JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 02
2024
Research Article
https://doi.org/10.56979/702/2024

Enhancing Breast Cancer Diagnosis with Integrated Dimensionality Reduction and Machine Learning Techniques

Aqeel Ahmed Khan[1], and Muhammad Abu Bakr[1] *

[1]Electrical Engineering Department, National University of Technology (NUTECH), Islamabad, Pakistan.
*Corresponding Author: Muhammad Abu Bakr. Email: muhammadabubakr@nutech.edu.pk

Received: May 01, 2024 **Accepted:** August 21, 2024 **Published:** September 01, 2024

Abstract: Breast cancer remains a significant cause of cancer-related mortality worldwide, highlighting the critical need for advancements in diagnostic techniques. Recent diagnostic methods, while effective, often face limitations in accuracy and efficiency. This paper aims to differentiate between tumorous (malignant) and non-tumorous (benign) cases of breast cancer using three publicly available datasets: Wisconsin Breast Cancer (WBC), Wisconsin Diagnostic Breast Cancer (WDBC), and Wisconsin Prognostic Breast Cancer (WPBC) datasets. We applied popular supervised machine learning classifiers, including Multi-layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), and K-Nearest Neighbor (KNN), in combination with dimensionality reduction techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Factor Analysis (FA). The classifiers were evaluated based on accuracy, precision, recall and F1 score. The results show that, due to FA's emphasis on feature selection and noise reduction, the SVM with FA achieved the highest accuracy of 98.64% on the WBC dataset. MLP without any dimensionality reduction performed best with an accuracy of 98.26% on WDBC. Conversely, MLP and SVM with LDA yield an accuracy of 89.80% on the more complex and noisy WPBC dataset.

Keywords: Breast Cancer Classification; Machine Learning; Multi-layer Perceptron; Dimensionality Reduction.

1. Introduction

Cancer is a disease that causes uncontrolled development and spread of abnormal cells in the body which have the ability to form tumors, attack surrounding tissues, and transit through various parts of the body via circulation of blood and lymphatic system [1]. Breast cancer generally begins with a lump or change in the appearance of breast tissue [2], and is the most common cancer affecting women worldwide, particularly those over the age of 50 [3].

Breast cancer can occur either with malignant or benign tumors. Although benign tumors are more painful, it is the malignant tumors that pose a greater risk to patients due to their potential to spread to other cells in the body. Therefore, early tumor diagnosis is crucial to ensure that a patient suffering carcinoma of the breast receives the appropriate treatment [4].

Screening for new tumor biomarkers is still an important task in order to improve diagnostic rates, and multiple markers or a super new marker may be required to obtain definitive diagnosis results [5]. Whereas the imaging devices will continue to be the usual strategy for monitoring breast cancer in the near future, new markers have the potential to improve throughput, speed, sensitivity, and specificity, as well as evaluate treatment efficacy and assess different types of breast cancer [8].

Artificial Intelligence (AI) has rapidly transformed the domain of automated breast cancer detection, with research showing promising results when compared to conventional CAD-e/CAD-x methods [6] [7]. Recent research indicates that AI algorithms surpass humans in retrospective data sets, with more mature and commercial products now available. Whereas more research is required, it is apparent that AI will be

the key to the monitoring of breast cancer in the future, and studies are currently evaluating various implementation options [9] [10].

Datasets are essential for the development and testing of algorithms for classification and prediction tasks in the field of Machine Learning (ML). Several publicly available datasets have been utilized for breast cancer research, including the Wisconsin Breast Cancer (WBC), Wisconsin Diagnostic Breast Cancer (WDBC), and Wisconsin Prognostic Breast Cancer (WPBC) datasets. All the three datasets have been extensively studied for classification using various machine learning classifiers. This paper provides an extensive evaluation of machine learning integrated with dimensionality reductions methods for breast cancer diagnosis and prediction. This paper makes the following key contributions:

- Extensive performance evaluation of various ML classification models across three different breast cancer datasets.
- Assessing the impact of integrating dimensionality reduction with supervised learning on breast cancer classification, highlighting improvements in model accuracy and efficiency.
- Identifying optimal classifier-dimensionality reduction pairs, with SVM-FA achieving the highest accuracy of 98.64% on WBC dataset at a contamination level of 0.2, proving its ability to deal with outliers.

The paper is structured as: Section 2 discusses the related literature while section 3 describes the material and methods. The results and discussion on research are described in section 4. Section 5 provides the conclusion.

2. Literature Review

Various studies have investigated the use of ML and dimensionality reduction methods to improve the accuracy and efficiency of breast cancer diagnosis. For instance, in [11] the authors utilized PCA-based dimensionality reduction techniques with classifiers such as Support Vector Machine (SVM,) K-Nearest Neighbor (KNN), Decision Tree (DT) etc. to identify malignant and benign tumors in WBC dataset. The study compared the accuracy achieved by all these classifiers with the sigmoid based Naive Bayes (NB) achieving an accuracy of 99.20%. A comparative study of Random Forest (RF), KNN and NB for breast cancer was conducted in [12], where the performance of the algorithms was examined based on various performance parameters. KNN was identified as the most effective algorithm achieving an accuracy of 95.90%.

In [13], a comparative analysis was conducted among four ML algorithms – SVM, Logistic Regression (LR), NB, and RF - utilizing the WBC dataset. Experimental results revealed that RF surpassed its counterparts, achieving an exceptional accuracy of 99.76% and minimizing error rates. All experiments were meticulously executed within a simulated environment on the ANACONDA Data Science Platform. In [14], the authors categorized the research into three areas: cancer prediction, diagnosis and treatment prediction, and outcome prediction. Four ML algorithms, namely LR, DT, RF, and SVM, were used with DT method achieved 100% accuracy [14]. A maximum voting based Ensemble Learning (EL) method using the predictions from KNN, SVM, and RF was presented in [15]. The method performed quite well with an accuracy rate of 98.9%.

Numerous features associated with the WBDC dataset was narrowed down to 17 features using 5 feature selection techniques [16]. Of the classifiers tested which included MLP, SVM, stack classifiers the SVM approach gave the best performance of 97.7% accuracy. In the literature [17], ANN, SVM, and KNN were reported to be the most common algorithms for cancer diagnosis. These methods also proved to have a better performance with average accuracies reaching between 83.45% and 99.30%.

Several researchers have analyze the performance of various ML models publicly available datasets for breast cancer detection where SVM and RF models achieved high accuracy of 96.5% [18]. Some researchers have also used Deep Learning (DL) techniques like Convolutional Neural Networks (CNNs) and Artificial Neural Networks (ANNs) for breast cancer diagnosis. For instance, a unique DL approach combined neural networks with optimal feature selection for classification of WBC dataset and achieved an accuracy of 99.67% [19]. A study utilizing 10-fold cross-validation on the WBC data demonstrated impressive outcomes, with SVMs reaching an accuracy of 98.57% and a specificity of 95.65%. In comparison, KNN achieved an accuracy of 97.14% and a specificity of 92.31% [20].

In [21], a comparative study was carried out on various ML algorithms utilizing the WBC dataset, revealing that SVM emerged as the most precise classifier, achieving an accuracy of 97.13%. Additionally, a new methodology called "SSL Co-training" was introduced in [22], which employs semi-supervised learning and pseudo-labeling techniques to forecast breast cancer survivability. This approach makes use of unlabeled patient data and assigns virtual labels, effectively tackling the issue of scarce labeled data. The method attained a mean accuracy of 76% and an AUC of 0.81. Furthermore, a comparative evaluation of five ML algorithms was conducted on the Breast Cancer Wisconsin Diagnostic Dataset (BCWDD), with SVM reaching the highest accuracy of 97.2% [23, 24].

