IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 18, 2020, accepted March 10, 2020, date of publication March 16, 2020, date of current version March 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980942

Analysis of Dimensionality Reduction Techniques on Big Data

G. THIPPA REDDY[1], M. PRAVEEN KUMAR REDDY[1], KURUVA LAKSHMANNA[1], RAJESH KALURI[1], DHARMENDRA SINGH RAJPUT[1], GAUTAM SRIVASTAVA[2,3], (Senior Member, IEEE), AND THAR BAKER[4]

[1]School of Infromation Technology and Engineering, VIT, Vellore 632014, India
[2]Department of Mathematics and Computer Science, Brandon University, Brandon, R7A 6A9, Canada
[3]Research Center for Interneural Computing, China Medical University, Shenyang 10122, China
[4]Department of Computer Science, Liverpool John Moores University, Liverpool L3 3AF, U.K.

Corresponding authors: Gautam Srivastava (srivastavag@brandonu.ca) and Thar Baker (t.baker@ljmu.ac.uk)

ABSTRACT Due to digitization, a huge volume of data is being generated across several sectors such as healthcare, production, sales, IoT devices, Web, organizations. Machine learning algorithms are used to uncover patterns among the attributes of this data. Hence, they can be used to make predictions that can be used by medical practitioners and people at managerial level to make executive decisions. Not all the attributes in the datasets generated are important for training the machine learning algorithms. Some attributes might be irrelevant and some might not affect the outcome of the prediction. Ignoring or removing these irrelevant or less important attributes reduces the burden on machine learning algorithms. In this work two of the prominent dimensionality reduction techniques, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are investigated on four popular Machine Learning (ML) algorithms, Decision Tree Induction, Support Vector Machine (SVM), Naive Bayes Classifier and Random Forest Classifier using publicly available Cardiotocography (CTG) dataset from University of California and Irvine Machine Learning Repository. The experimentation results prove that PCA outperforms LDA in all the measures. Also, the performance of the classifiers, Decision Tree, Random Forest examined is not affected much by using PCA and LDA.To further analyze the performance of PCA and LDA the eperimentation is carried out on Diabetic Retinopathy (DR) and Intrusion Detection System (IDS) datasets. Experimentation results prove that ML algorithms with PCA produce better results when dimensionality of the datasets is high. When dimensionality of datasets is low it is observed that the ML algorithms without dimensionality reduction yields better results.

INDEX TERMS Cardiotocography dataset, dimensionality reduction, feature engineering, linear discriminant analysis, machine learning, principal component analysis.

I. INTRODUCTION

Machine Learning (ML) is one of the rapid growing technologies in the past 15 years. It has numerous applicaions in various fields like computer vision, bioinformatics, business analytics, healthcare, banking sector, fraud detection, prediction of trends etc. ML allows a computer to learn from a huge data samples and predicts the patterns that exisit within the data [1]. Machine learning algorithms are used in the different research areas to predict and classify the test data to produce the accurate results. There is a constant growth in the use of Machine Learning classifier models in medical field [2]. They have proved to be very helpful in diagnosis of various medical and clinical datasets [3].

During pregnancy periods, women are mostly affected with jaundice, high blood pressure and diabetes. These will affect growth of the baby in the womb. Cardiotocography (CTG) is used to diagnose the tests under the formation of fetal health and pregnancy periods [4]. CTG produces a recorded paper result of the mother uterine contractions signals and fetal heart rate [5]. CTG can also be used continuously if any irregularities occur during the fetal auscultation. This

The associate editor coordinating the review of this manuscript and approving it for publication was Michael Lyu.

practice has also become a standard procedure in many nations [6].

For any type of human disease prediction, datasets need to be pre-processed. In this connection, dimensionality reduction plays a vital role in reducing the high-dimensional data into reduced dimensionality [7]. In the past few decades, numerous dimensionality reduction techniques have been used for filtering the data samples of the considered dataset. Reduction of dimensionality requires mapping of inputs which are of high-dimensionality to a lesser-dimensionalility so that similar points in the input space are mapped to neighboring points on the manifold.

Feature engineering is an important pre-processing step that helps in extraction of transformed features from the raw data that will simplify the ML model and also improves the quality of the results of a machine learning algorithm. Machine Learning practitioners spend majority of their time on data cleaning and feature engineering [8]. Some of the examples of feature engineering are given below

1) Decomposition of categorical attributes.
2) Decomposition of Date-Time attribute into hour of the day, part of the day, day of the week etc.
3) Transforming attributes with numerical values to reduce the complexity of the input data.

In the current digital world huge amount of data is generated from almost all sectors. Machine learning algorithms are being used to extract meaningful patterns from this data which will aid in making executive and business decisions. Dimensionality reduction techniques can tremendously reduce the time complexity of training phase of ML algorithms hence reducing the burden of the machine learning algorithms. The main motivation behind this paper is to study the impact of dimensionality reduction techniques on the performance of the ML algorithms.

In this paper, two of the popular dimensionality reduction techniques namely Linear Discernment Analysis (LDA) and Principle Component Analysis (PCA) are investigated on widely used ML algorithms namely Decision Tree, Navie Bayes, Random Forest and Support Vector Machine using publicly available Cardiotocography (CTG) dataset from UCI machine learning repository [9]. Then experimentation is repeated on two other datasets namely, Diabetic Retinopathy Dataset from UCI machine learning repository [10] and Intrusion Detection Dataset from UCI machine learning repository [11].

The first step in the proposed work is to apply feature engineering to improve the quality of CTG dataset. In the next step the dimensionality reduction techniques, PCA and LDA are applied individually on the CTG dataset that will extract most important attributes. Next the extracted features are fed to train the aforementioned ML algorithms. In the next step, the performance of ML algorithms without application of the dimensionality reduction techniques is compared with the performance of ML algorithms with application of LDA and PCA separately using several performance metrics. Then the impact of feature engineering and dimensionality reduction techniques on the performance of ML algorithms is analysed in detail. The same steps are appliedon DR and IDS datasets.

The main contributions of the paper are as follows:

- Presenting a thorough and systematic review of the literature and background on dimensionality reduction techniques.
- Investigate the performance of PCA and LDA on ML algorithms against several metrics (Accuracy, F1-score, Precision, Recall, Sensitivity, Specificity).
- Prove that the dimesionality reduction techniques do not degrade the performance of ML algorithms to a great extent.
- Use feature engineering techniques (like transformation) during pre-processing phase which can reduce the burden on ML algorithms.
- Study the effect of dimensionality reduction on several datasets with varied dimensions.

The rest of the paper is organized as follows – Section II reviews the literature. Dimensionality reduction techniques used in this work and the methodology used in this work are detailed in Section III. In Section IV, experimentation results are analyzed, followed by the conclusion and future work in Section V.

II. LITERATURE REVIEW

In this section, a survey on several articles on ML algorithms and dimensionality reduction techniques are presented.

The authors in [12] embrace feature engineering principles to improve the quality of machine learning model to accelerate the discovery of potentially interesting catalytic materials.

