JAMI
Journal of Applied Mathematics and Informatics

J. Appl. Math. & Informatics Vol. **42***(2024), No. 6, pp. 1275 - 1291*
https://doi.org/10.14317/jami.2024.1275

ENHANCING MACHINE LEARNING MODEL EFFICIENCY THROUGH MODIFIED LDA FEATURE REDUCTION FOR DISEASE DIAGNOSIS

P. SARANYA, D. VIJI*, B. DHIYANESH, K. MURUGAN

Abstract. Feature selection is vital in building a machine learning model to recognize and choose the relevant data samples and eliminate the irrelevant or less significant features. In developing a machine learning model, feature selection and feature extraction significantly affect the model's outcome. It increases the model efficiency and reduces the computational speed cost. The capacity of machine learning to solve challenging problems in dynamic environments has led to its use in disease diagnosis with various improved methodologies. Although this has been attempted in conventional research, the feature reduction method could have been more effective, which has a detrimental impact on the system. This work suggests that Modified LDA (Modified Linear Discriminant Analysis) reduces the dimension of features to address such implications. It reduces the data dimensionality effectively while choosing features. It helps to improve the overall system metrics significantly. The effectiveness of this novel method in lowering the dimensions of the different cancer datasets like lung cancer, cervical cancer, and breast cancer is investigated. The proposed algorithm is analysed by comparing its performance with existing models based on the standard metrics, time, and number of features reduced.

AMS Mathematics Subject Classification : 11R29, 52A05, 15A15.
Key words and phrases : Feature selection, machine learning model, disease diagnosis, modified LDA (Modified Linear Discriminant Analysis), dimension reduction.

1. Introduction

In the past two decades, machine learning has grown from an academic interest to a corporate tool. For various industries, ML is the AI approach of

Received March 23, 2024. Revised June 24, 2024. Accepted July 1, 2024. *Corresponding author.

choice (Jordan et al., 2015). Many AI system experts now agree that training a system by exhibiting desirable input-output behaviour is far easier than manually programming. Machine learning has affected computer science and data-intensive industries like consumer services, complex system defect detection, and logistics management.

In most industries like biology, cosmology, social science, and many others, machine learning algorithms have been used to experiment with high-throughput samples. Recent examples of successful uses of machine learning include robotics, speech processing, and upcoming technical areas. Also, with the ImageNet database, the annotation of recognized objects Video, photo, text, audio, and social media data have been created at a rate that has never been seen before by new computer and internet programs. High-dimensional data make analysis and decision-making difficult. Feature selection (Yang et al., 2018) improves learning efficiency in processing high-dimensional data in theory and practice.

The feature selection process uses a selection criterion to choose the dataset's relevant characteristics. Removing redundant and superfluous information reduces the processing scale of data. Feature selection can pre-process learning algorithms, improving accuracy, speed, and simplicity. Feature selection and extraction reduce dimensionality. Unlike feature selection, feature extraction frequently transforms weak pattern recognition features into strong ones. For decades, image processing, text processing, and others have involved feature selection, a methodological and practical research area (Liu et al., 2012). The theoretical premise states that feature selection methods can be based on statistics, information theory, manifolds, and rough sets and classified according to various standard approaches.

In this paper is organized into the following sections. Section 2 elaborates on the proposed methodology and explains the algorithm. Sections 3 and 4 illustrate the experimental results and analysis of the proposed algorithm with three cancer datasets. Conclusion and future enhancements are given in last section.

1.1. Carcinoma Sub-types. The following five sub-types of carcinoma

$(i).Carcinoma : Impactsglandsandorganslikeskin, breast, lung, andpancreas.$

$(ii).Softorconnectivetissuesareaffectedbysarcoma.Example : bloodvessels, bones, muscles, andfat.$

$(iii).Melanoma : Cancerofthepigment - producingcellsoftheskin.$

$(iv).Lymphoma : Thisdiseaseaffectswhitebloodcells.$

$(v).Leukemiaaffectstheblood.$

A prevalent form of cancer that accounts for millions of deaths is called carcinoma, wherein lung cancer accounts for the most significant number of deaths. Cancer of the colon and rectum is the second leading cause of death, behind breast cancer. Each year, 400000 young people are diagnosed with cancer, according to WHO data.

The possibility of developing a chronic disease can be identified and treated significantly earlier. More than fifty percent of the populations in nations considered to be well-developed are affected by chronic diseases, and those populations spend most of their income on treatment for those ailments (Chen et al., 2017).

