The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem

Mustafa Abdul Salam[1]
Artificial Intelligence Dept
Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

Ahmad Taher Azar[2]
Faculty of Computers and Artificial Intelligence, Benha, University, Egypt.and College of Computer and Information Sciences, Prince Sultan University, Riyadh, Kingdom of Saudi Arabia

Mustafa Samy Elgendy[3]
Scientific Computing Dept.
Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

Khaled Mohamed Fouad[4]
Information Systems Dept
Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

Abstract—**In most conditions, it is a problematic mission for a machine-learning model with a data record, which has various attributes, to be trained. There is always a proportional relationship between the increase of model features and the arrival to the overfitting of the susceptible model. That observation occurred since not all the characteristics are always important. For example, some features could only cause the data to be noisier. Dimensionality reduction techniques are used to overcome this matter. This paper presents a detailed comparative study of nine dimensionality reduction methods. These methods are missing-values ratio, low variance filter, high-correlation filter, random forest, principal component analysis, linear discriminant analysis, backward feature elimination, forward feature construction, and rough set theory. The effects of used methods on both training and testing performance were compared with two different datasets and applied to three different models. These models are, Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest classifier (RFC). The results proved that the RFC model was able to achieve the dimensionality reduction via limiting the overfitting crisis. The introduced RFC model showed a general progress in both accuracy and efficiency against compared approaches. The results revealed that dimensionality reduction could minimize the overfitting process while holding the performance so near to or better than the original one.**

Keywords—Dimensionality reduction; feature subset selection; rough set; overfitting; underfitting; machine learning

I. Introduction

Overfitting could be defined as the curse of a machine learning classifier and would probably be considered as the most common problem for beginners. It was a challenging problem with enthralling solutions that lied in dealing with the procedure's arrangements. Overfitting was an essential trouble which appeared illogically from outside; it occurred when the model proved its data accurately. [1][2].

The only service of the leaning-difference crisis was for observing when the model stepped into underfitting or overfitting. This Bias variance trouble is basic for a guarded machine learning. It's a method to identify the outcomes of the algorithm via dividing the evaluation error down. There are three kinds of error to be expected:

1) Bias error: The bias error was calculated by indicating the difference between the model's predicted evaluation and the real value that the model had been testing to reach.

2) Variance error: According to a certain data point of view, the variance error came from the turbulence of a sample predictions.

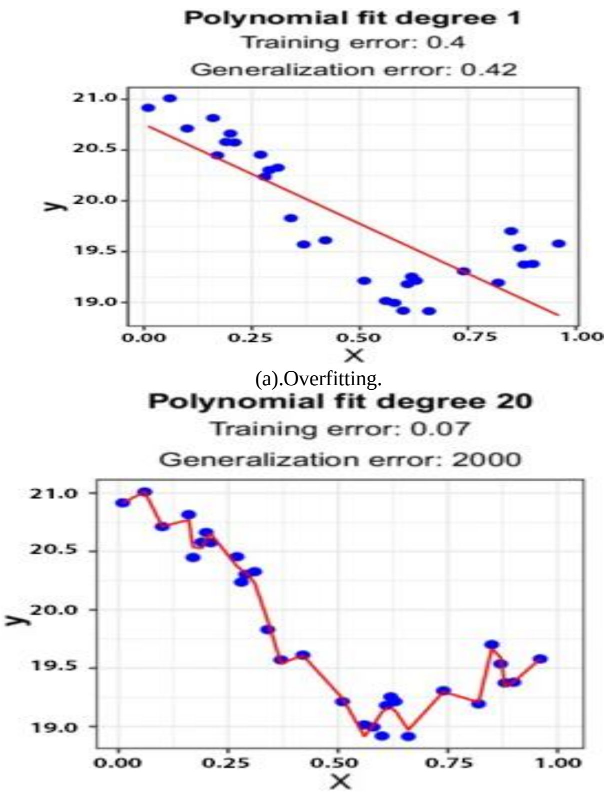
3) The irreducible error: It was likely to find out overfitting and underfitting Since training error and non-test error were gained by the low overfitting results. On the contrary, underfitting led to great training and a collection of test errors, as shown in figure 2 below.[2][3].

However, overfitting occurred in case that the model matched with the data very well, as illustrated in figure 1 (a). Underfitting took place whenever the model or algorithm was not applied to the data typically as cleared in figure 1 (b).

By investigation, many ways were reached to skip the overfitting

1) Regularization: In machine learning, regularizing the criterion was regarded as one technique in order to reduce, regularize, or narrow the coefficient value into zero. Such a method hindered exploring a more elaborated or even an all-purpose model. It suppressed the occurrence of overfitting, as revealed in Figure 3.

- Greenline reflected the coefficient before regularization
- Blue Line conveyed the coefficient after regularization



(a).Overfitting.

(b). Underfitting.

Fig. 1. Overfitting and Underfitting Problem Curves.

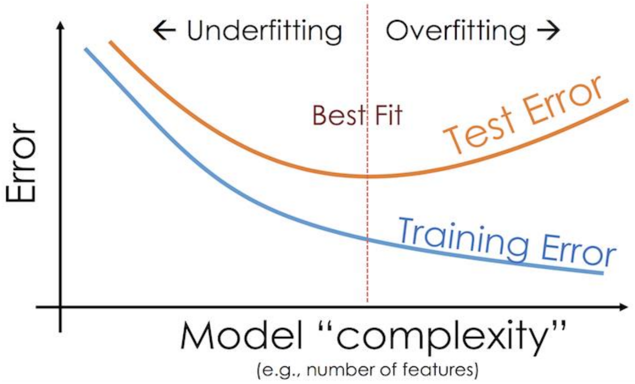


Fig. 2. The Relation between Train and Test Error and Model Order.