UCI laboratory data is used in [13] to identify patterns using Artificial Neural Networks (ANN), Decision Trees (DT), SVM and Naive Bayes algorithms. The performance of these algorithms is compared with the proposed method in which an ANN with 18 neurons in the hidden layer is used. The proposed method gave better results compared to other methods.

The authors in [14] use a Convolutional Neural Networks (CNN) to predict heart disease based on ECG signals. This technique takes the heart cycles in the training phase with different starting points from the Electrocardiogram (ECG) signals. CNN can generate features with different positions in testing phase.

A neural networks based classification system of Fetal Heart Rate (FHR) signals to reduce the errors caused by human examination is presented by the authors in [15]. This algorithm is trained on FHR data and diagnoses FHR data in real time. The authors have designed a recurrent neural network called ''MKRNN'' and a convolution neural network classification method called ''MKNet''. Analysis of the proposed method on different classification models for fetal heart rate monitoring proved that neural network is innovative and feasible. Also MKNet is one of the best fetal cardiac monitoring model for classification.

In [16], the authors propose a new method for selecting a subset of optimal functions to increase the accuracy using pathological classification of the CTG. Feature selection performed with PCA as a pre-processing stage in machine training has proven to be very effective in improving computational time and accuracy. This method helps medical staff to make medical decisions more efficiently and quickly by interpreting CTG readings.

In [17], the authors present a computer-aided diagnostic support system that uses digital signal processing techniques for FTH automation, uterine tone signal detection and segmentation resulting in high sensitivity (SE) and positive predictivity (PPV). In this work, the main aim of the authors is to improve diagnostic accuracy.

An approach for digitalization of CTG signals using image processing techniques is proposed in [18]. This approach comprises four main phases: preprocessing, segmentation of images, signal removal and signal calibration. Limited adaptive histogram equalization contrasts and median filtering are used during the pre-processing phase to reduce noise and contrast enhancement. Otsu threshold algorithm is used to partition CTG images during image segmentation. The correlation coefficient is used to calculate whether original signals and CTG signals are similar. The experimental analysis shows that the digitization of the CTG signals is better.

A new methodology (CWV-BANNSVM) for diagnosing breast cancer (BC) is introduced by researchers in [19]. Wisconsin Breast Cancer Dataset (WBCD) was analyzed using Artificial Neural Networks and SVM. The proposed method integrated these two techniques. In the first technique SVM algorithm was applied to the BC dataset and the accuracy was 99.71%. In the second technique, the authors used the boosting ensemble technique and confidence-weighted voting method for BC detection. To evaluate the performance of the proposed method, a number of metrics such as Specificity, AUC, Gini, FPR, Tolerance, Precision, FNR, F1-score, and Accuracy were used. Accuracy of the proposed approach was 100%.

Integrated SVM and simulated annealing algorithm is introduced in [20] to diagnose hepatitis disease. The authors reviewed the test analysis using the real-time UCI dataset. The proposed hybrid model achieves a 97% accuracy rate. In this approach genetic algorithm and SVM are hybridized to optimize the parameters for dimensionality reduction. Experimental results indicate that the new algorithm significantly improves the accuracy.

In [21], the authors investigate the Naive Bayes classification for the diagnosis of coronary heart disease. The authors performed a test analysis using a longitudinal dataset in real time. The authors compared two methods for describing the characteristics of a single patient using horizontal aid and average temporal association rules period. The results obtained show that the proposed classifier performed better when compared to baseline classifiers.

The authors in [22] propose a private-key, fully homomorphic encryption algorithm to ensure the confidentiality of medical data for the private Naive Bayes classification. This approach allows data owner to classify information privately without accessing the trained model. The authors tested this algorithm for breast cancer data and achieved fast, accurate results.

A system for retrieving medical data is proposed by the researchers in [23],. This work tried to recover missing medical images with several images with linguistic information. Decision support tools and Decision Trees are used to collect information. Experimental results indicate that most recovered medical cases belong to the correct class.

In [24], the authors propose a Decision Tree based data mining technique to determine risk factors for CHD. The analysis was performed using the C4.5 Decision Tree algorithm, with different separation parameters based on risk factors. Experimental results show that CHD diagnosis can be reduced by this approach.

In [25], the authors integrate PCA and K-means algorithm to predict diabetes. In this work, the authors first applied PCA for dimensional reduction and then applied K-means to cluster the data. The hybridization of these two techniques provides a higher accuracy rate.

The authors in [26] investigated the performance of PCA clustering based on brain tumour images. In this model, the PCA method is first applied to MRI images of different sizes and applied clustering using K-means and FCM. The integration of PCA and K-means leads to a higher performance rate.

An integrated PCA-SVM algorithms is used in [27] to enhance digital image recognition. Experimental results show that hybridizing these two strategies provides better accuracy and recognition speed.

Several models are proposed to use meta heuristic algorithms such as firefly, BAT, cuckoo search along with popular classifiers like fuzzy rule based classifiers, ANN etc. to classify diabetic and heart disease datasets. Popular dimensionality reduction techniques like Locality Preserving Projections, rough sets etc. are used in these works for feature selection [28]–[32].

Sweta Bhattacharya *et al.* [33] present a PCA-Firefly algorithm to classify intrusion detection dataset. In this work, the transformation is performed by one-hot encoding, dimensionality reduction is performed by PCA-Firefly algorithm. The dimensionally reduced dataset is then classified by XGBoost classifier. The superiority of the proposed model is established by experimental results.Several researchers proposed machine learning models for effective classification of intrusion detection datasets [34]–[36].

The authors in [37] present a PCA-Firefly based Deep Learning Model for early detection of diabetic retinopathy. Optimal features are selected by PCA-Firefly algorithm and the Deep Neural Networks classify the diabetes retinopathy dataset. The proposed approach yielded good classification results when compared to other machine learning approaches considered.

TABLE 1. Summary of papers surveyed.