Dimensionality reduction (DR) techniques like Principal Component Analysis (PCA) can make ML models more efficient by simplifying data. A recent study [25] employed PCA with ML for breast cancer diagnosis, aiming to improve accuracy. The researchers compared KNN, LR, and Ensemble Learning (EL) methods with EL outperforming the others by achieving an impressive accuracy of 99.30%. This combined PCA and ML approach shows great promise for enhancing breast cancer detection, potentially enabling earlier diagnosis and treatment. Other ML algorithms such as KNN and LR achieved an accuracy of 98.60% and 97.90%. A novel approach combining NB, Chi-squared attribute selection, and extended Kernel principal component analysis (K-PCA) with a sigmoid kernel is proposed in [26]. This integrated method achieved a high accuracy of 99.28% in breast cancer cell classification. The sigmoid K-PCA and feature selection contributed to its exceptional performance, outperforming recent state-of-the-art studies. In [27], RF, LR, Xtreme Gradient and Ada-Boost Classifiers are trained on BCWD, and their efficiency is evaluated by comparing this paper using ensemble classifier and ML techniques. The aim of the study was to figure out the most efficient ensemble and ML classifier for breast cancer diagnosis based on accuracy and AUC score.

Another study [28] worked on enhancing the breast cancer recurrence prediction by addressing dataset challenges including: (a) applying data oversampling to tackle class imbalance, and (b) using PCA and a Genetic Algorithm (GA)-based dimensionality reduction technique to manage excessive data elements. It also integrates outputs from RF and SVM classifiers using neural networks. The proposed framework achieved notable improvements, with an accuracy of 98.3%, AUC of 99%, and precision, recall, and F1 scores of 98%. The related work is compiled in Table 1 below.

Table 1. Literature Review

References	Model	Dataset	Year	Accuracy
[11]	NB(Sigmoid)	WBC	2019	99.20%
[12]	KNN	WDBC	2018	95.90%
[13]	RF	WBC	2019	99.76%
[14]	DT	WBC	2020	100.00%
[15]	SVM	WBC	2016	----
[16]	SVM	WDBC	2023	97.70 %
[17]	SVM	WDBC	2023	83.45%
	KNN			99.30%
[18]	ANN	WBCD	2020	99.30%
[19]	CNN	WBCD	2018	99.67%
[20]	KNN	WBC	2017	98.57%
[21]	SVM	WBC	2016	97.13%
[22]	SSL Co-training	Breast cancer survivability dataset (1973–2003)	2013	76.00%
[23]	SVM	BCWG	2011	97.20%
[24]	SVM	BCWD	2022	97.20%
[25]	Ensemble Learning	WDBC	2023	99.30%
[26]	Sigmoid K-PCA	WBC	2022	99.28%
[27]	LR	BCWD	2023	96.49%
[28]	RF	WDBC	2024	98.30%
	SVM			99.00%

3. Materials and Methods

3.1. Datasets Description

The WBC [29] dataset contains information about 699 instances with 10 attributes and one target variable. Based on these characteristics, the data aims to classify whether the cell is 'benign or malignant'. The distribution of the class variable consisted of 65.5% of the cases falling into the "malignant" category, and 34.5% falling into the "benign" category. Missing values are indicated in the data by the symbol "?" in the bare nuclei attribute. This dataset has been utilized in numerous studies to construct and assess predictive models and is frequently used for teaching and research purposes.

The WDBC dataset contains information about breast tumor samples collected through fine needle aspiration technique. The dataset contains a total of 32 attributes for each of the 569 instances. The diagnosis column has two classes: "M" for malignancy and "B" for benignity. Out of the total 569 instances, 357 cases are classified as benign and 212 as malignant. Dataset also includes the means, standard deviations, and maximum values for each of the 10 attributes. The final 10 attributes correspond to simple ID numbers and do not provide any useful information for classification. The WDBC dataset is regarded as a high-quality dataset since there are no missing values. The dataset is often utilized for diagnosis of breast tumors based on offered variables.

The WPBC dataset was also collected from fine needle aspiration samples of breast tumor masses, and was used to predict patient outcomes. It consists of data from 198 patients, and includes 34 features. The outcome variable is recurrence status, with "R" indicating recurrence and "N" indicating no recurrence. The WPBC dataset contains both clinical and demographic information about the patients, as well as features describing the properties of the tumor itself. Some of the features include tumor size, shape, and texture, as well as the subject's age, menstruation state, and estrogen receptor status. Dataset also includes information about the patient's treatment, such as radiation therapy and chemotherapy. Like the WDBC dataset, the WPBC dataset does not have any missing data. However, some of the features in the WPBC dataset are unnamed and do not have clear descriptions, which makes the dataset more challenging to work with.

Table 2(a). WBC Dataset

Dataset Features	Range of Values
Clump Thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
Marginal adhesion	1-10
Single epithelial cell size	1-10
Bare nuclei	1-10(missing value'?')
Bland chromatin	1-10
Normal nucleoli	1-10
Mitoses	1-10
Class	2 for 'Benign', 4 for 'Malignant'

Table 2(a) provides the description of the WBC dataset. The term "Clump Thickness" refers to the thickness of the cell clump inside the body. The variation in cell sizes inside a mass is referred to as "uniformity of cell size". The variance in the shape of the mass's cells is considered to be "uniformity of cell shape". The degree whereby the cells adhere to the other cells near the mass's edge is determined by "marginal adhesion". The size of a "Single epithelial cell" is used to describe the mass of cells. "Bare Nuclei" describes the appearance of the nuclei of cells in the mass. The uniformity of the chromatin (genetic material) in the cells is referred to as "bland chromatin". "The nucleoli" (structures within the nuclei) in cells are described as normal in terms of size and shape. "Mitoses" signifies the number of mitotic figures in a mass, where the range of values for each characteristic is between 1 and 10.

Table 2(b). WDBC Dataset

Dataset Features	Range of Mean	Range of Standard error
Radius	6.981 - 28.11	0.1121 - 2.873
Texture	9.71 - 39.28	0.3602 - 4.885

Perimeter	43.79 - 188.5	0.757 - 21.98
Area	143.5 – 2501	6.802 - 542.2
Smoothness	0.05302 - 0.1634	0.002 - 0.01
Compactness	0.01938 - 0.3454	0.00225 - 0.1354
Concavity	0 - 0.4268	0 - 0.2
Concave Points	0 - 0.2012	0 - 0.053
Symmetry	0.106 - 0.304	0.0079 - 0.0789
Fractal Dimension	0.04996 - 0.09744	0.0009 - 0.03

Table 2(b) describes the WBDC dataset. Radius Mean describes the average distance across the cell nucleus's center and its perimeter locations. The term "texture mean" describes the gray-scale values in the standard deviation of the cell nucleus image. The average perimeter of the cell nucleus is described by the perimeter mean. The average size of the cell nucleus is described by Area Mean. Smoothness The range of cell nucleus radius lengths is described by the mean. Compactness the mean, defined as the perimeter squared divided by the area less one, describes the compactness of the cell nucleus. Concavity The severity of concave regions of the cell nucleus is described by mean. The cell nucleus's concavity is expressed as a number by the term mean. The symmetry of the cell nucleus is described by Symmetry Mean. Dimension of fractals. The cell nucleus's "coastline approximation," which measures its complexity, is described by means.

Table 2(c). WPBC Dataset

Dataset Features	Minimum	Maximum	Range
Radius	7.760	36.0400	28.2800
Texture	10.380	39.2800	28.9000
Perimeter	47.980	251.2000	203.220
Area	170.400	4254.000	4083.60
Smoothness	0.052	0.1634	0.1114
Compactness	0.019	0.3454	0.3264
Concavity	0.000	0.4275	0.4275
Concave points	0.000	0.2012	0.2012
Symmetry	0.106	0.3040	0.1980
Fractal dimension	0.050	0.0974	0.0474

Table 2(c) provides the description of the WPBC dataset. The radius, perimeter, and area of the mass include information about the size and shape of the mass. “Texture, smoothness, compactness, concavity, and concave points” provide insight into the structural properties of the mass. “Symmetry” and “fractal dimension” are additional measures that may provide useful information about the mass. “Lymph node” is an important indicator of the extent of breast cancer and may inform treatment decisions. The number of positive axillary lymph nodes can help predict the likelihood of metastasis and may inform the need for additional treatments such as chemotherapy or radiation. “Tumor size” is also an important variable as it can inform the extent and severity of the cancer. Larger tumors may indicate a more advanced stage of the cancer and may require more aggressive treatment approaches.

