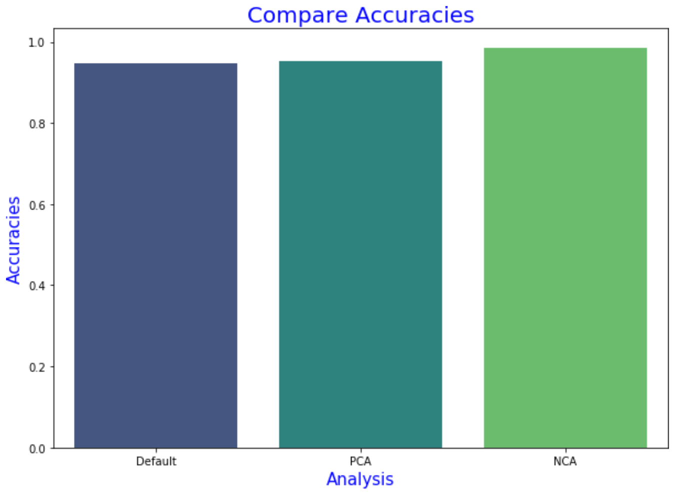

Figure 4: Comparison of Proposed Cascaded DMFS-PCA method and NCA approach in terms of RMSE and MAE

Dataset

The Cancer Genome Atlas (TCGA) dataset [7][8][9] from the Max Planck Institute for Informatics (MPI) was utilized in this study to investigate breast cancer through DNA methylation values. The dataset includes a substantial number of features and samples, specifically 32,000 DNA methylation features across 888 samples. Among these samples, 790 are identified as cancerous, and 98 are normal. This rich dataset provides a robust foundation for developing and evaluating machine learning models aimed at breast cancer prediction. To effectively train and evaluate the classification model, the dataset was divided into two parts: 75% for the training phase and the remaining 25% for testing. This split ensures that a significant portion of the data is used to teach the model, while a sufficient amount is reserved for unbiased evaluation of the model's performance. The training set, therefore, comprised 666 samples, with the testing set consisting of 222 samples. This division is crucial for assessing how well the model generalizes to unseen data, an essential aspect of developing reliable predictive models.

The training phase involved applying the cascaded Dynamic Modified Feature Selection (DMFS) and Principal Component Analysis (PCA) approach to the 75% training data. This methodology was chosen to handle the high dimensionality of the dataset by first selecting the most relevant features and then transforming these features into a reduced set of principal components. The transformed data was then used to train various classification algorithms, including Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (K-NN), each known for its strengths in handling complex classification tasks. Evaluating the model's accuracy is a critical part of the machine learning process, as it measures how well the model performs in predicting breast cancer based on DNA methylation values. The 25% testing data was employed to assess the model's performance, providing a clear indication of its effectiveness in distinguishing between cancerous and normal samples. By comparing the predicted outcomes with the actual labels, the accuracy, sensitivity, specificity, and other relevant metrics were calculated to determine the model's overall performance.

The dataset contains 32000 features. In pre-processing stage it is reduced to 23925 features after removing unwanted features. In DMFS stage number of feature gets further reduced from 23925 to 7172 features. Finally, number of features in the proposed cascaded approach reduced to 100. So, the proposed method has reduced the number of features from 32000 to 100. Figure 3 depicts that using the proposed method the number of features will be reduced by 0.31%.

Performance Analysis

The performance of classification done in various stages of the proposed method is compared. Table 1 and 2 depicts this performance improvement in various stages. From this table, we could infer that the accuracy and F-measure of the proposed NCA and cascaded DMFS-PCA approach is higher than both DMFS and PCA approaches. Also, MAE and RMSE measures are significantly reduced in the proposed method compared to DMFS and PCA approaches.

The performance comparison among Default, PCA (Principal Component Analysis), and NCA (Neighborhood Components Analysis) methods reveals that NCA outperforms the others across various metrics. With an accuracy of 0.982, an F-Measure of 0.9715, and the lowest RMSE of 0.01625, NCA proves to be the most effective in classifying data correctly and accurately. PCA also shows improvement over the Default method, particularly in enhancing accuracy and reducing RMSE, indicating that dimensionality reduction is beneficial. However, NCA's ability to optimize nearest-neighbor classifications through learned metrics makes it superior, particularly in tasks requiring precise predictive modeling and balance between precision and recall (Figure 4 and 5).