ML algorithms are utilized to perform healthcare data analytics and healthcare data management. Machine learning algorithms use patients' statistical information to determine how the disease works. They then use prediction techniques to apply their knowledge to future patient data. Machine learning algorithms and Data mining methods dramatically improve data analytics (Mir et al., 2018).

1.2. Literature survey. SVM with random forest algorithms is used (Sivaranjani et al., 2021) to figure out how likely someone will get a disease related to diabetes. An LDA-based framework was suggested to reduce the number of wrong classifications made with the sample ECG data (Masoomi et al., 2015). Linear discriminant analysis combined with a hybrid feature selection strategy and an ensemble method is the best way to find coronary heart disease (Kolukisa et al., 2018) in the medical literature. It has been shown that linear discriminant analysis works better than artificial neural networks at finding breast cancer disease (Pereira et al., 2020).

The problem of reducing the number of dimensions in the medical data used in the study (Hariharan et al., 2022) was solved using a modified form of linear discriminant analysis called Hybrid LDA. Kernel Principal Component Analysis (KPCA) reduces the study's dimensionality and features better than PCA (Dinesh et al., 2021). Breast cancer prediction using linear discriminant analysis for feature selection was carried out, and results with various classifiers were analysed in the literature (Gayathri et al., 2018).

According to Kernel PCA with SVM and recursive feature elimination are compared in this study (Octaria et al., 2020). The LDA-based classifier used in the research (Khalid et al., 2015) was utilized to finish the classification required for disease diagnosis. An extreme learning classifier was utilized for disease classification, and experiments were carried out to compare the suggested system with PCA and GDA, with the results indicating that the specified system performed better (Avc et al., 2019). An algorithm for the diagnosis of coronary heart disease was developed in the paper (Kolukisa et al., 2018) using a hybrid approach to feature selection. The research looked into the clinical and anthropometric usefulness of measurements in the context of breast cancer prediction (Bikesh et al., 2019).

A multi-stage learning approach combined statistical analysis with a wrapper method to determine the possibility of breast cancer recurrence (Alwohaibi et al., 2021). This was done in order to determine the possibility of breast cancer recurrence. In implementing a hybrid feature selection algorithm, a combination of optimization algorithms based on statistical analysis, correlational analysis,

and brainstorming are used to select features most effectively. Various classification algorithms, such as SVM, LR, and LDA, were utilized in the hybrid feature selection model experiments. It is stated in this paper that Linear Discriminant Analysis (LDA) performs better than the other two classifiers, and it also presents the idea of combining LDA with hybrid algorithms.

A solution known as the Doubly Regularized Linear Discriminant Analysis Classifier (Zaib et al., 2021) could address challenges of this nature. A novel method called PLDA is used to extract features from high-dimensional data spaces (Ahmad et al., 2020). A novel method for selecting features was founded on a variation of linear discriminant analysis with L2-norm (Yang et al., 2020). R2LDA, which stands for "double regularized linear discriminant analysis," was an improvement made to RLDA (Zaib et al., 2021). ANNs, FFNNs, and a Kernel Principal Component Analysis were used to create an intelligent tool to predict performance metrics using the health index.

According to the study's findings (Zheng et al., 2018), utilizing a linear discriminant analysis based on a harmonic mean makes it possible to integrate harmonic means between the calculations of classes. A speedy technique known as incremental linear discriminate analysis with QR factorization (Chu et al., 2015) was used to lessen the computational and space complexity involved in the feature extraction and classification.

This study (Korns et al., 2017) applies Linear Discriminant Analysis (LDA) in conjunction with genetic programming and symbolic classification to solve problems involving multiple classes of financial information. The accurate extraction of the underlying sub-manifolds of data in intra-class can be achieved through optimal graph embedding. In the meantime, a solution to the minimization problem in the form of an effective iterative optimization algorithm has been suggested. In order to evaluate how successful the specified method is, some encouraging experimental results are presented based on synthetic and real-world datasets. UCI Machine Learning provided the model's experiments with four small-scale datasets: Australian, Heart, Pima, and Diabetes. These datasets were obtained from UCI Machine Learning.

1.3. Objectives & Problem Identification. The main objectives are,

(i).To conduct pre-processing, which includes scaling the data and checking for missing values, to eliminate unwanted data and improve the performance rate.

(ii).To reduce the dimensionality of the data and convert them into low dimensional data using modified linear discriminant analysis.

(iii).To reduce the processing time for dimensionality reduction.