3.2. Proposed Framework

Figure 1 illustrates the outline of our research work. Different stages of the framework are discussed in detail below.

3.2.1. Data Preprocessing

Preprocessing is the preliminary step to manage missing values to reduce bias and improve result accuracy. The WBC dataset contains dummy/missing values in the 'Bare Nuclei' column, personified by the symbol '?'. Omitting these values results in loss of information, hence an unsuitable approach. A suitable technique is to substitute values that are not present with the mean value of corresponding features. To facilitate analysis of the dataset, the Sample ID column should be removed before cleaning the WBC dataset, as it does not offer any helpful information.

3.2.2. SMOTE Sampling

The Synthetic Minority Over-Sampling Technique (SMOTE) [30] is a prevalent machine learning method for addressing class imbalance by producing artificial samples of the minority class instead of

solely replicating the current samples. To create novel samples, the approach interpolates between current minority class samples. The objective is to equalize the distribution of the classes and minimize the over-representation of the dominant class in the training dataset. The SMOTE technique uses the "fit-resample" method to generate replicas of the minority class that subsequently fit the SMOTE model to the input data. This approach returns both target variable, target, and resample data to balance the class distribution. Figure 2, 3 and 4 depict the SMOTE sampling of WBC, WDBC and WPBC datasets features respectively.

3.2.3. Outlier Detection

Outlier identification is a frequently used technique to find such instances that considerably differ from the rest of the data points, which have the potential to alter the overall evaluation. Machine learning-based approaches using isolation forest have gained popularity for their ability to identify outliers in a given dataset. Before pre-processing a dataset, the Isolation Forest [31] is often used, with the target of removing outlier applicants. It is an unsupervised outlier detection method that is ensemble-based with precision and linear time complexity. The forest is formed by an assortment of binary trees that were assembled using the dataset's random property. Then, stroll through the forest, estimating the anomaly score of each data point in each tree. This ensemble method involves generating multiple decision trees and scoring each data point based on its distance from the root of the tree, thereby isolating outliers with a unique score. The contamination parameter, which measures the proportion of outliers in the dataset, is initialized to 0.05, and the model is tailored to the data using the fit ( ) method. The Numpy technique is used to locate the indices of the dataset's outliers based on their scores, which can be used to isolate them.

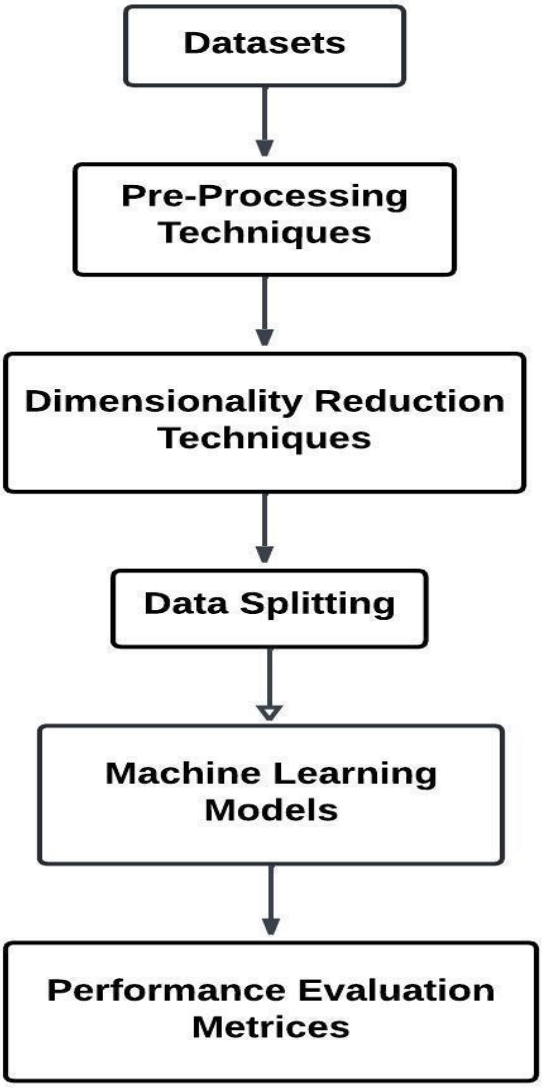
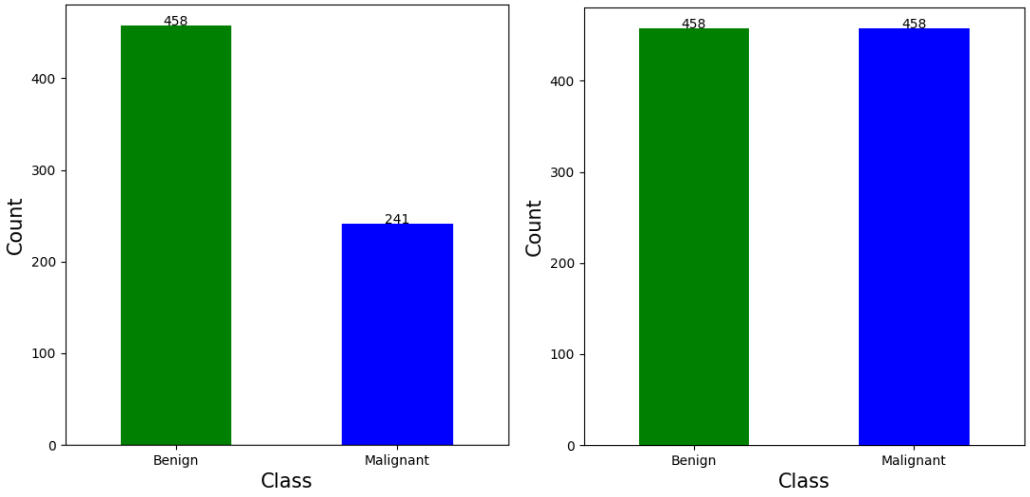
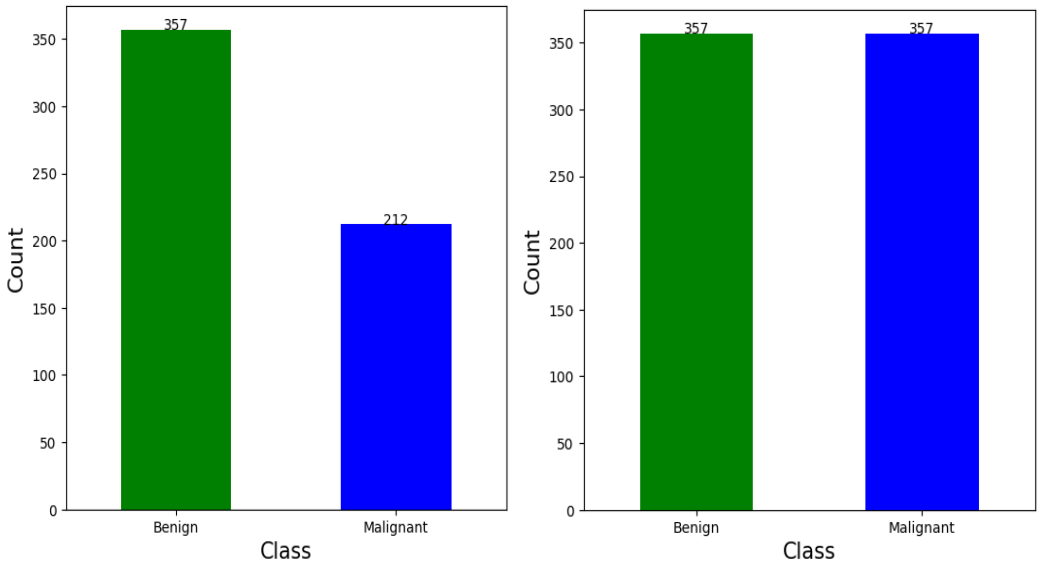