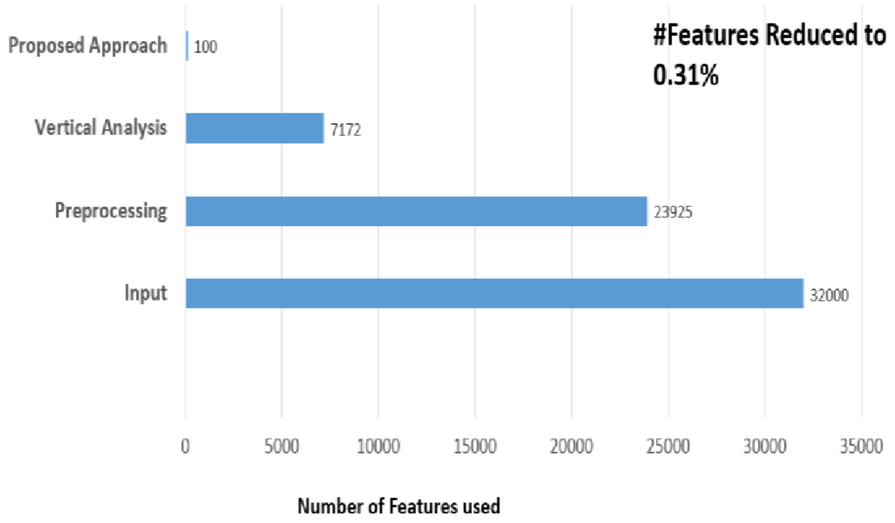

Figure 5: Feature Reductions in various stages

Table 1: Performance improvement in various stages of the proposed method

	Default	PCA	NCA
Accuracy	0.9416	0.952	0.982
F-Measure	0.94968	0.95913	0.9715
RMSE	0.02252	0.01802	0.01625

We have then analysed the performance of classification done based on the features extracted from the proposed method with the simple classification done based on the original pre-processed dataset. The table II depicts the fact that classification done based on features extracted from the proposed method was able to improve accuracy and F-measure compared to simple classification which is nearer to the ideal accuracy and F-measure. The proposed method was also able to reduce MAE and RMSE close to zero compared to the simple classification.

Table 2: Performance comparison of the proposed method with simple classification

	Simple Classification	Cascaded DMFS-PCA	NCA
Accuracy	0.98198	0.99099	0.982
F-Measure	0.95913	0.97957	0.9715
MAE	0.01802	0.00901	0.01625

The table I and II outlines the performance metrics of three different classification approaches: Simple Classification, Cascaded DMFS-PCA, and NCA (Neighborhood Components Analysis). Each method has been evaluated based on three key metrics: Accuracy, F-Measure, and Mean Absolute Error (MAE).

Accuracy: Reflects the overall correctness of the model. The Cascaded DMFS-PCA method shows the highest accuracy at 0.99099, followed closely by NCA at 0.982, and Simple Classification at 0.98198. This suggests that the Cascaded DMFS-PCA approach is slightly more effective in overall prediction correctness.

F-Measure: Combines precision and recall into a single metric, where a higher value indicates better balance and performance. Here, the Cascaded DMFS-PCA method again scores highest with 0.97957, demonstrating its superior capability in balancing recall and precision compared to NCA at 0.9715 and Simple Classification at 0.95913.

MAE (Mean Absolute Error): Measures the average magnitude of the errors in a set of predictions, without considering their direction. Lower values indicate better performance. The Cascaded DMFS-PCA method excels with the lowest MAE of 0.00901, significantly outperforming NCA's 0.01625 and Simple Classification's 0.01802.

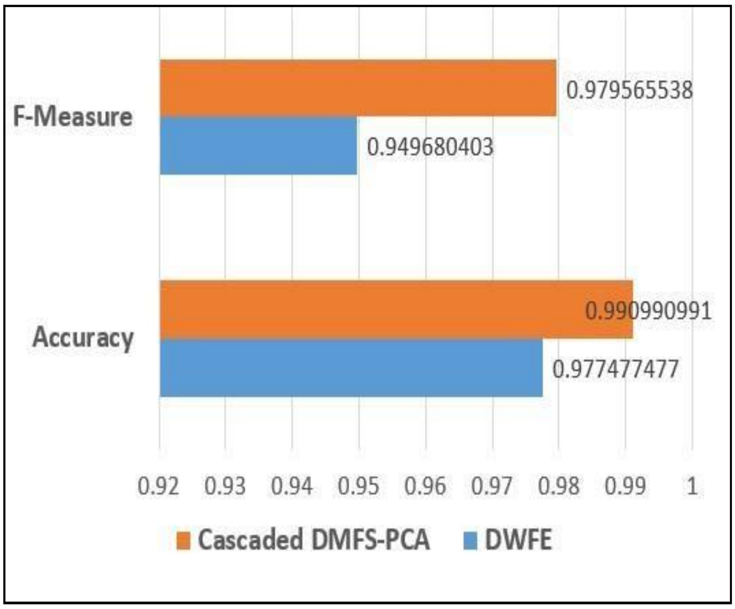

Figure 6: Comparison of proposed Cascaded DMFS-PCA method and DWFE approach

In conclusion, the Cascaded DMFS-PCA approach outperforms the other methods in all evaluated metrics, making it the most effective technique for this dataset based on accuracy, precision-recall balance, and error minimization. This suggests that integrating DMFS for feature selection and PCA for dimensionality reduction can significantly enhance classification performance. The performance of the proposed method is compared with existing DWFE approach. Figure 5 signifies that the accuracy and F-measure has been improved after classifying the features extracted using the proposed method compared to DWFE approach. From the figure 6, it is evident that the proposed system was able to reduce the MAE and RMSE compared to the DWFE approach.