Paper cited	Methods used	Key findings	Limitations
[12]	Neural networks	Authors have followed an approach which was easily accessible with the geometric features like local electronegativity, effective coordination number of an adsorption sites.	Method is used to screen only 100-terminated multi-metallic alloys for CO2 electro reduction.
[13]	ANN model for multilayer perception	Multilayer perceptron was developed in the hidden layer using 18 neurons.	Specificity and accuracy is low for the considered dataset.
[14]	Back propagation Neural networks	Precision percentage is better on Hypertension diagnosis dataset.	Considered only a minimum number of samples.
[15]	Neural networks	Input layers for Long short term memory using the different parameters for Recurrent Neural Network have been addressed.	Noise reduction in the data is not so clear.
[16]	PCA and AdaBoost	With the help of AdaBoost model it has outperformed other models considered with a computation time of 2.4 and 11.6 seconds.	For the considered dataset, proposed system is not able to process the mixed abnormal values.
[17]	Hilbert Transform and Adaptive Threshold Technique	Achieved positive predictivity value (PPV) of 96.80%.	For original signals, transient changes and signal classification are very weak compared to the other parameters.
[18]	Effective method for digitizing the CTG signals	Proposed a good approach for digitizing the CTG signals.	Less number of samples considered.
[19]	Support Vector Machines and Artificial Neural Networks	Better results have been obtained for FPR, FNR, F1 score.	SVM and ANN were applied to the lesser attributes of dataset.
[20]	Support vector machine and simulated annealing (SVM-SA)	Attained classification accuracy of 96.25%.	Classification process was carried out with only 10-fold cross validation.
[21]	Genetic algorithm with Support vector machine	To construct chromosome, Kernel function and Kernel penalty factor parameters are used.	Dimensionality reduction is high, hence expected results are not achieved.
[22]	Naive Bayes classifier	Temporal association rules were used for coronary heart disease diagnosis and achieved high performance.	Temporal abstraction occurrences for multiple window system is slower than the single window system.
[23]	Convolution Neural Network (CNN) models	Improved the classification at different stages of DR.	Less accuracy in early stages.
[24]	Decision Tree	Classification assessment has been done for various Risk Factors of Coronary Heart Events.	Extracted models are not much used for reducing CHD morbidity.
[25]	PCA and K-means	Logistic regression model is well Improved and also prediction percentage is quite encouraging.	Measured only a minimum number of samples.
[26]	PCA and Fuzzy-means	Algorithms like EM-PPCA and PPCA worked effectively for the clustering algorithms such as K-Means and FCM.	Number of iterations considered are very less, so the sensitivity and accuracy values are not high.
[27]	PCA-SVM	Efficiency in average run time for the digital image recognition using coupling algorithm is achieved.	Only few samples are considered for the extraction process.
[28]–[32]	Firefly, BAT, Cuckoo Search, LPP, Artificial Neural Networks, Rough Sets	Used LPP, Rough Sets-Cuckoo Search for dimensionality reduction and meta-heuristic algorithms during classification.	Datasets considered have less instances and dimensions.
[33]	PCA, Firefly	Used PCA-Firefly for dimensionality reduction and XGBoost for classification.	Increase in time complexity as the dimensionality reduction and training phases of classification consume more time.
[37]	PCA, Firefly	PCA-Firefly is used for dimensionality reduction and Deep Neural Networks for classification.	Number of features in the dataset is very less. The model has to be tested on dataset with huge dimensions to test its scalability and robustness.

The papers studied in this section along with the key findings and their limitations is summarized in Table 1.

III. PRELIMINARIES AND METHODOLOGY

In this section, the two popular dimensionality reduction techniques, Principal Component Analysis and Linear Discriminant Analysis are discussed followed by discussion on the proposed methodology.

A. DIMENSIONALITY REDUCTION TECHNIQUES

1) PRINCIPAL COMPONENT ANALYSIS

PCA is a statistical procedure which uses an orthogonal transformation. PCA converts a group of correlated variables to a group of uncorrelated variables [38]. PCA is used for exploratory data analysis. Also PCA can be used for examination of the relationships among a group of variables. Hence it can be used for dimensionality reduction.

Assume that a dataset $x^{(1),}x^{(2),\ldots\ldots}x^{m}$ has n dimension inputs. n-dimension data has to be reduced to k-dimension ($k \ll n$) using PCA. PCA is described below:

1) Standardization of the raw data: The raw data should have unit variance and zero mean.

$$x_j^i = \frac{x_j^i - \overline{x}_j}{\sigma_j} \quad \forall j$$

2) Calculate the co-variance matrix of the raw data as follows:

$$\sum = \frac{1}{m}\sum_{i}^{m}(x_i)(x_i)^T, \sum \in R^{n*n}$$

3) Calculate the eigenvector and eigenvalue of the co-variance matrix as given in Equation 1.

$$u^T \sum = \lambda\mu$$
$$U = \begin{bmatrix} | & | & | \\ u_1 & u_2 \ldots & u_n \\ | & | & | \end{bmatrix}, \quad u_i \in R^n \quad (1)$$

4) Raw data has to be projected into a k-dimensional subspace: Top k eigenvector of co-variance matrix are chosen. These will be the new, original basis for the data. The Calculation of corresponding vector is given in Equation 2.

$$x_i^{new} = \begin{bmatrix} u_1^T x^i \\ u_2^T x^i \\ ::::: \\ ::::: \\ u_k^T x^i \end{bmatrix} \in R^k \quad (2)$$

In this way if the raw data is with n dimensionality, it will be reduced to a new k dimensional representation of the data.

2) LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA is another popular dimensionality reduction approach for pre-processing step in data mining and machine learning applications [39]. The main aim of LDA is to project a dataset with high number of features onto a less-dimensional space with good class-separability. This will reduce computational costs.

The approach followed by LDA is very much analogous to that of PCA. Apart from maximizing the variance of data (PCA), LDA also maximizes separation of multiple classes. The goal of Linear Discriminant Analysis is to project a dimension space onto a lesser subspace i (where $i \leq x - 1$) without disturbing the class information.

The 5 steps for performing a LDA are listed below.

1) For every class of dataset, a d-dimensional mean vectors is computed.
2) Computation of scatter matrices is carried out.
3) The eigenvectors $(E_1, E_2, E_3, \ldots .E_d)$ and their corresponding eigenvalues $(\psi_1, \psi_2, \psi_3, \ldots .\psi_d)$ of the scatter matrices are computed.
4) Sort the eigenvectors in descending order of eigenvalues and then opt for k eigenvectors which have maximum eigenvalues in order to form a $d * i$ matrix WW.
5) Use the above $d * i$ eigenvector matrix for transforming the input samples into a new subspace. i.e., $YY = XX * WW$.

PCA vs. LDA: Both LDA and PCA are linear transformation techniques which can be used to reduce the dimensionality/number of features. PCA is an ''unsupervised'' algorithm whereas LDA is ''supervised''.

B. METHODOLOGY

This work investigates the effect of feature engineering and dimensionality reduction techniques on the performance of ML algorithms on CTG dataset. The various steps used in this work are discussed below:

1) In step-1 feature engineering techniques, normalization and conversion of categorical data to numeric data is applied on CTG dataset. To normalize the input dataset, min-max standardscaler normalization method is used.
2) In step-2, the normalized dataset is experimented using ML algorithms, Decision Tree, Naive Bayes, Random Forest and SVM. The performance of these classifiers is then evaluated on the metrics, Precision, Recall, F1-Score, Accuracy, Sensitivity and Specificity.
3) In step-3 LDA is applied on normalized dataset to extract the most prominent features. The resultant dataset is then experimented on the ML algorithms. The ML algorithms using LDA are evaluated on the metrics mentioned in second step.
4) In the step-4, PCA is applied on the normalized dataset. The resultant dimensionally reduced dataset is then experimented using the aforementioned ML algorithms. The results obtained are again evaluated using the metrics mentioned in step 2.
5) In step-5 the results obtained by the ML algorithms without dimensionality reduction and also ML algorithms with LDA and PCA are analyzed. The effect of dimensionality reduction on the performance of ML algorithms is investigated.
6) Repeat steps 1 to 5 on Diabetic Retinopathy and Intrusion Detection Datasets to analyze the performance of PCA and LDA on differnt varities of datasets.

The proposed methodology is shown in Figure 1. The effect of dimensionality reduction techniques on ML algorithms is evaluated in the next section.