Figure 1. The proposed framework of our research work



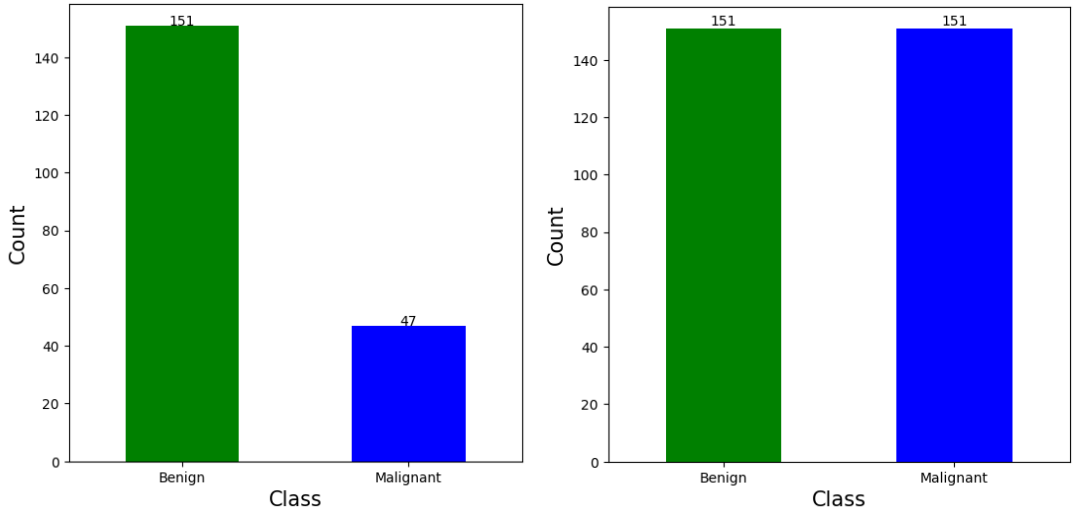
(a) (b)

Figure 2(a) (b). Before and After SMOTE Sampling of WBC



(a) (b)

Figure 3(a) (b). Before and After SMOTE Sampling of WDBC



(a) (b)

Figure 4(a) (b): Before and after SMOTE Sampling of WPBC

3.2.4. Dimensionality Reduction

Dimensionality reduction (DR) is a technique that involves transforming a large set of high-dimensional data into a lower-dimensional space while maintaining the fundamental form of the original data [32]. This can be accomplished through feature extraction, which guides the unique attributes to a new set of features, or feature selection, which maintains certain aspects of the original features. The main aim of dimensionality reduction is to reduce the dataset's complexity and size while trying to retain as much information as possible in order to improve computational efficiency and facilitate data visual display, exploration, and analysis. Dimensionality reduction is widely used in research fields, including machine learning, signal processing, computer vision, and data mining. Different DR methods used in this work are discussed below.

3.2.4.1. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a method used to decrease the dimensionality of datasets, which improves both performance and computational efficiency [32]. LDA operates under the premise that the mean values of each attribute differ across the target classes. This technique converts the original feature set into a new one that preserves a significant amount of information while optimizing the separation between the target classes. It computes the covariance matrix and mean vectors for each class within the dataset, then determines the linear combinations of features that enhance the differentiation between the target classes. The dataset is divided into two exclusive subsets, with 20% of the instances designated for testing and the remaining 80% used for training. LDA aims to generate a single feature from a linear combination of the original features to effectively separate the two breast cancer classes in the WBC dataset. Different performance parameters have been utilized to examine the performance of the LDA technique, which provide an estimate of its ability to accurately classify instances of breast cancer into their respective classes. These parameters include precision, recall, accuracy and F1 score. Figure 5 (a), (b), and (c) represent the LDA results for two classes, 0 (benign) and 1 (malignant) for the three datasets, using a single component. In this case, LDA struggled due to insufficient variance between the classes, resulting in overlap. The high number of features and noise further hindered LDA's ability to achieve optimal separation.

3.2.4.2. Principal Component Analysis

PCA reduces the dimensionality of datasets, aiming to preserve as much of the original data variation as possible. PCA allows us to reduce the dataset's dimension by identifying the principal components, which are linear combinations of the initial traits that capture the most data variance [32]. This led to the dataset's simplification and potentially improved the effectiveness of a model developed using ML by lowering the probability of overfitting and reducing computing complexity. Identifying the primary components led to feature selection, which can increase the performance of a ML model. Figure 6 (a), (b), and (c) depict the results of PCA for two classes, 0 (benign) and 1 (malignant) of the three datasets. In Figure 6 (a) and (b), the classes are clearly separated with minimal overlap, indicating that PCA successfully identified a linear transformation that enhances class reparability. However, in Figure 6 (c), overlap occurs due to noise in the data, which hinders PCA from capturing enough information to fully distinguish between the classes.

3.2.4.3. Factor Analysis

Factor Analysis (FA) is a method of classifying causal factors that may be affecting observed variation in a dataset. FA can be performed to minimize the dataset features by identifying latent factors that seize the significant number of the diversity in the data [32]. The dataset becomes easier to interpret as the number of instances reduces. This in turn also reduces the possibility of over-fitting. FA helps to enhance the evaluation of a ML model by identifying factors that capture the most of the inequality in the data. This is due to the model being trained on a smaller set of more informative instances. Figure 7 (a), (b), and (c) show the results of FA for two classes, 0 (benign) and 1 (malignant), using a single component. The data points are clustered along the y-axis at 0, indicating FA's inability to separate the classes. Similar to LDA, proper data scaling and normalization are crucial. This suggests that FA fails to recognize any variance between the classes due to improper data scaling.