Following the evaluation and analysis of conventional methods for extracting and selecting features for classification, several issues were found to exist.

(i).LDA has some computational issues; to solve the issues, pixel grouping is used. It is a pre-processing step used to avoid the problems of large scatter

matrices when it comes to computing. This technique requires the use of heuristics in order to be successful. Also, LDA does not guarantee outlier's class in calculations scatter matrices.

(ii).When dealing with high-dimensional data, the particle swarm optimization technique can efficiently converge on solutions. As a result of the impossibility of striking a balance between the priorities of global search and those of local search, the problem can only be solved by finding the optimal local solution.

(iii).GBM's evaluation speed is affected by its high memory consumption. One must evaluate all ensemble-based learners to use the fitted GBM model for predictions. Despite the simplicity of each base-learner, a large ensemble can make quick predictions difficult.

2. Proposed methodology

The LDA method seeks to reduce the number of dimensions used by the original data matrix. Three actions had to be taken to accomplish this objective. The inter-class variance, known as the between-class matrix, is initially computed to determine the degree of separability between classes. Second, it needs to determine the within-class variance or within-class matrix. Third, it creates the lower-dimensional space with the calculated matrices.

Between-class variance calculation is the first step in the dimensionality reduction process using linear discriminant analysis. The difference between the mean of the ith class (μi) and the overall mean (mu) is what is meant to be represented by the between-class variance of the ith class, abbreviated as SBi. The LDA technique looks for a lower-dimensional space in order to maximize the between-class variance, which is another way of saying that it maximizes the distance that separates the different classes. Assume that,

Actual data is represented as $X = \{x1, x2, x3, x4 \ldots\ldots xn\}$ whereas xi denotes jth sample in the actual data.

x has M number of features, represented as xjϵRM

Between class variance be

$$\text{SBj} = (\mu\text{j} - \mu)(\mu\text{j} - \mu)\text{T} \tag{2.1}$$

and $\text{SB} = \sum_{j=1}^{c} n_j S_{Bj}$

$$\text{SB} = \sum_{j=1}^{c} n_j (\mu\text{j} - \mu)(\mu\text{j} - \mu)\text{T} \tag{2.2}$$

where SB is total between class variance and W is transformation matrix of LDA. Also mean values of classes' μ i and total mean μ are calculated as

$$\text{mj} = \frac{1}{n_i} \sum_{x_j \in \omega_i} x\text{j} \tag{2.3}$$

$$\mathrm{m} = \sum_{j=1}^{c} \frac{n_j}{N} mi \tag{2.4}$$

The second step in LDA is calculating within class variance SW, that is, the difference between the mean and samples of the classes. LDA looks to minimize the difference between the mean and class samples; that is, it has to minimize the within-class variance value.

Within-class variance is calculated as follows,

$$\mathrm{SWi} = \sum_{x_j \epsilon \omega_j, i=1\ldots c} \left(W^T x_j - m_i\right) 2$$

Can be expanded and written as follows

$$\mathrm{SWi} = \sum_{x_j \in \omega_i, i=1\ldots c} W^T S_{Wi} W$$

That can be

$$\mathrm{SWi} = \mathrm{diT} * \mathrm{di} = \sum_{j=1}^{n_i} \left(x_{ji} - m_i\right) \left(x_{ji} - m_i\right)^T$$

Where di is centering data of the respective sample i.

Step 1: Given a set of N samples {xi | yj, i = 1 ton, j = 1 ton where x and y are rows and columns in the data set }

Step 2: Mean value for each class as

$$mi = (1/ni) \sum xj \in \omega i xj \tag{2.9}$$

Step 3: Total mean for all classes

$$m = (1/N) \sum j = 1 \text{ toN } xj = \sum j = 1 \text{ to } c(nj/N)mj \tag{2.10}$$

Step 4. Within-class matrix as

$$\mathrm{Sw} = \sum_{i=1}^{c} a_k \left(b_k - \mathrm{c}\right) \left(b_k - \mathrm{c}\right)^T \tag{2.11}$$

Step 5: Between-class matrix as

$$\mathrm{Sb} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(m_{ij} - m_j\right) \left(m_{ij} - m_j\right)^T \tag{2.12}$$

where xi j represents the ith sample in the jth class.