IV. RESULTS AND DISCUSSIONS

The experimentation is performed on CTG dataset which is collected from publicly available UCI machine learning repository using Python 3. A personal laptop with Windows 10 operating system and 8 GB RAM is used for this experimentation. The important attributes in the dataset are explained in Table 2.

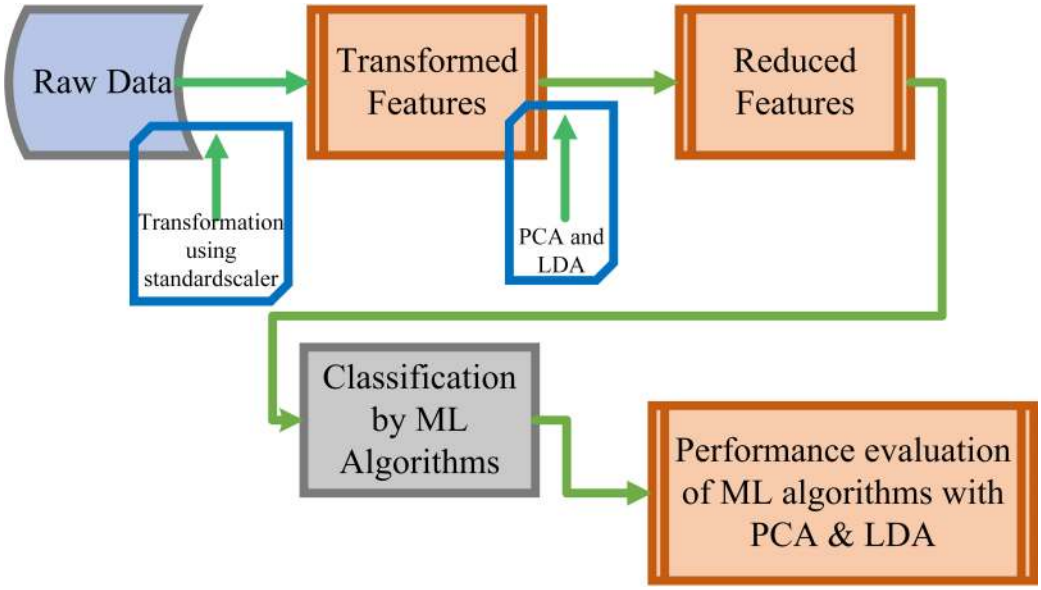

FIGURE 1. Proposed model based on PCA and LDA dimensionality reduction techniques.

TABLE 2. Major attributes in the dataset.

LB	FHR baseline (beats per minute)
AC	Accelerations per second
DL	Light decelerations per second
DS	Severe decelerations per second
DP	Prolongued decelerations per second
ASTV	percentage of time with abnormal short term variability
MSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
MLTV	mean value of long term variability

A. DATASET DESCRIPTION

Many women feel uneasy during the trimester of their pregnancy time. During this time, a fetus heart rate also has some problems with respect to the oxygen supply. Cardiotocogram tracing is used to crisscross the signs of an unborn baby's heart rate. Cardiotocography is used to observe the fetal heart and contractions of the uterus. Cardiotocography dataset is considered from UCI machine learning repository, which has 2126 instances and 23 attributes. The major attributes that are used for contractions of the uterus and the fetal heart are UC (uterine contractions per second) and FM (fetal movements per second). Few other attributes also play a role in recognizing the fetal heart.

B. METRICS FOR EVALUATION OF THE MODEL

1) ACCURACY

It is the percentage of correct predictions that a classifier has made when compared to the actual value of the label in the testing phase. Accuracy can be calculated as below:

Accuracy = (TN + TP)/(TN+TP+FN+FP) Where, TP is true positives, TN is true negatives, FP is false positives, FN is false negatives.

If the class label of a record in a dataset is positive, and the classifier predicts the class label for that record as positive, then it is called as true positive. If the class label of a record in a dataset is negative, and the classifier predicts the class label for that record as negative, then it is called as true negative. If the class label of a record in a dataset is positive, but the classifier predicts the class label for that record as negative, then it is called as false negative. If the class label of a record in a dataset is negative, but the classifier predicts the class label for that record as positive, then it is called as false positive.

2) SENSITIVITY

It is the percentage of true positives that are correctly identified by the classifier during testing. It is calculated as given below: TP/(TP + FN)

3) SPECIFICITY

It is the percentage of true negatives that are correctly identified by the classifier during testing. It is calculated as given below: TN/(TN + FP)

C. PERFORMANCE EVALUATION OF CLASSIFIERS WITH PCA AND LDA

The results of experimentation are discussed in this section. First the dataset, without dimensionality reduction is experimented using the following machine learning algorithms: Decision Tree, Naive Bayes, Random Forest, and SVM. Table 3, Table 4, Table 5, Table 6 show the confusion matrices for these algorithms. The confusion matrices show that SVM and Random Forest algorithms perform slightly better than Decision Tree and Naive Bayes in terms of Precision, Recall and F1-score.

Decision Tree Confusion Marix: $\begin{bmatrix} 323 & 3 & 0 \\ 4 & 54 & 0 \\ 0 & 0 & 42 \end{bmatrix}$

TABLE 3. Decision_Tree confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.99	0.99	0.99	326
2.0	0.95	0.93	0.94	58
3.0	1.00	1.00	1.00	42
micro average	0.98	0.98	0.98	426
macro average	0.98	0.97	0.98	426
weighted average	0.98	0.98	0.98	426

Naive Bayes Confusion Matrix: $\begin{bmatrix} 336 & 2 & 2 \\ 4 & 52 & 0 \\ 0 & 6 & 24 \end{bmatrix}$

TABLE 4. Naive bayes confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.99	0.99	0.99	340
2.0	0.87	0.93	0.90	56
3.0	0.92	0.80	0.86	30
micro average	0.97	0.97	0.97	426
macro average	0.93	0.91	0.91	426
weighted average	0.97	0.97	0.97	426

Random Forest Confusion Matrix: $\begin{bmatrix} 325 & 1 & 0 \\ 5 & 53 & 0 \\ 0 & 0 & 42 \end{bmatrix}$

TABLE 5. Random forest confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.98	1.00	0.99	326
2.0	0.98	0.91	0.95	58
3.0	1.00	1.00	1.00	42
micro average	0.99	0.99	0.99	426
macro average	0.99	0.97	0.98	426
weighted average	0.99	0.99	0.99	426

SVM Confusion Matrix: $\begin{bmatrix} 332 & 1 & 0 \\ 4 & 54 & 0 \\ 0 & 1 & 34 \end{bmatrix}$

TABLE 6. SVM confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.99	1.00	0.99	333
2.0	0.96	0.93	0.95	58
3.0	1.00	0.97	0.99	35
micro average	0.99	0.99	0.99	426
macro average	0.98	0.97	0.98	426
weighted average	0.99	0.99	0.99	426

Figure 2 shows the performance of aforementioned algorithms on the dataset based on accuracy, sensitivity and specificity measures. The Figure 2 shows that all the algorithms perform equally good in all these measures. Accuracy of Decision Tree, Naive bayes, Random Forest and SVM are 98.35%, 96.71%, 98.59%, and 98.59% respectively.