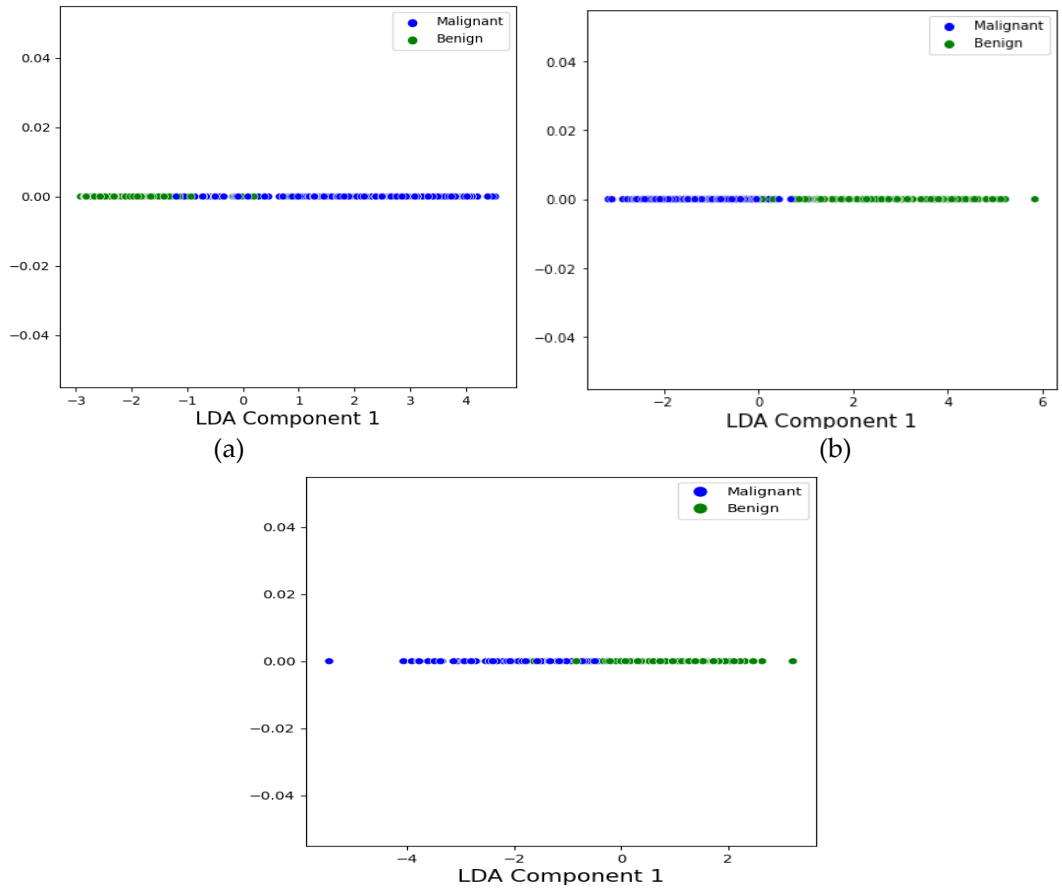


Figure 5 (a) (b) (c). LDA of WBC, WDBC & WPBC

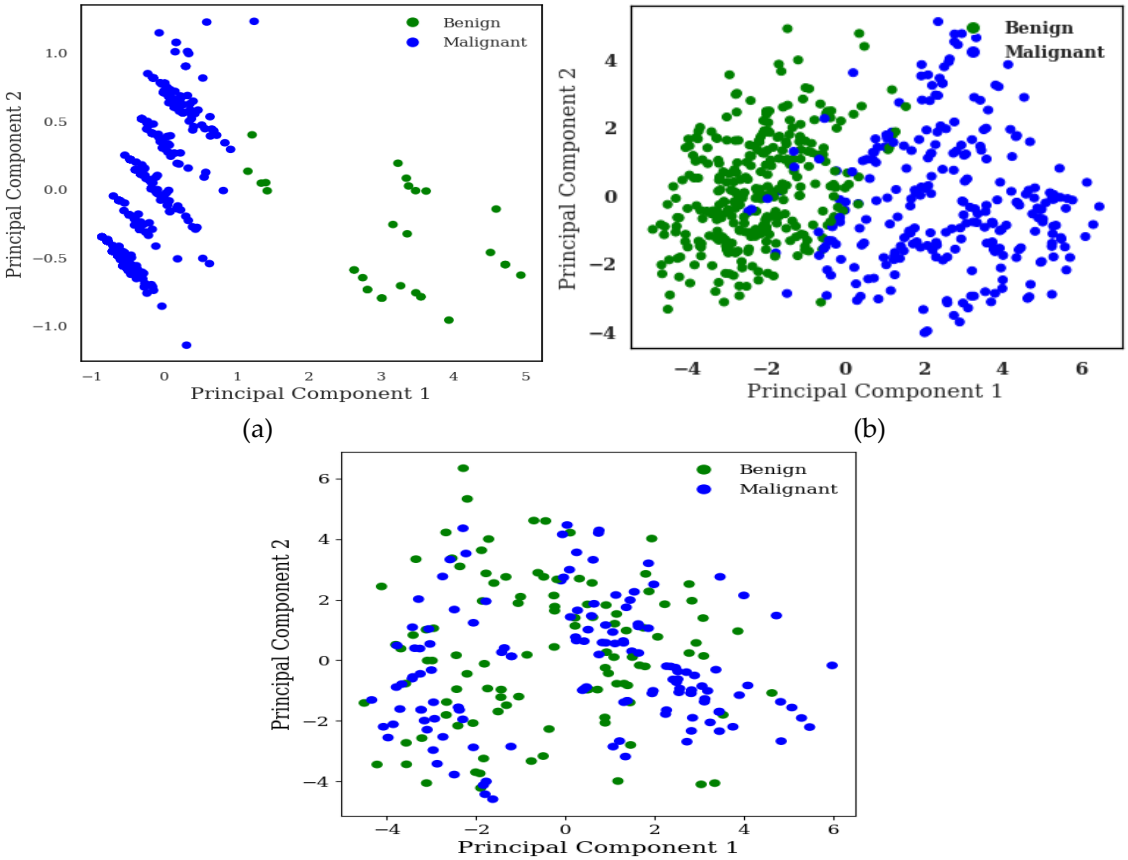


Figure 6 (a) (b) (c). PCA of WBC, WDBC, WPBC

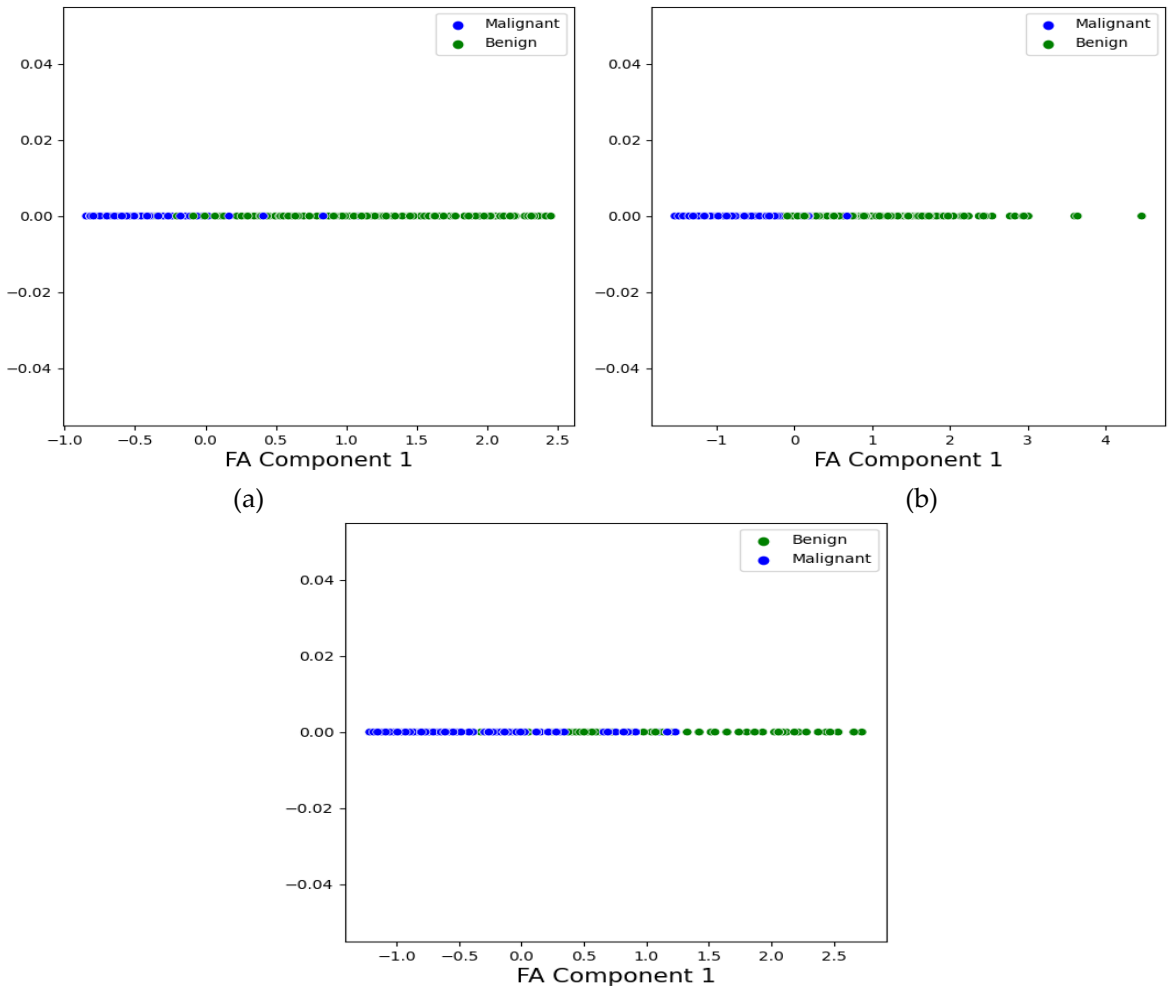


Figure 7 (a) (b) (c): FA of WBC, WDBC, and WPBC

3.2.5. Classifiers

This section introduces various ML techniques employed in this study.