Step 6. Compute $\sigma(\Delta \mathrm{ij})$ authority function for each class as follows:

$$\Delta_{ij} = \sqrt{\left(m_j - m_i\right)^T S_w^{-1} \left(m_j - m_i\right)} \tag{2.13}$$

Step 7: Modified within-class matrix as

$$\widehat{S_w} = \sum_{k=1}^{c} Q_k \left(\sigma\left(\Delta_{ik}\right)\right) + R_k \Bigg) S_w \tag{2.11}$$

$$\text{where Sw } = \sum_{i=1}^{c} w_k \left(x_k - \bar{x}\right) \left(x_k - \bar{x}\right)^T$$

Step 8: Modified between-class matrix as

$$\widehat{S}_b = \sum \mathrm{j} = 1 \text{ to } \mathrm{Pj}\,\sigma(\Delta \mathrm{ij})\mathrm{Sb} \tag{2.14}$$

where xi j represents the ith sample in the jth class and

$$\mathrm{Sb} = \sum_{\mathrm{j}=1}^{\mathrm{c}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\mathrm{x_{ij}} - \mathrm{m_j})\,(\mathrm{x_{ij}} - \mathrm{m_j})^{\mathrm{T}}$$

Step 9: From above equations, the matrix W calculated using fisher criteria

$$\mathrm{W} = \mathrm{S} - 1 \text{ WSB} \tag{2.15}$$

The Eigen values and vector as λ and V of W respectively are then calculated.

Step 10: Low dimensional space is obtained from sorted eigen vectors.

Step 11: Lower dimensional space of LDA is states as Y = XVk.

Similarly, for the classes, the values are calculated between class variance and within-class variances in previous steps; the third step is to generate the transformation matrix W of LDA. This is Fisher criteria.

Like the Fishers criteria is said to be $\arg\max_W \frac{W^T S_B W}{W^T S_W W}$

$\mathrm{S_W} = \lambda \mathrm{S_B W}$ where λ is the Eigen values of transformation matrix W.

Eigen vectors has to be calculated; Let Eigen vectors represented as V. Then,

$$\mathrm{V} = \{\mathrm{v_1}, \mathrm{v_2}, \mathrm{v_3} \ldots \mathrm{v_M}\}$$

which are non-zero vectors give information about the LDA space. Each Eigen vector has its own Eigen values and represents each one of the axes in LDA space. Eigen vector which has highest values in Eigen value matrix will construct the low dimensional space Vk. That states that actual data with X with N^*M number of features are reduced into k^*M numbers of features and other features are discarded.

Let Y = XVk denote the low dimensional data with k∗M numbers of features.

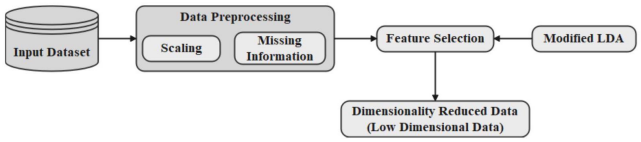

FIGURE 1. **Modified LDA feature Selection**

Class parameters like within-class and between-class do not determine classification accuracy, and they also insist that the manipulation of distances should not lead to overlapping classes. This is because the transformation will preserve the distances between already well-separated classes. When the so-called outlier class is prominent in estimating the scatter matrices, LDA does not guarantee it will discover the optimal subspace. LDA does not provide any assurance that it will locate the best subspace when an outlier is present.

To the improve the overall performance, Sb and Sw values are calculated based on the following formula:

$\widehat{S_w} = \sum_{j=i}^{c} p_j \sigma\left(\Delta_{ij}\right) S_{wj}$

Using Eqn.2.17 substituting S_{wj} in above formula, we get

$\widehat{S_w} = \sum_{j=1}^{c} p_j \sum_{i=1}^{N_j} \sigma\left(\Delta_{ij}\right) \left(x_i^j - \mu_j\right) \left(x_i^j - \mu_j\right)^T$

$\widehat{S_b} = \sum_{j=i}^{c} p_j \sigma\left(\Delta_{ij}\right) S_{bj}$

Eqn.2.19 substituting S_{bj} in above formula, we get

$\widehat{S_b} = \sum_{j=1}^{c} p_j \sum_{i=1}^{N_j} \sigma\left(\Delta_{ij}\right) \left(\mu_j - \mu_i\right) \left(\mu_j - \mu_i\right)^T$

Here, $\sigma\left(\Delta_{ij}\right)$ is said to be authority function.

$$\Delta_{ij} = \sqrt{\left(\mu_j - \mu_i\right)^T S_w^{-1} \left(\mu_j - \mu_i\right)} \tag{2.21}$$

This is the reasoning behind Δ_{ij}. It should come as no surprise that if $\sigma(\Delta)$ is a constant, S_b and S_w will be equal to the actual result obtained.