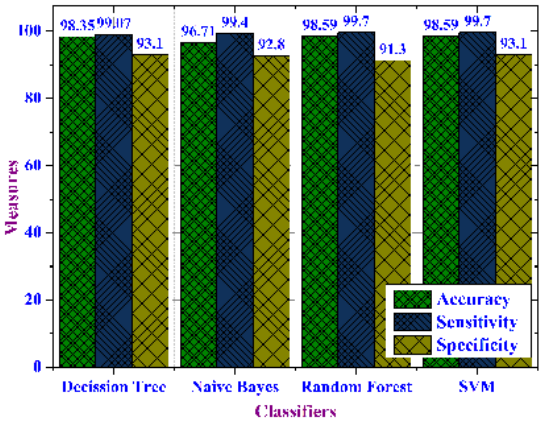

FIGURE 2. Performance evaluation of classifiers without dimensionality reduction.

Sensitivity achieved by these algorithms is 99.07%, 99.4%, 99.7% and 99.7% respectively. Specificity achieved is 93.1%, 92.8%, 91.3% and 93.1% respectively. The above results show that random forest and SVM outperform the other two algorithms in terms of accuracy and sensitivity, where as Decision Tree and SVM perform better than other two algorithms in terms of specificity.

In the next phase, the dimensions of the dataset is reduced using Linear Discriminant Analysis (LDA). The number of non target dimensions/attributes which were 36 is reduced to 1 using LDA. Then the dataset with reduced attributes is evaluated using Decision Tree, Naive Bayes, Random Forest and SVM classifiers.

The confusion matrices for these experimentations are shown in Table 7, Table 8, Table 9, Table 10. The confusion matrices show that Decision Tree, SVM and Random Forest algorithms perform better than Naive Bayes in terms of Precision, Recall and F1-score.

Decision Tree-LDA Confusion Matrix: $\begin{bmatrix} 324 & 2 & 0 \\ 8 & 49 & 1 \\ 0 & 0 & 42 \end{bmatrix}$

TABLE 7. Decision Tree-LDA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.98	0.99	0.98	326
2.0	0.96	0.84	0.90	58
3.0	0.98	1.00	0.99	42
micro average	0.97	0.97	0.97	426
macro average	0.97	0.95	0.96	426
weighted average	0.97	0.97	0.97	426

Naive Bayes-LDA Confusion Matrix: $\begin{bmatrix} 326 & 0 & 0 \\ 19 & 39 & 0 \\ 0 & 42 & 0 \end{bmatrix}$

Random Forest-LDA Confusion Matrix: $\begin{bmatrix} 324 & 2 & 0 \\ 8 & 49 & 1 \\ 0 & 0 & 42 \end{bmatrix}$

SVM-LDA Confusion Matrix: $\begin{bmatrix} 325 & 1 & 0 \\ 8 & 50 & 0 \\ 0 & 0 & 42 \end{bmatrix}$

TABLE 8. Naive Bayes-LDA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.94	1.00	0.97	326
2.0	0.48	0.67	0.56	58
3.0	0.00	0.00	0.00	42
micro average	0.86	0.86	0.86	426
macro average	0.48	0.56	0.51	426
weighted average	0.79	0.86	0.82	426

TABLE 9. Random Forest-LDA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.98	0.99	0.98	326
2.0	0.96	0.84	0.90	58
3.0	0.98	1.00	0.99	42
micro average	0.97	0.97	0.97	426
macro average	0.97	0.95	0.96	426
weighted average	0.97	0.97	0.97	426

TABLE 10. SVM-LDA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.98	1.00	0.99	326
2.0	0.98	0.86	0.92	58
3.0	1.00	1.00	1.00	42
micro average	0.98	0.98	0.98	426
macro average	0.99	0.95	0.97	426
weighted average	0.98	0.98	0.98	426

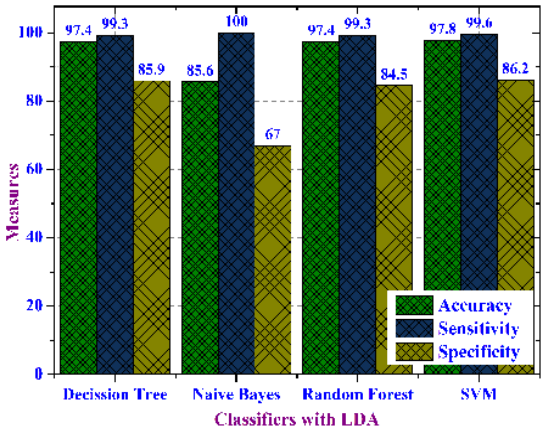

FIGURE 3. Performance evaluation of classifiers using LDA for dimensionality reduction.

Figure 3 shows the performance of these classifiers on the reduced dataset in terms of accuracy, sensitivity and specificity measures. Accuracy of Decision Tree, Naive Bayes, Random Forest and SVM are 97.4%, 85.6%, 97.4% and 97.8% respectively. Sensitivity achieved by these algorithms is 99.3%, 100%, 99.3% and 99.6% respectively. Specificity achieved is 85.9%, 67%, 84.5%, 86.2% respectively. These results show that Naive Bayes with LDA performs relatively less in terms of accuracy and specificity, whereas its sensitivity is 100%. The other 3 algorithms perform equally good in terms of all the three measures.

The dataset is then reduced using Principal Component Analysis (PCA) dimensionality reduction. The dataset which had 36 non-target dimensions is reduced to 26 non-target dimensions with PCA. The resultant dataset with reduced dimensions is then experimented using Decision Tree, Naive Bayes, Random Forest and SVM classifiers. The confusion matrices for these experimentations are shown in Table 11, Table 12, Table 13, Table 14. The confusion matrices show that Decision Tree, SVM and Random Forest algorithms perform better than Naive Bayes in terms of Precision, Recall and F1-score.

Decision Tree-PCA Confusion Matrix: $\begin{bmatrix} 324 & 2 & 0 \\ 5 & 53 & 0 \\ 0 & 1 & 41 \end{bmatrix}$

TABLE 11. Decision Tree-PCA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.98	0.99	0.99	326
2.0	0.95	0.91	0.93	58
3.0	1.00	0.98	0.99	42
micro average	0.98	0.98	0.98	426
macro average	0.98	0.96	0.97	426
weighted average	0.98	0.98	0.98	426

Naive Bayes-PCA Confusion Matrix: $\begin{bmatrix} 338 & 0 & 0 \\ 11 & 45 & 0 \\ 10 & 0 & 22 \end{bmatrix}$

TABLE 12. Naive Bayes-PCA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.94	1.00	0.97	338
2.0	1.00	0.80	0.89	56
3.0	1.00	0.69	0.81	32
micro average	0.95	0.95	0.95	426
macro average	0.98	0.83	0.89	426
weighted average	0.95	0.95	0.95	426

Random Forest-PCA Confusion Matrix: $\begin{bmatrix} 324 & 1 & 1 \\ 5 & 53 & 0 \\ 0 & 0 & 42 \end{bmatrix}$

TABLE 13. Random Forest-PCA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.98	0.99	0.99	326
2.0	0.98	0.91	0.95	58
3.0	0.98	1.00	0.99	42
micro average	0.98	0.98	0.98	426
macro average	0.98	0.97	0.97	426
weighted average	0.98	0.98	0.98	426

SVM-PCA Confusion Matrix: $\begin{bmatrix} 325 & 3 & 0 \\ 5 & 57 & 0 \\ 0 & 0 & 36 \end{bmatrix}$

Figure 4 shows the performance of these classifiers on the reduced dataset in terms of accuracy, sensitivity and specificity measures. Accuracy of Decision Tree, Naive Bayes, Random Forest and SVM are 98.1%, 95%, 98.3% and 98.1% respectively. Sensitivity achieved by these algorithms is 99.3%, 100%, 99.6% and 99% respectively. Specificity achieved is 91.3%, 80.3%, 91.3% and 92% respectively.