3.2.5.1. Multi-Layer Perceptron

MLP is an artificial neuron that executes particular mathematical operations to recognize all of the instances in a dataset. MLP works with further perceptual neurons to deal with difficult issues. This combination is known as a multi-layer perceptron MLP or ANN [33]. MLP employs a back-propagation model to alter the biases of connection between nodes in sequence to minimize the changes between predicted and actual model outcomes.

3.2.5.2. Support Vector Machine

This is a robust automated learning model that is utilized primarily for classification purposes. The algorithm generates a hyperplane for splitting data into classes and optimizes the distance between it and the nearest points in each class [34]. SVM excels with high-dimensional data where other algorithms may face challenges.

3.2.5.3. Random Forest

Random Forest (RF) is a ML method that integrates several decision trees to generate predictions. This technique entails building numerous decision trees, each trained on distinct subsets of the dataset while utilizing a random assortment of features. It is particularly effective for classification and regression problems and is recognized for its resilience to outliers and noisy data.

3.2.5.4. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a versatile technique applicable to both classification and regression problems [36]. The method operates by locating the k nearest data points in the training set relative to a specified input sample, subsequently utilizing the labels of these neighboring points to forecast the label for the new data instance. KNN is straightforward to implement and demonstrates effective performance

with smaller datasets; however, its computational demands increase significantly as the size of the dataset expands.

3.2.5.5. Decision Tree

Decision Tree (DT) is frequently used for both classification and regression purposes. It segments the data according to the attributes of the input features and systematically creates sub-trees until all data points in each sub-tree are classified into the same category or show comparable predicted results [37]. Although DT is user-friendly and capable of managing non-linear relationships among input features, it can be prone to overfitting if the tree's complexity increases significantly.

3.2.6. Performance Metrics

Performance metrics are quantitative measures that measure how effectively a ML model performs on a specific dataset. These measures help to analyze the model's efficiency, accuracy, and effectiveness in making predictions or classifications. Performance measures vary according to the type of model (classification, regression, etc.). However, these are some of the metrics used in this paper:

1. Accuracy determines the model overall's correctness and is calculated as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

2. Precision is calculated as the fraction of accurate predictions among all positive predictions made by the model:

$$\text{Precision} = TP / (TP + FP)$$

3. Recall is the proportion of truly positive predictions among all real positive observations in the dataset:

$$\text{Recall} = TP / (TP + FN)$$

4. F1 Score represents the weighted harmonic mean of precision and recall. It maintains an appropriate mix between precision and recall:

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

4. Results

In this section, we present the detailed results of our experimentation. The results of different ML algorithms on WBC, WDBC and WPBC are depicted in Figure 8 (a), (b) and (c) respectively. The table further shows the impact of dimensionality reduction on the performance of various techniques.

Figure 8(a), shows performance of various classifiers on the WBC dataset, both with and without dimensionality reduction. MLP, RF, KNN classifiers without DR demonstrated best performance among models, all achieving 97.28% accuracy. When applied with DR models such as PCA, the SVM classifier showed notable improvement, achieving the highest accuracy of 97.96%. PCA improved the performance of SVM and DT models, while MLP, RF and KNN remained the same in terms of accuracy. Meanwhile, LDA reduced the performance of KNN, while improvement was noted in the accuracy of MLP and SVM. In the case of FA, the SVM classifier accuracy improved to 98.64%. FA provided a performance boost for DT but decreased the effectiveness of RF and KNN to 96.60% and 91.84% respectively. The integration of outlier detection using Isolation Forest with a contamination rate of 0.2 contributed to the robustness of the results, particularly by enhancing the accuracy and reliability of most classifiers. These findings highlight the significance of selecting appropriate classifiers and dimensionality reduction techniques. Overall SVM achieved the highest accuracy of 98.64% with FA.

Wisconsin Breast Cancer																
DR Algorithms	**Without DR**				**With DR (PCA)**				**With DR (LDA)**				**With DR (FA)**			
Classifiers	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1(%)
MLP	97.28	97.15	97.15	97.15	97.28	97.00	97.00	97.00	97.96	97.64	98.00	97.81	97.28	96.83	97.73	97.20
SVM	95.92	95.47	96.02	95.73	97.96	97.99	97.61	97.80	96.60	96.63	96.47	96.55	98.64	98.92	98.21	98.55
RF	97.28	97.07	98.07	97.07	97.28	96.49	97.87	97.09	97.28	96.89	97.46	97.15	96.60	96.43	96.62	96.52
KNN	97.28	97.17	97.17	97.17	97.28	97.46	96.73	97.07	96.60	96.59	96.32	96.45	91.84	91.33	92.33	91.67
DT	94.56	93.92	94.91	94.34	97.28	97.20	97.20	97.20	94.56	95.41	93.09	94.05	95.24	94.98	95.20	95.08

Figure 8(a). The results of WBC using outlier detection with contamination of 0.2

Wisconsin Diagnosis Breast Cancer																
DR Algorithms	**Without DR**				**With DR (PCA)**				**With DR (LDA)**				**With DR (FA)**			
Classifiers	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)
MLP	98.26	98.26	98.26	98.26	94.78	94.77	94.77	94.77	96.70	96.11	96.74	96.41	92.31	92.52	90.49	91.37
SVM	97.39	97.43	97.35	97.38	93.04	93.03	93.11	93.04	97.39	97.43	97.35	97.38	89.57	89.64	89.47	89.53
RF	93.91	93.89	93.94	93.91	93.91	93.89	93.94	93.91	95.65	95.63	95.68	95.65	91.30	91.29	91.36	91.30
KNN	94.78	94.77	94.77	94.77	94.78	94.77	94.77	94.77	97.39	97.43	97.35	97.38	88.70	88.84	88.56	88.64
DT	92.17	92.20	92.27	92.17	92.17	92.15	92.20	92.16	95.65	95.63	95.68	95.65	90.43	90.41	90.45	90.42

Figure 8(b). The results of WDBC using outlier detection with contamination of 0.2

Information about various classifiers' effectiveness is derived from their performance study on the WDBC dataset, both with and without DR in Figure 8(b). MLP classifiers consistently performed the best without dimensionality reduction, achieving 98.26% across all parameters. When DR techniques such as PCA, LDA and FA applied, a noticeable decline in MLP's performance occurred, with accuracy decreasing to 94.78%, 96.70% and 92.31% respectively. Similarly, the SVM classifier, which performed strongly without dimensionality reduction with 97.39% accuracy, saw reduced performance with PCA and FA to 93.04% and 89.57% respectively. SVM with LDA gave accuracy of 97.39% similar to KNN but still better than other models. When employed without DR, PCA, LDA, or FA, the RF classifier maintained 93.91%, 93.91%, 95.65%, and 91.30% performance, respectively, thus never outperform as compared to other models. KNN classifier performed best when used LDA with an accuracy of 97.39%. Meanwhile, the DT model performed best with LDA achieving an accuracy of 95.65%. Overall, MLP achieved the highest accuracy of 98.26% without dimensionality reduction and SVM-LDA and KNN-LDA achieved accuracy of 97.39% as compared to other classifiers.