If $\sigma(\Delta)$ differs in each class, it will have an impact on the categorization of overlapping data from nearby classes.

In this case, the observation that, for the sake of categorization, if Δ is larger, $\sigma(\Delta)$ should be smaller. The value of $\sigma(\Delta)$ will be normalized so that the biggest value is 1 . Here, $\sigma(\Delta)$ is defined as:

$$\sigma\left(\Delta_{ij}\right) = \frac{1}{\Delta_{ij}^2} \tag{2.22}$$

3. Experimental Result

3.1. Dataset description.

The performance of the proposed methods is analysed with three data sets from the University of California, Irvine. Results are experimented for dimensionality reduction using modified linear discriminant analysis. Three cancer datasets for cancers like breast cancer, lung cancer, and cervical cancer from the UCI repository are trained and tested with the proposed model.

Table 3.1 Data set description for Dimensionality reduction using Modified LDA

Data set	No. of Instances	No. of Attributes	Class
Breast cancer	286	9	**2**
Cervical cancer	858	36	**2**
Lung cancer	32	56	**2**

3.2. Breast cancer dataset.

The Breast cancer dataset obtained from the UCI repository has 357 benign and 212 malignant cases. It is a multivariate dataset used for classification task analysis and has categorical type data attributes. Table 3.2 shows the attribute and its data type information of the breast cancer dataset.

Table 3.2 Sample breast cancer attributes

Column	
ID number	compactness_se
diagnosis	concavity_se
radius mean	concave points_se
texture mean	symmetry_se
perimeter mean	fractal_dimension_se
area mean	radius_worst
smoothness mean	texture_worst
compactness mean	perimeter_worst
concavity mean	area_worst
concave points mean	smoothness_worst
symmetry mean	compactness_worst
fractal dimension mean	concavity_worst
radius se	concave points_worst
texture se	symmetry_worst
perimeter se	fractal_dimension_worst
area se	Unnamed: 32
smoothness se	

3.3. Cervical cancer. The cervical cancer dataset was obtained from the UCI repository with 36 columns. It is taken from Universitario de Caracas in Caracas, Venezuela. The dataset has 858 patients' medical records. Many columns have null values, as some patient values are not stated. This data set is a multi-variant data set and is used for classification. Attributes are accurate and integer values. Table 3.3 shows the attributes and data type information of cervical cancer data set.

Table 3.3 Sample cervical cancer dataset attributes

Column	
Age	IUD
Number of sexual partners	IUD (years)
First sexual intercourse	STDs
Num of pregnancies	STDs (number)
Smokes	STDs: condylomatosis
Smokes (years)	STDs: cervicalcondylomatosis
Smokes (packs/year)	STDs: vaginalcondylomatosis
Hormonal Contraceptives	STDs: vulvo-perinealcondylomatosis
Hormonal Contraceptives (years)	STDs: syphilis

3.4. Lung cancer. Lung cancer dataset Kaggle datasets (download-d shuvojitdas/ lung-cancer-dataset) were obtained from the Kaggle repository. The lung cancer dataset has 16 columns. It is a multi-variant data set, and it is used

for classification tasks. Data attributes are types of integers. Table 3.4 shows the attributes and data type information of the lung cancer dataset.

Table 3.4 Sample lung cancer dataset attributes

Column	
Gender	Allergy
Age	Wheezing
Smoking	Alcohol Consuming
Yellow Fingers	Coughing
Anxiety	Shortness Of Breath
Peer Pressure	Swallowing Difficulty
Chronic Disease	Chest Pain
Fatigue	Lung_Cancer

3.5. Performance metrics. Two critical metrics have been considered for performance analysis: the number of features reduced and the time taken for the dimensionality reduction.

4. Performance analysis

4.1. Comparative analysis of Modified LDA with conventional methods. Table 4.1 shows the reduced number of features in the lung cancer dataset with PCA, LDA, and the proposed algorithm for dimensionality reduction and the time taken for their execution. Modified LDA has obtained 11 features for 16 features in the original data set, which is better than the other two methods implemented.