TABLE 14. SVM-PCA confusion marix.

	Precision	Recall	F1-score	Support
1.0	0.98	0.99	0.99	328
2.0	0.95	0.92	0.93	62
3.0	1.00	1.00	1.00	36
micro average	0.98	0.98	0.98	426
macro average	0.98	0.97	0.97	426
weighted average	0.98	0.98	0.98	426

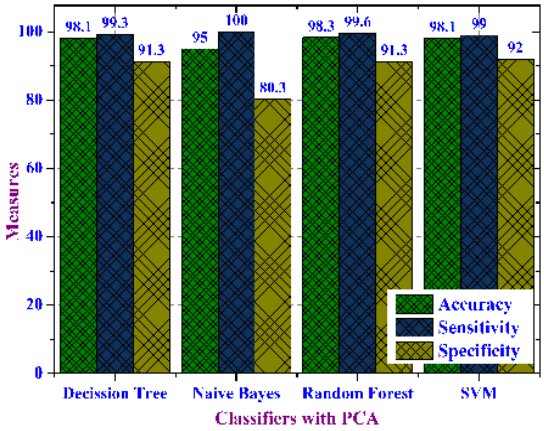

FIGURE 4. Performance evaluation of classifiers using PCA for dimensionality reduction.

As can be observed from Figure 4, Naive Bayes algorithm fares poorly with respect to other three algorithms against accuracy and specificity measures. Whereas its sensitivity is 100%. The other three algorithms perform equally good in terms of all the three measures.

The experimantation results on Cardiotocography datasets are summarized in Table 15.

To further analyze the performance of PCA and LDA on datasets with varying dimensions with several ML algorithms, the experimentation is repeated on two other datasets namely, Diabetic Retinopathy (DR) dataset and Intrusion Detection system (IDS) dataset. The diabetic retinopathy dataset has 1151 instances and 20 attributes. The IDS dataset has 125973 instances and 43 attributes. After applying one-hot encoding, to transform the data in categorical form into numericl data, these attributes were increased to 3024.

The experimentation results on Diabetic Retinopathy Dataset are summarized in Table 16.

The experimentation results of IDS Dataset are summarized in Table 17.

From the above investigations the following points can be observed as per Table 15, Table 16 and Table 17.

CTG Dataset (Table 15)

1) Decision tree, Naive Bayes, Random Forest and SVM perform almost the same when without dimensionality reduction techniques.
2) When the dimensionality of the dataset is reduced using PCA as well as LDA: Decision Tree, Random Forest and SVM classifiers fare better with respect to all the three measures. Performance of Naive Bayes is dropped in terms of accuracy and specificity when the dimensions are reduced, but its sensitivity is 100% even when the dimensions are reduced.
3) It can also be observed that dataset with reduced dimensions using PCA performs better when compared with that of LDA.
4) The performance of Decision Tree and Random Forest ML algorithms is almost similar even when the less significant features are eliminated by both PCA and LDA.
5) As the number of dimensions and the number of instances in this dataset is not that huge, dimensionality reduction is not having a positive impact on the results.

DR Dataset (Table 16)

1) SVM yielded best accuracy, specificity and sensitivity for DR dataset without dimensionality reduction.
2) When dimensionality reduction is applied on DR dataset, the perforance of the ML algorithms dropped

TABLE 15. Summary of results for CTG dataset.

	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Number of Features
DT	98	98	98	98.35	99.07	93.1	36
NB	97	97	97	96.71	99.4	92.8	36
RF	99	99	99	98.59	99.7	91.3	36
SVM	99	99	99	98.59	99.7	93.1	36
DT+LDA	97	97	97	97.4	99.3	85.9	1
NB+LDA	79	86	82	85.6	100	67	1
RF+LDA	97	97	97	97.4	99.3	84.5	1
SVM+LDA	98	98	98	97.8	99.6	86.2	1
DT+PCA	98	98	98	98.1	99.3	91.3	26
NB+PCA	95	95	95	95	100	80.3	26
RF+PCA	98	98	98	98.3	99.6	91.3	26
SVM+PCA	98	98	98	98.1	99	92	26

TABLE 16. Summary of results for DR dataset.

	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Number of Features
DT	57	57	57	57.1	54.5	59.2	19
DT + PCA	67	67	67	67.09	66.3	67.6	12
DT+LDA	67	67	67	66.6	63.3	69.2	1
NB	64	63	63	63.2	64.2	62.4	19
NB+PCA	61	60	60	59.7	59.7	59.7	12
NB+LDA	77	70	69	69.6	92	52.3	1
RF	70	69	69	68.8	76.2	63	19
RF+PCA	69	69	69	69.2	67.8	70.4	12
RF+LDA	68	68	68	67.2	64.3	70	1
SVM	79	76	76	76.1	89.4	65.3	19
SVM+PCA	71	66	66	65.8	81.3	55.7	12
SVM+LDA	73	70	70	70.12	81.1	61.5	1

TABLE 17. Summary of results for IDS dataset.

	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Number of Features
DT	99	99	99	99.1	98.4	94.2	3023
DT + PCA	99	99	99	99.3	99.5	95.3	2965
DT+LDA	87	88	88	87.6	92.1	77.2	1
NB	97	97	97	97.2	98.2	90.7	3023
NB+PCA	97	97	97	97.8	98.7	91.6	2965
NB+LDA	85	85	85	85.2	90.2	74.3	1
RF	99	99	99	99.4	99.4	98.2	3023
RF+PCA	99	99	99	99.7	99.8	98.4	2965
RF+LDA	88	88	88	88.2	92	85	1
SVM	99	99	99	99.6	99.7	98.9	3023
SVM+PCA	99	99	99	99.8	99.8	99	2965
SVM+LDA	89	89	89	89	92.6	85.8	1

TABLE 18. Summary of results for IDS dataset (1151 records and 20 attributes).