Wisconsin Prognosis Breast Cancer																
DR Algorithms	**Without DR**				**With DR (PCA)**				**With DR (LDA)**				**With DR (FA)**			
Classifiers	Ac(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)	Ac.(%)	Pr.(%)	Re.(%)	F1.(%)
MLP	81.63	83.14	79.76	80.46	44.90	43.36	43.45	43.39	89.80	90.77	88.69	89.35	44.90	42.47	42.86	42.50
SVM	87.76	91.18	85.71	86.83	59.18	57.48	55.95	54.96	89.80	90.77	88.69	89.35	57.14	28.57	50.00	36.36
RF	85.71	85.60	85.12	85.32	53.06	51.25	51.19	51.02	83.67	83.77	82.74	83.10	57.14	57.00	57.14	56.86
KNN	71.43	78.19	67.26	66.37	42.86	40.79	41.07	40.86	87.76	89.25	86.31	87.11	42.86	41.67	41.67	41.67
DT	69.39	69.25	69.64	69.18	55.10	54.17	54.17	54.17	83.67	83.77	82.74	83.10	57.14	57.00	57.14	56.86

Figure 8(c). The results of WPBC using outlier detection with contamination of 0.2

The performance evaluation of various classifiers on the WPBC dataset, with and without various DR, reveals significant differences in its effectiveness. In Figure 8(c), SVM classifier achieved the highest accuracy of 87.76% without DR. When DR techniques such as PCA are applied, performance of all the classifiers showed a huge decline with SVM achieving highest accuracy of 59.18 %. On the other hand, LDA provided a performance boost for most models with SVM and MLP achieving a similar accuracy of 89.80 % as compared to others. Meanwhile, FA DR decreased accuracy of all classifiers like PCA did, and SVM, RF and DT models achieved a similar accuracy of 57.14% highest as compared to other models. The SVM and MLP classifiers exhibited highest accuracy with LDA technique as compared to other classifiers. However, the performance of all models dropped dramatically while comparing their accuracy with WBC and WDBC. This drop of accuracy is due to the less number of instances in the dataset. These findings highlight the challenges of using dimensionality reduction techniques in prognosis-related datasets.

In general, all of the models applied on these datasets achieved an accuracy of ranging between 98.26% - 69.39% without DR techniques, but it is still preferable to use DR techniques because they significantly improve model performance by reducing the computational complexity and it allows the data to take up

the less space. Furthermore, it also reduces dataset overfitting, by identifying hidden patterns in the data, while working with high-dimensional datasets with more features than data points.

4.1. Performance Comparison on WBC Dataset

Table 3(a) shows the results of the proposed SVM-FA model in comparison with previous studies on WBC dataset. The proposed classifier attained the highest accuracy in the WBC dataset, indicating its effectiveness in correctly classifying breast cancer instances. FA reduces the original features into a lower-dimensional space, preventing noise and redundancy in the data and enable the model to locate the underlying patterns. Compared to SVM, the SVM-FA has the advantage of reducing the dimensionality of the dataset and enhancing the model's generalization capability. KNN is a simple and intuitive algorithm that can work well on small datasets, but it can suffer from the curse of dimensionality and become computationally expensive on large datasets. ANN can handle complex relationships between features, but it requires more training data and can be prone to overfitting. NB is a probabilistic algorithm that can work well with high-dimensional data and can handle dummy/missing values, but it makes strong assumptions about the distribution of features that may not hold in some datasets. Thus, SVM-FA model stands out for its accuracy.

Table 3(a). Performance Comparison on WBC Dataset

Reference	Classifier	Accuracy	Year
[38]	SVM	97.07 %	2020
[39]	SVM (Sequential Minimal Optimization)	96.90 %	2019
[40]	KNN	97.00 %	2019
	ANN	97.00 %	
	PCA+ANN	97.00 %	
[41]	NB	97.36 %	2018
[42]	DT	94.00 %	2018
	ANN	95.40 %	
This Study	**SVM-FA**	**98.64 %**	**2024**

4.2. Performance Comparison on WDBC Dataset

Table 3(b) compares the accuracy of several ML models that have been used over time by different researchers on the WDBC dataset. The MLP model achieved an accuracy of 98.26% at a contamination value of 0.2 among all the models without any DR as shown in Table 4(b). Compared to SVM and RF, the MLP has the advantage of enhancing the model's generalization capability. SVM and RF are powerful algorithms that can handle both linear and nonlinear datasets and have been widely used in classification tasks. However, they can be sensitive to the decision to use hyper-parameters. Overall, MLP performs better than the other models even without DR. This model can be a useful tool for diagnosing breast cancer and could potentially aid healthcare professionals in making more accurate and timely decisions.

Table 3(b). Performance Comparison on WDBC Dataset

Reference	Classifier	Accuracy	Year
[43]	KNN-Euclidean	95.68 %	2015
[44]	SVM	96.50 %	2021
	RF	96.50 %	
[45]	RF	96.10 %	2018
[46]	SVM	97.20 %	2019
	RF	97.20 %	
This Study	**MLP**	**98.26 %**	**2024**

4.3. Performance Comparison on WPBC Dataset

Table 3(c) compares the accuracy of various ML models applied to the WPBC dataset by researchers over time. The SVM and MLP models integrated with LDA achieved an impressive 89.80% accuracy, outperforming KNN, RF-RFE, LDA-SVM, and MLP. The GMDH neural network demonstrated even higher accuracy due to its ability to handle small and noisy datasets effectively [38]. Compared to other

models, SVM-LDA and MLP-LDA exhibited lower error rates while effectively reducing the dataset's dimensionality. This reduction helped minimize overfitting and enhanced the models' computational efficiency and predictive performance, making them stand out among the evaluated approaches.

Table 3(c). Performance Comparison on WPBC Dataset

Reference	Classifier	Accuracy	Year
[43]	KNN-Euclidean	72.00 %	2015
[47]	GMDH	96.90 %	2020
[48]	RF-RFE	74.13 %	2021
[49]	LDA-SVM	79.50 %	2022
[50]	MLP	78.60 %	2022
This Study	**SVM-LDA MLP-LDA**	**89.80 %**	**2024**

4.4. Discussion

SVM-FA performed well on the WBC dataset with an accuracy of 98.64%, as SVM is able to effectively capture the clear distinctions between benign and malignant cases. Meanwhile, MLP without any dimensionality reduction technique achieved the best accuracy of 98.26% on WDBC. These datasets have relatively straightforward, linearly or non-linearly separable patterns. However, the WPBC dataset is a more complex and noisy dataset, although SVM-LDA and MLP-LDA achieved the highest accuracy of 89.80% among other classifiers, it was still lower as compared to GMDH claimed accuracy of 96.60%. Our model could have performed better if we had selected a higher value, but it was nonetheless accurate at a contamination level of 0.2. SVM classifier is considered more robust to highly complex dataset and can handle high-dimensional, noisy data better, making it a superior choice for the WPBC dataset. Additionally, MLP depends on the quality of the input data. This difference in the nature of the datasets explains why MLP excels in WDBC, and it shows why SVM's strengths are more aligned with the data's challenges.

5. Conclusions

Machine learning has become increasingly prominent in breast cancer detection due to its robustness and effectiveness. In this study, different ML algorithms are applied to classify malignant and benign tumors in three breast cancer datasets. After preprocessing techniques including handling missing values and scaling of data, different dimensionality reduction techniques were applied on these algorithms. Performance variations across models on various breast cancer detection datasets are clearly evident based on their results. SVM with FA achieved the best accuracy of 98.64% on WBC dataset. On the other hand, MLP without any dimensionality reduction technique achieved the best accuracy of 98.26% on WDBC compared to the other classifiers. In the case of WPBC, SVM-LDA and MLP-LDA performed better than other ML algorithms and achieved a similar accuracy of 89.80%. While PCA causes models to underperform due to loss of critical information, SVM with LDA handles high-dimensional noisy data better, making it a superior choice for WPBC dataset.