Table 4.1 Dimensionality reduction for lung cancer dataset using Modified LDA

Dimensionality Reduction Method	Number of Features Reduced	Time taken/ sec
PCA	14	20
LDA	12	16
Modified LDA	11	10

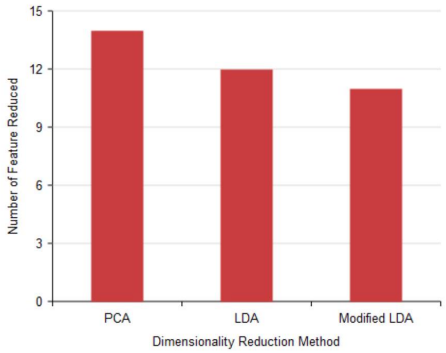

FIGURE 2. **Number of Reduced Features by Various Methods for Lung Cancer Dataset**

Figure 2 shows the number of features reduced by various methods for lung cancer, and the proposed algorithm has a minimum number of reduced features compared to other methods.

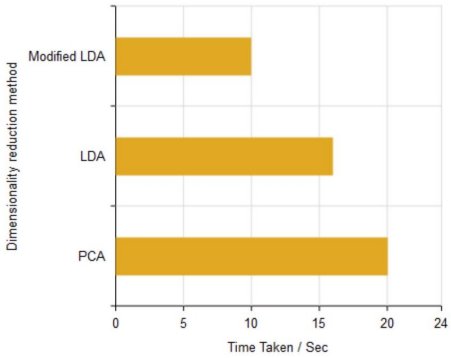

FIGURE 3. **Time Taken for Feature Reduction for Lung Cancer Dataset**

Figure 3 shows the time taken for feature reduction by PCA, LDA, and modified LDA; modified LDA takes less time than the other two methods.

4.2. Dimensionality reduction for cervical cancer dataset. Table 4.2 shows the reduced number of features in the cervical cancer dataset with PCA, LDA, and the proposed algorithm for dimensionality reduction and the time taken for their execution. Modified LDA has obtained 25 features for 30 features in the original data set, which is better than the other two methods implemented.

Table 4.2 Dimensionality Reduction for Cervical Cancer Dataset using Modified LDA

Dimensionality reduction method	Number of features reduced	Time taken/ sec
PCA	34	25
LDA	32	23
Modified LDA	29	18

Figure 4.2 shows the time taken for feature reduction by PCA, LDA, and modified LDA; modified LDA takes less time than the other two methods.

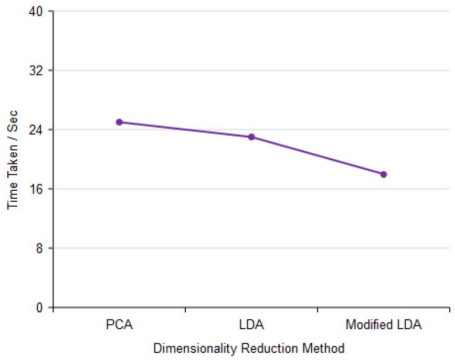

FIGURE 4. **Number of Reduced Features by Various Methods for Cervical Cancer Dataset**

Figure 4 shows the number of features reduced by various methods for cervical cancer, and the proposed modified LDA has a minimum number of reduced features compared to other methods.

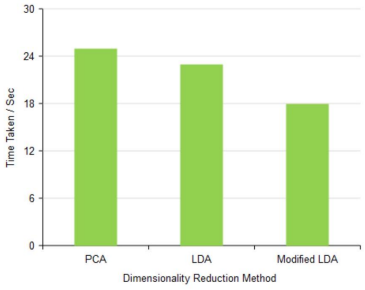

FIGURE 5. **Time Taken for Feature Reduction for Cervical Cancer Dataset**

Figure 5 shows the time taken for feature reduction by PCA, LDA, and modified LDA; modified LDA takes less time than the other two methods.

4.3. Dimensionality reduction for breast cancer dataset. Table 4.3 shows the reduced number of features in the breast cancer dataset with PCA, LDA, and the proposed algorithm for dimensionality reduction and the time taken for their execution. Modified LDA has obtained 28 features for 31 features in the original data set, which is better than the other two methods implemented.

Table 4.3 Dimensionality Reduction for Breast Cancer Dataset using Modified LDA

Dimensionality reduction method	Number of feature reduced	Time taken/ sec
PCA	29	24
LDA	28	21
Modified LDA	25	15

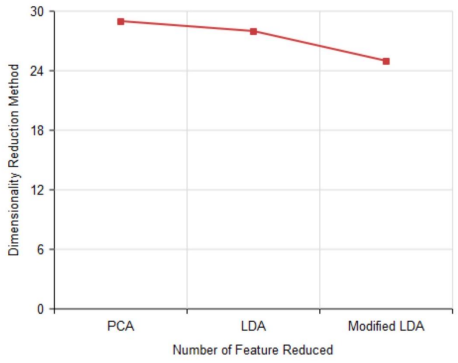

FIGURE 6. **Number of Reduced Features by Various Methods for Breast Cancer Dataset**

The Figure 6 shows the number of features reduced by various methods for breast cancer, and the proposed modified LDA has a minimum number of reduced features compared to other methods.