	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Number of Features
DT	54	54	54	54.2	53	55	20
DT+PCA	58	58	58	59.8	58.8	60.3	12
DT+LDA	56	56	56	56.4	57.9	56.3	1
NB	57	57	57	57.2	58.7	59.3	20
NB+PCA	56	56	56	56.2	57.8	58.1	12
NB+LDA	63	63	63	63.3	64.2	64.7	1
RF	58	58	58	58.2	59.1	60.1	20
RF+PCA	57	57	57	57.9	58.5	59.2	12
RF+LDA	55	55	55	55.1	56.8	57.2	1
SVM	61	61	61	61.8	62.1	62.9	20
SVM+PCA	57	57	57	57.8	58.2	59.6	12
SVM+LDA	58	58	58	58.8	59.6	60.1	1

significantly. Random Forest classifier yielded superior performance with PCA on this datasets, where as SVM permed best when LDA is applied on DR dataset.

IDS Dataset (Table 17)

1) SVM and Random Forest classifiers outperformed the other classifiers considered.
2) Classifiers with PCA yielded better results with respect to all the metrics than Classifiers without dimensioanlity reduction.
3) When LDA is applied, there is a dip in the performance of the classifiers.

From Table 16 and Table 17 it can be observed that the performance of the ML algorithms along with both PCA and LDA for IDS dataset outperform that of DR dataset as the number of instances and attributes are huge for IDS dataset. To observe the performance of these ML models on IDS dataset with similar dimensionality as that of DR dataset, a random 1151 records and 20 feataures are fed to these models. The experimental results of the reduced IDS dataset are depicted in Table 18. As some important features are removed in order to make the IDS dataset similar to that of DR dataset, the performance of the ML models on IDS dataset is reduced when compared to that of the DR dataset.

To summarize the above discussion, when the size of the dataset is too less, dimeensionality reduction techniques have negative impact on teh performance of the ML algorithms. When the size and dimensions of the dataset are significant PCA performs better than pure classifiers without dimensionality reduction and also classifiers with PCA. Hence, as the size of the dataset increases, it is suggested to use PCA for better results in terms of specificity, sensitivity and accuracy metrics. Also Random Forest and SVM algorithms with PCA yield best results. Hence it is recommended to use either Random forest or SVM classifiers with PCA when the dataset is of high-dimensions.

V. CONCLUSION AND FUTURE WORK

In this work, the effect of two pioneer dimensionality reduction techniques, namely Principal Component Analysis and Linear Discriminant Analysis on ML algorithms have been investigated. These dimensionality reduction techniques are applied on Cardiotocography dataset which is available in UCI machine learning repository. This dataset has 36 dependent attributes. By choosing to retain 95% of the components using PCA, number of dependent attributes has been reduced to 26, whereas LDA reduced the dependent attributes to 1. This reduced dataset is trained using four popular classifiers, Decision Tree classifier, Naive Bayes classifier, Random Forest classifier and SVM. From the results, it is observed that the performance of classifiers with PCA is better than that of with LDA. Also Decision Tree and Random Forest classifiers outperform the other two algorithms without using dimensionality reduction as well as with both PCA and LDA. When the same experimentation is performed on Diabetic Retinopathy dataset, it is observed that both PCA and LDA had negative performance on the results as the size of the dataset is less. Whereas for IDS dataset, the performance of the classifiers with PCA is better than that of classifiers without dimensionality reduction and classifiers with LDA.

In future, the effectiveness of these dimensionality reduction techniques can be tested on high dimensionality data such as images, text data etc. Also these techniques can be used on more complex algorithms like Deep Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, etc.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[2] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Berlin, Germany: Springer, 2003, pp. 63–71.
[3] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.* Berlin, Germany: Springer, 2000, pp. 1–15.
[4] R. M. Grivell, Z. Alfirevic, G. M. Gyte, and D. Devane, "Antenatal cardiotocography for fetal assessment," *Cochrane Database Systematic Rev.*, vol. 9, pp. 1–48, Sep. 2015.
[5] Z. Alfirevic, G. M. Gyte, A. Cuthbert, and D. Devane, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour," *Cochrane Database Systematic Rev.*, vol. 2, pp. 1–108, Feb. 2017.
[6] D. Ayres-de-Campos, C. Y. Spong, E. Chandraharan, and F. Intrapartum Fetal Monitoring Expert Consensus Panel, "FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography," *Int. J. Gynecol. Obstetrics*, vol. 131, no. 1, pp. 13–24, Oct. 2015.
[7] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative," *J. Mach. Learn. Res.*, vol. 10, nos. 66–71, p. 13, 2009.
[8] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Newton, MA, USA: O'Reilly Media, 2018.
[9] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-Sá, and L. Pereira-Leite, "Sisporto 2.0: A program for automated analysis of cardiotocograms," *J. Maternal-Fetal Med.*, vol. 9, no. 5, pp. 311–318, Sep. 2000.
[10] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowl.-Based Syst.*, vol. 60, pp. 20–27, Apr. 2014.
[11] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT—Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervas. Comput.*, vol. 17, no. 3, pp. 12–22, Jul./Sep. 2018.
[12] Z. Li, X. Ma, and H. Xin, "Feature engineering of machine-learning chemisorption models for catalyst design," *Catal. Today*, vol. 280, pp. 232–238, Feb. 2017.
[13] C.-A. Cheng and H.-W. Chiu, "An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 2566–2569.
[14] S. Zaman and R. Toufiq, "Codon based back propagation neural network approach to classify hypertension gene sequences," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2017, pp. 443–446.
[15] H. Tang, T. Wang, M. Li, and X. Yang, "The design and implementation of cardiotocography signals classification algorithm based on neural network," *Comput. Math. Methods Med.*, vol. 2018, Dec. 2018, Art. no. 8568617.
[16] Y. Zhang and Z. Zhao, "Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2017, pp. 1–6.
[17] J. A. Lobo Marques, P. C. Cortez, J. P. D. V. Madeiro, S. J. Fong, F. S. Schlindwein, and V. H. C. D. Albuquerque, "Automatic cardiotocography diagnostic system based on Hilbert transform and adaptive threshold technique," *IEEE Access*, vol. 7, pp. 73085–73094, 2019.

[18] Z. Cömert, A. engür, Y. Akbulut, Ü. Budak, A. F. Kocamaz, and S. Güngör, "A simple and effective approach for digitization of the CTG signals from CTG traces," *IRBM*, vol. 40, no. 5, pp. 286–296, Oct. 2019.

[19] M. Abdar and V. Makarenkov, "CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer," *Measurement*, vol. 146, pp. 557–570, Nov. 2019.

[20] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 570–579, Nov. 2012.

[21] Z. Tao, L. Huiling, W. Wenwen, and Y. Xia, "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling," *Appl. Soft Comput.*, vol. 75, pp. 323–332, Feb. 2019.

[22] K. Orphanou, A. Dagliati, L. Sacchi, A. Stassopoulou, E. Keravnou, and R. Bellazzi, "Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis," *J. Biomed. Informat.*, vol. 81, pp. 74–82, May 2018.

[23] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. Ahmed Khan, and W. Jadoon, "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530–150539, 2019.

[24] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 559–566, May 2010.

[25] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and *K*-means techniques," *Informat. Med. Unlocked*, vol. 17, 2019, Art. no. 100179.