Funding: This research was funded by Higher Education Commission (HEC) of Pakistan, 20-17332/NRPU/R&D/HEC/2021-2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. American Cancer Society. What is Cancer? https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html
2. National Breast Cancer Foundation. What is Breast Cancer? https://www.nationalbreastcancer.org/breast-cancer-awareness-month
3. National Cancer Institute. Breast Cancer. https://www.cancer.gov/types/breast.
4. Ara, Sharmin, Annesha Das, and Ashim Dey. "Malignant and benign breast cancer classification using machine learning algorithms." 2021 International Conference on Artificial Intelligence (ICAI). IEEE, 2021.
5. He, Z., Chen, Z., Tan, M., Elingarami, S., Liu, Y., Li, T., ... & Li, W. (2020). A review on methods for diagnosis of breast cancer cells and tissues. Cell proliferation, 53(7), e12822.
6. Sechopoulos, Ioannis, Jonas Teuwen, and Ritse Mann. "Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art." Seminars in Cancer Biology. Vol. 72. Academic Press, 2021.
7. Bharati, S., Podder, P., & Mondal, M. (2020). Artificial neural network based breast cancer screening: a comprehensive review. arXiv preprint arXiv:2006.01767.
8. Iranmakani, S., Mortezazadeh, T., Sajadian, F., Ghaziani, M. F., Ghafari, A., Khezerloo, D., & Musa, A. E. (2020). A review of various modalities in breast imaging: technical aspects and clinical outcomes. Egyptian Journal of Radiology and Nuclear Medicine, 51(1), 1-22.
9. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.
10. Marinovich, M. L., Wylie, E., Lotter, W., Lund, H., Waddell, A., Madeley, C., ... & Houssami, N. (2023). Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. Ebiomedicine, 90.
11. Z. Mushtaq, A. Yaqub, A. Hassan and S. F. Su, "Performance Analysis of Supervised Classifiers Using PCA Based Techniques on Breast Cancer," 2019 International Conference on Engineering and Emerging Technologies (ICEET), 2019, pp. 1-6, doi: 10.1109/CEET1.2019.8711868.
12. S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.
13. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S "Breast Cancer Prediction using Machine Learning" (2019).
14. Ramik Rawal "Breast Cancer Prediction Using Machine Learning" (2020).
15. Dana Bazazeh and Raed Shubair "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis" (2016).
16. Elsadig, Muawia A., Abdelrahman Altigani, and Huwaida T. Elshoush. "Breast cancer detection using machine learning approaches: a comparative study." International Journal of Electrical & Computer Engineering (2088-8708) 13.1 (2023).
17. Abunasser, Basem S., et al. "Literature review of breast cancer detection using machine learning algorithms." AIP Conference Proceedings. Vol. 2808. No. 1. AIP Publishing, 2023.
18. TIWARI, MONIKA and Bharuka, Rashi and Shah, Praditi and Lokare, Reena, Breast Cancer Prediction Using Deep Learning and Machine Learning Techniques (March 22, 2020).
19. N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018, pp. 98-103, doi: 10.1109/ICACCCN.2018.8748777.
20. M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226-229.
21. R. MurtiRawat, S. Panchal, V. K. Singh and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 534-540, doi: 10.1109/ICESC48915.2020.9155783.
22. Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" Procedia Computer Science, vol. 83, 2016. doi:10.1016/j.procs.2016.04.224.
23. Juhyeon Kim, Hyunjung Shin, Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data, Journal of the American Medical Informatics Association, Volume 20, Issue 4, July 2013, Pages613–

618, https://doi.org/10.1136/amiajnl-2012-001570.
24. S. Gupta, D. Kumar, and A. Sharma, "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis," J.Comput. Sci., vol. 2, no. 2, pp. 188–195, 2011.
25. Mushtaq, Zohaib & Qureshi, Muhammad Farrukh & Abbass, Muhammad & Al-Fakih, Sadeq. (2023). Effective kernel-principal component analysis based approach for wisconsin breast cancer diagnosis. Electronics Letters. 59. 10.1049/ell2.12706.
26. Arshad, Muhammad Waqas. (2023). PREDICTION AND DIAGNOSIS OF BREAST CANCER USING MACHINE LEARNING AND ENSEMBLE CLASSIFIERS. 4. 49-56. 10.17605/OSF.IO/9CFN6.
27. Singhal, Vattsal & Chaudhary, Yuvraj & Verma, Sanidhya & Agarwal, Umang & Sharma, Mr. (2022). Breast Cancer Prediction using KNN, SVM, Logistic Regression and Decision Tree. International Journal for Research in Applied Science and Engineering Technology. 10. 1877-1881. 10.22214/ijraset.2022.42688.
28. Hussein, Mahmoud, Mohammed Elnahas, and Arabi Keshk. "A framework for predicting breast cancer recurrence." Expert Systems with Applications 240 (2024): 122641.
29. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
30. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
31. Cheng, Z., Zou, C., & Dong, J. (2019, September). Outlier detection using isolation forest and local outlier factor. In Proceedings of the conference on research in adaptive and convergent systems (pp. 161-168).
32. Sorzano, Carlos Oscar Sánchez, Javier Vargas, and A. Pascual Montano. "A survey of dimensionality reduction techniques." arXiv preprint arXiv:1403.2877 (2014).
33. Thilagam, P. Santhi. "Multi-layer perceptron based fake news classification using knowledge base triples." Applied Intelligence 53.6 (2023): 6276-6287.
34. Reyaz, Nahida, et al. "SVMCTI: support vector machine based cricket talent ıdentification model." International Journal of Information Technology 16.3 (2024): 1931-1944.
35. Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.
36. Xing, Wenchao, and Yilin Bei. "Medical health big data classification based on KNN classification algorithm." Ieee Access 8 (2019): 28808-28819.
37. Chen, Caixia, Liwei Geng, and Sheng Zhou. "Design and implementation of bank CRM system based on decision tree algorithm." Neural Computing and Applications 33 (2021): 8237-8247.
38. Shamrat, FM Javed Mehedi, et al. "An analysis on breast disease prediction using machine learning approaches." International Journal of Scientific & Technology Research 9.02 (2020): 2450-2455.
39. Bayrak, Ebru Aydındag, Pınar Kırcı, and Tolga Ensari. "Comparison of machine learning methods for breast cancer diagnosis." 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT). IEEE, 2019.
40. Sahu, Bibhuprasad, Sachi Mohanty, and Saroj Rout. "A hybrid approach for breast cancer classification and diagnosis." EAI Endorsed Transactions on Scalable Information Systems 6.20 (2019).
41. Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." Journal of Algorithms & Computational Technology 12.2 (2018): 119-126.
42. Higa, Autsuo. "Diagnosis of breast cancer using decision tree and artificial neural network algorithms." cell 1.7 (2018): 23-27.
43. Rana, Mandeep, et al. "Breast cancer diagnosis and recurrence prediction using machine learning techniques." International journal of research in Engineering and Technology 4.4 (2015): 372-376.
44. Ara, Sharmin, Annesha Das, and Ashim Dey. "Malignant and benign breast cancer classification using machine learning algorithms." 2021 International Conference on Artificial Intelligence (ICAI). IEEE, 2021.
45. Li, Yixuan, and Zixuan Chen. "Performance evaluation of machine learning methods for breast cancer prediction." Appl Comput Math 7.4 (2018): 212-216.
46. Yadav, Akshya, et al. "Comparative study of machine learning algorithms for breast cancer prediction-a review." International Journal of Scientific Research in Computer Science, Engineering and Information Technology (2019): 979-985.
47. Khandezamin, Z., Naderan, M., & Rashti, M. J. (2020). Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. Journal of Biomedical Informatics, 111, 103591.

48. El_Rahman, Sahar A. "Predicting breast cancer survivability based on machine learning and features selection algorithms: a comparative study." Journal of Ambient Intelligence and Humanized Computing 12 (2021): 8585-8623.
49. Egwom, Onyinyechi Jessica, et al. "An LDA–SVM Machine Learning Model for Breast Cancer Classification." Bio Med Informatics 2.3 (2022): 345-358.
50. Zeid, M., D. El-Bahnasy, and S. E. Abu-Youssef. "An efficient optimized framework for analyzing the performance of breast cancer using machine learning algorithms." Journal of Theoretical and Applied Information Technology 100.14 (2022).