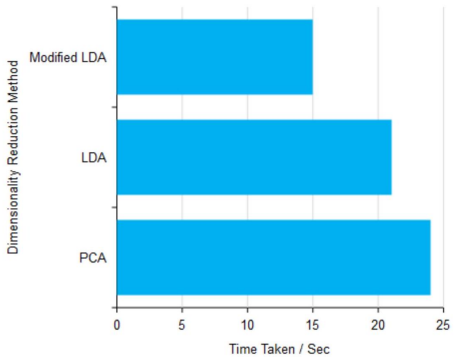

FIGURE 7. **Time taken for Feature Reduction for Breast Cancer Dataset**

Also, Figure 7 shows the time taken for feature reduction by PCA, LDA, and modified LDA; modified LDA takes less time than the other two methods.

5. Conclusion

The study's results show that a unique QPSO-modified LDA model is used to find and choose features for disease prediction models. In the process of disease prediction, the classification step uses gradient boosting in conjunction with weighted column sub-sampling. Predictions become more accurate when high-dimensional data is turned into low-dimensional data through dimensionality reduction using modified linear discriminant analysis. The QPSO-modified LDA is used for both the selection of features and the optimization of those features. The QPSO-modified LDA combination is utilized for the dimensionality reduction process, and QPSO is the optimization method used for feature selection optimization. GBA with weighted sub-sampling has been implemented to perform classification on the feature subset.

It is possible to perform classification with a weighted sub-sampling gradient-boosting model and base learners in the form of a decision tree. Concerns about computational memory could be resolved using the innovative approach. The results achieved by collecting data from three different datasets, such as those pertaining to breast, lung, and cervical cancer, demonstrate that the suggested framework has greater detection accuracy than previous models. Precision, the F1-score, and Recall are some of the performance measures used in formulating an evaluation of the performance capability of the suggested framework. Based on the findings of the comparative study, the work that has been proposed attains a better rate of Precision, F1-score, and Recall value than many other text categorization methods that are now in use.

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

Data availability : Not applicable

Acknowledgments : The healthcare facilities have access to a wealth of patient data that may be used for various objectives, such as disease prediction, improving patient care, assisting patients in receiving better care, etc. Worldwide, millions of people have diabetes, which is a condition that is considered to be one of the deadliest diseases (Mir et al., 2018). According to the WHO, diabetes is the leading cause of renal failure, blindness, heart attacks, and strokes. In recent years, millions of deaths have been attributed to diabetes and its consequences. The risks of mortality and other diseases brought on by this disease can be reduced with an early diagnosis.

The Cancer is a disease that leads to death, and it is one of the most prominent deadly diseases worldwide. In the year 2020, approximately 10 million people will have passed away as a result of this fatal illness. Breast cancer, lung cancer, colon and rectum cancer, and prostate cancer are the most prevalent forms. Using tobacco, having an imbalanced BMI, and eating an imbalanced diet are all significant contributors to infection. Early cancer detection allows for effective treatment in a hospital setting, dramatically lowering the death rate. The condition frequently manifests symptoms in people between 55 and 65 (Qiang et al., 2007). Diagnostic procedures for cancer include positron emission tomography scans, CT scans, MRI scans, ultrasound scans, and X-rays. With many different detection methods, disease early-stage prediction needs even more optimization and improved time complexity. The diagnosis of cancer at the initial stage is essential in order to lower the mortality rate. Deep learning algorithms for cancer prediction are currently being researched (Shahin et al., 2020) with a wide range of cancer types.}