[26] I. E. Kaya, A. Ç. Pehlivanlı, E. G. Sekizkarde , and T. Ibrikci, "PCA based clustering for brain tumor segmentation of T1w MRI images," *Comput. Methods Programs Biomed.*, vol. 140, pp. 19–28, Mar. 2017.

[27] L. Hu and J. Cui, "Digital image recognition based on Fractional-order-PCA-SVM coupling algorithm," *Measurement*, vol. 145, pp. 150–159, Oct. 2019.

[28] G. Thippa Reddy and N. Khare, "FFBAT-optimized rule based fuzzy logic classifier for diabetes," *Int. J. Eng. Res. Afr.*, vol. 24, pp. 137–152, Jun. 2016.

[29] N. Khare and G. T. Reddy, "Heart disease classification system using optimised fuzzy rule based algorithm," *Int. J. Biomed. Eng. Technol.*, vol. 27, no. 3, pp. 183–202, 2018.

[30] G. T. Reddy and N. Khare, "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model," *J. Circuits, Syst. Comput.*, vol. 26, no. 04, Apr. 2017, Art. no. 1750061.

[31] T. R. Gadekallu and N. Khare, "Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction," *Int. J. Fuzzy Syst. Appl.*, vol. 6, no. 2, pp. 25–42, Apr. 2017.

[32] G. T. Reddy and N. Khare, "Hybrid firefly-bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis," *Int. J. Intell. Eng. Syst.*, vol. 10, no. 4, pp. 18–27, 2017.

[33] S. Bhattacharya, S. R. K. S, P. K. R. Maddikunta, R. Kaluri, S. Singh, T. R. Gadekallu, M. Alazab, and U. Tariq, "A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU," *Electronics*, vol. 9, no. 2, p. 219, 2020.

[34] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019.

[35] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust intelligent malware detection using deep learning," *IEEE Access*, vol. 7, pp. 46717–46738, 2019.

[36] S. Kaur and M. Singh, "Hybrid intrusion detection and signature generation using deep recurrent neural networks," *Neural Comput. Appl.*, vol. 1, pp. 1–19, Apr. 2019.

[37] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I.-H. Ra, and M. Alazab, "Early detection of diabetic retinopathy using PCA-firefly based deep learning model," *Electronics*, vol. 9, no. 2, p. 274, 2020.

[38] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[39] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1569–1576.

G. THIPPA REDDY received the B.Tech. degree in computer science and engineering (CSE) from Nagarjuna University, the M.Tech. degree in CSE from Anna University, Chennai, India, and the Ph.D. degree from VIT, Vellore, India. He has 14 years of experience in teaching. He produced more than 25 international/national publications. He is currently working as an Assistant Professor (Senior) with the School of Information Technology and Engineering, VIT. He is currently working in the area of machine learning, the Internet of Things, deep neural networks, blockchain.

M. PRAVEEN KUMAR REDDY received the B.Tech. degree in computer science and engineering (CSE) from JNT University, the M.Tech. degree in CSE from VIT, Vellore, India and the Ph.D. degree from VIT, Vellore, India. He had worked as Software Developer with IBM, in 2011. He worked in Alcatel-Lucent, in 2013. He was a Visiting Professor with the Guangdong University of Technology, China, in 2019. He is currently working as an Assistant Professor with the School of Information Technology and Engineering, VIT. He produced more than 15 international/national publications. He is currently working in the area of energy aware applications for the Internet of Things (IoT) and high-performance computing.

KURUVA LAKSHMANNA received the B.Tech. degree in computer science and engineering from the Sri Venkateswara University College of Engineering, Tirupathi, India, in 2006, the M.Tech. degree in computer science and engineering(information security) from the National Institute of Technology at Calicut, India, 2009, and the Ph.D. degree from VIT, India, 2017. He was a Visiting Professor with the Guangdong University of Technology, China, in 2018. He is currently working as an Assistant Professor Senior with VIT, India. His research interests are data mining in DNA sequences, algorithms, and knowledge mining.

RAJESH KALURI received the B.Tech. degree in computer science and engineering (CSE) from JNTU, Hyderabad, the M.Tech. degree in CSE from ANU, Guntur, India, and the Ph.D. degree in computer vision from VIT, India. He is having 8.5 years of teach experience. He was a Visiting Professor with the Guangdong University of Technology, China, in 2015 and 2016. He is currently working as an Assistant Professor (Senior) with the School of Information Technology and Engineering, VIT, India. He has published research articles in various reputed international journals. His current researches are in the areas of computer vision and human computer interaction.

DHARMENDRA SINGH RAJPUT received the Ph.D. degree from NIT Bhopal, India, in 2013. He is currently working as an Associate Professor with VIT, India. His research areas are data mining, artificial intelligence, soft computing, automata, and natural language processing.

GAUTAM SRIVASTAVA (Senior Member, IEEE) received the B.Sc. degree from Briar Cliff University, USA, in 2004, and the M.Sc. and Ph.D. degrees from the University of Victoria, Victoria, BC, Canada, in 2006 and 2011, respectively. He then taught for three years at the Department of Computer Science, University of Victoria, where he was regarded as one of the top undergraduate professors in the Computer Science Course Instruction at the University. In 2014, he joined a tenure-track position at Brandon University, Brandon, MB, Canada, where he currently is active in various professional and scholarly activities. He was promoted to the rank an Associate Professor in January 2018. He is popularly known as an active in research in the field of data mining and big data. In his eight-year academic career, he has published a total of 60 articles in high-impact conferences in many countries and in high-status journals (SCI and SCIE) and has also delivered invited guest lectures on big data, cloud computing, Internet of Things, and cryptography at many Taiwanese and Czech universities. He received the Best Oral Presenter Award in FSDM 2017 which was held at the National Dong Hwa University (NDHU) in Shoufeng (Hualien County), Taiwan, in November 2017. He currently has active research projects with other academics in Taiwan, Singapore, Canada, Czech Republic, Poland, and USA. He is constantly looking for collaboration opportunities with foreign professors and students. He is an Editor of several international scientific research journals.

THAR BAKER received the Ph.D. degree in autonomic cloud applications from LJMU, in 2010. He was working as a Postdoctoral Research Associate in autonomic cloud computing with LJMU, where he built the first private cloud computing research platform for the Department of Computer Science. He worked as a Lecturer with the Department of Computer Science, Manchester Metropolitan University (MMU), in 2011. He became a Senior Fellow with the Higher Education Academy (SFHEA), in 2018. He is currently a Reader in cloud engineering and the Head of the Applied Computing Research Group (ACRG), Department of Computer Science, Liverpool John Moores University. He has published numerous refereed research articles in multidisciplinary research areas including: cloud computing, distributed software systems, big data, algorithm design, green and sustainable computing, and autonomic web science. He has successfully completed the Strategic Executive Development for Diverse Leasers in Higher Education (StellarHE) Course in 2016. He has been actively involved as a member of editorial board and review committee for a number of peer reviewed international journals, and is on programme committee for a number of international conferences. He was appointed as an Expert Evaluator in the European FP7 Connected Communities CONFINE Project from 2012 to 2015.