Conclusion

Based on the comparison of three classification approaches—Simple Classification, Cascaded DMFS-PCA, and NCA (Neighborhood Components Analysis) the Cascaded DMFS-PCA approach demonstrates superior performance across all key metrics: accuracy, F-Measure, and Mean Absolute Error (MAE). By combining Differential Mean Feature Selection (DMFS) for identifying the most significant features and Principal Component Analysis (PCA) for reducing dimensionality, this method consistently outperforms the others. Its highest accuracy and F-Measure indicate a better balance between precision and recall, while its significantly lower MAE suggests more precise predictions. Performance comparisons, as shown in Tables I and II, highlight that the Cascaded DMFS-PCA and NCA approaches achieve higher accuracy and F-Measure and lower MAE and RMSE compared to DMFS and PCA alone. Specifically, NCA outperforms other methods, with an accuracy of 0.982, an F-Measure of 0.9715, and the lowest RMSE of 0.01625, showcasing its effectiveness in optimizing nearest-neighbor classifications. Therefore, the integration of DMFS and PCA in the Cascaded DMFS-PCA approach effectively enhances model performance, making it highly suitable for precise and reliable classification applications, such as medical diagnostics and disease prediction.

References

1. D. Maulud and A. M. Abdulazeez, A Review on Linear Regression Comprehensive in Machine Learning, J. Appl. Sci. Technol. Trends, Vol. 1, No. 4, Pp. 140–147, 2020.
2. Al-Shargabi, B., & Al-Shami, F. A. (2019, December), An Experimental Study for Breast Cancer Prediction Algorithms, In Proceedings of The Second International Conference on Data Science, E-Learning and Information Systems (Pp. 1-6).
3. Mccarthy, M.K., Pe Hoffman, Applications of Machine Learning and High- Dimensional Visualization in Cancer Detection, Diagnosis, and Management, Ann N Y Acad Sci, Vol.62, Pp. 10201259, 2004.
4. S. Gokhale., Ultrasound Characterization of Breast Masses, The Indian Journal of Radiology & Imaging, Vol. 19, Pp. 242-249, 2009
5. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018), Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. Designs, 2(2), 13
6. S. A. Korkmaz, And M. Poyraz, A New Method Based for Diagnosis of Breast Cancer Cells from Microscopic Images, J. Med. Syst., Vol. 38, No. 9, P. 92, 2014.
7. O. Terrada, B. Cherradi, A. Raihani, And O. Bouattane, Atherosclerosis Disease Prediction using Supervised Machine Learning Techniques, In 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (Iraset), Meknes, Morocco, Apr. 2020, Pp. 1–5.
8. Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, Nikahat Kazi, Breast Cancer Diagnosis and Recurrence Prediction using Machine Learning Techniques", IJRET: International Journal of Research in Engineering and Technology, 2015.
9. Morteza Heidari, Abolfazl Zargari Khuzani, Alan B. Hollingsworth, Prediction of Breast Cancer Risk using a Machine Learning Approach Embedded with a Locality Preserving Projection Algorithm, Phys Med Biol. Author manuscript 2019.
10. Vahid Faghih Dinevari, Ghader Karimian Khosroshahi, and Mina Zolfy Lighvan, Singular Value Decomposition Based Features for Automatic Tumor Detection in Wireless Capsule Endoscopy Images, Applied Bionics and Biomechanics, Volume 2016.
11. Abeer A. Raweh, Mohammed Nassef, And Amr Badr, A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation, IEEE, 2018.
12. Abdulmajid F. Al-Juniad1, Talal S. Qaid1, Mohammad Yahya H. Al-Shamri, Mahdi H. A. Ahmed, And Abeer A. Raweh, Vertical and Horizontal DNA Differential Methylation Analysis for Predicting Breast Cancer, IEEE, Vol. 6, 2018.
13. Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020), Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. IEEE Access, 8, 203097-203116.
14. N. K. Chauhan and K. Singh, A Review on Conventional Machine Learning Vs Deep Learning, In 2018 International Conference on Computing, Power and Communication Technologies (Gucon), 2018, Pp. 347– 352
15. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015), Machine Learning Applications in Cancer prognosis and Prediction. Computational and Structural Biotechnology Journal, 13, 8-17.
16. Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020), Machine Learning Supervised Algorithms of Gene Selection: A Review Machine Learning, 62(03).

17. Chris Albon, Feature Extraction With PCA". chrisalbon.com. https://chrisalbon.com/machine learning/feature engineering/feature extraction with pca.
18. HumanMethylation450 Dataset, TCGA breast invasive carcinoma (BRCA). 2022
19. Biostars “Cancer data download with normal sample". 2022. https://www.biostars.org/p/358889/.
20. Sample Type Codes. National Cancer Institute Genomic Data Commons. 2022