References

1. M. Alwohaibi, M. Alzaqebah, N.M. Alotaibi, A.M. Alzahrani and Zouch, *Journal of King Saud University", Computer and Information Sciences*, Math. Oper. Research **34** (2022), 5192-5203.
2. E. Avc, H. Kutlu, R. Çoteli and M. Ustundag, *A New Expert Hepatitis Diagnosis System Based on Linear Discriminant Analysis-Extreme Learning Machine Classifier*, 1st International Informatics and Software Engineering Conference (UBMYK), **7** (2019), 1-5.
3. Bikesh Kumar Singh, *Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm*, Biocybernetics and Biomedical Engineering **39** (2019), 393-409.
4. Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram, *Designing disease prediction model using machine learning approach*, IEEE-3rd International Conference on Computing Methodologies and Communication (ICCMC), **7** (2019), 1211-1215.
5. De Giorgi, Maria Grazia, *Intelligent Combined Neural Network and Kernel Principal Component Analysis Tool for Engine Health Monitoring Purposes*, Aerospace 9.3 118 **7** (2022).
6. M.G. Dinesh, and D. Prabha, *Diabetes mellitus prediction system using hybrid KPCA-GA-SVM feature selection techniques*, Journal of Physics: Conference Series **1767** (2021), 1211-1222.

7. He, Liang, *Local pairwise linear discriminant analysis for speaker verification*, IEEE Signal Processing Letters **25** (2018), 1575-1579.
8. B. Hariharan, *Dimensionality Reduction based Medical Data Classification using Hybrid Linear Discriminant Analysis*, 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), **978** (2022), 6654-6671.
9. Michael I. Jordan, and Tom M. Mtchell, *Machine learning: Trends, perspectives, and prospects*, Science 349.6245 **2015** (2015), 255-260.
10. Md. Ekramul Hossain, Arif Khan, Mohammad Ali Moni, and Shahadat Uddin, *Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review*, IEEE/ACM Trans. Comput. Biol. Bioinformatics **18** (2021), 745–758.
11. Octaria, Elke Annisa, *Kernel PCA and SVM-RFE based feature selection for classification of dengue microarray dataset*, AIP Conference Proceedings **2264** (2020).
12. J. Ravi, *A robust measure of pairwise distance estimation approach: RD-RANSAC*, International Journal of Statistics and Applied Mathematics **2** (2017), 31-34.
13. Shahin and S. Almotairi, *An accurate and fast cardio-views classification system based on fused deep features and LSTM*, IEEE Access **8** (2020), 135184-135194.
14. S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, *Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction*, 7th International Conference on Advanced Computing and Communication Systems, **7** (2021), 141-146.
15. Yang and Libo, *Robust and efficient linear discriminant analysis with l 2, 1-norm for feature selection*, IEEE Access **8** (2020), 44100-44110.
16. F Yang, K.Z. Mao, G.K.K. Lee and W. Tang, *Emphasizing Minority Class in LDA for Feature Subset Selection on High-Dimensional Small-Sized Problems*, IEEE Transactions on Knowledge and Data Engineering **27** (2018), 88-101.
17. T. Zaib, S. Ballal, Khattak and T.Y. Al-Naffouri, *A Doubly Regularized Linear Discriminant Analysis Classifier with Automatic Parameter Selection*, IEEE Access **9** (2021), 51343-51354.
18. Zheng and Shuai, *A harmonic mean linear discriminant analysis for robust image classification*, IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), **7** (2016), 2375-0197.

Dr. P. Saranya received Doctor of Philosophy in Faculty of Computer Science and Engineering from Sathaybama Institute of Science and Technology, Chennai. She did Master of Engineering in the year 2013 from Roever Engineering College. She had 11 years of teaching experience in reputed Engineering colleges. She is interested in Database management systems, Data structures, Machine Learning and Big Data Analytics and big data.

Assistant Professor of Computational Intelligence, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.
e-mail: saranya.pattusamy@gmail.com

Dr. D. Viji received Ph.D. degree with the Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai India. She did Master of Engineering Degree in Computer Science and Engineering from Adhi parasakthi Engineering College, Melmaruvathur in 2015. She had 7 years of teaching experience in reputed Engineering colleges. She is interested in Big data analytics, Data mining.

Assistant Professor of Computing Technologies, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.
e-mail: dviji2k@gmail.com

Dr. Dhiyanesh Balasubramaniyan received Ph.D. on cloud computing at Anna University in 2017. He did his Master of Technology (Computer Science and Engineering) at

PRIST University. He has 15 years of rich experience in teaching at various reputed institutions. His research interests include cloud computing, network security, cyber security, and Machine Learning.

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamil Nadu, India.
e-mail: dhiyanu87@gmail.com

Dr. K. Murugan received Ph.D. on Information and Communication at Anna University in 2020. He did his Master of Engineering (Applied Electronics) at Mohamed Sathak Engineering College. He has 17 years 6 months of experience in teaching at various reputed institutions. His research interests include cloud computing, IOT Systems.

Associate Professor of ECE, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India.
e-mail: murugank@bitsathy.ac.in