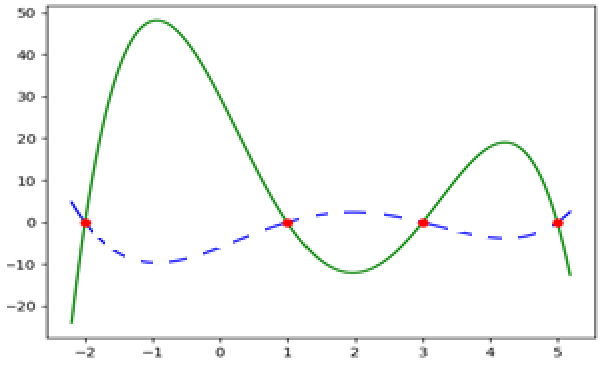


Fig. 3. Coefficients before and after Regularization.

2) Dimensionality Reduction.

3) It was hard to cope with more than a thousand features on a dataset, especially when it depended upon from where to begin! Dimensionality reduction was serving as an advantage and a defect at the same time to get a high number of variables. There was a plenty of data for the study, but the scale acted as an obstacle to get a precise information. The principle of Dimensionality reduction enabled us to to handle the extreme dimensional knowledge so that we can draw correlations and ideas from it easily. That reduction system also interfered to decrease the number of variables in the ordinary dataset by keeping a lot of data and via preserving (or improving) the model's efficiency. It was a successful attempt to operate over such huge datasets. Figure 4 illustrated that n data dimensions can be shortened into a subset of k dimensions (k<n).

This was called minimizing dimensionality. The advantages of dimensionality reduction on dataset were set as follows: [4]

1) The lower the number of measurements was, the less the space area of data storage was required.

2) Only Fewer measurements conform to less time for computation/training.

3) In the case with large dimensions, the algorithms did not accomplish well. Hence, dimensions reduction for a better performance for the algorithm was a must.

4) Multicollinearity was taken into consideration by excerpting superfluous features. For appetizers, two main characteristics were launched:' time cut in minutes on the treadmill' and 'the resulted burned calories. The cause was strongly related to its effect. As the time anyone killed increasingly over a treadmill, the more calories could be burned there. Accordingly, if just one of them was done there was no need to save them.

5) Data visualization Support. Focusing on details in larger dimensions was not easy at all. Therefore, minimizing the distance into 2D or 3D led to more accurate areas and noticed models.

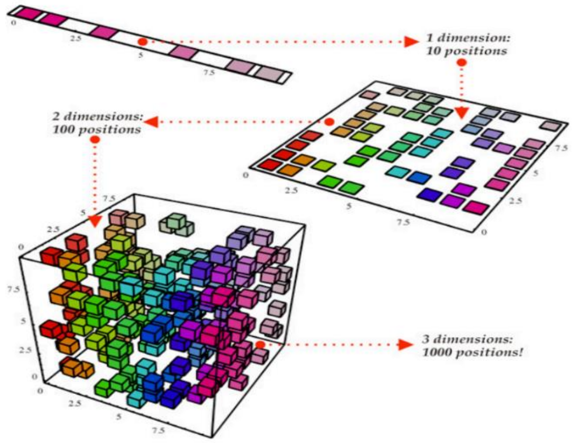


Fig. 4. Dimensionality Reduction.

The essential objective of this paper is to display how to avoid the overfitting problem for the model and improve its performance applying 9 Common dimensionality reduction techniques. The reduction technique was investigated in different dataset and their outcomes were compared through both training and testing. The evaluation was carried out depending on three separate models (ANN - SVM - RFC) with two datasets.

The remainder of discussion is arranged as follows: Section 2 contains the aimed work in brief. Section 3 tackles the description of dimensionality reduction techniques. Section 4 shows the datasets on which the experiments were taken place. It also summed up the preprocessing attitude and proved how we chose the dimensionality reduction algorithms' criterion. Section 5 traced the final results and concludes this paper.

II. Related Work

The dimension reduction helped in converting data which covered a large space into a tiny space of smaller dimension [5]. In different fields, the dimension reduction was a very effective step because it facilitated the classification, visualization or compression of huge data. The purpose was to eliminate the impact of issues caused by the high dimensionality data [6].

At last, numerous methods for dimension reduction have been approached. Such ways are worthy to handle any complicated non-linear problems. This reduction was presented as a substitution to the traditional linear techniques like the PCA which is the most common way of examining nonparametric data. Once a table of measurable data (unbroken or separate) was obtained, there were no observations (individuals).

Van Der Maatan et al. [7] High-dimensional data was everywhere in computing, and all of these datasets cover a lower dimensional area than that stretched by the whole dataset. A range of dimensional reduction methods had been improved to specify this lower-dimensional space. The data map minimized the number of indicators for the supervised learning problems; besides it developed the visual performance.

Chatfield et al. [8] Key Component Analysis (PCA) and Manifold Learning are two main techniques. Nonlinear subspace mapping was the pivot of Manifold Learning and not the linear subspace mapping. Dupont et al. in [9] describes the two manners had a prominent progress in a precise description of subspace which consumed most of the data variation. Moreover, the two techniques appealed to have a clear difference. Van Der Maatan et al. [7]; a current research showed the efficacy of PCA in real datasets. The process achieved a complete success after all.

Breiman et al. [10]; the large number of dimensional reduction methods involved a variety of global, linear, non-linear, and local ways. Every attitude gathered many data characteristics. These collective attitudes led to great success in supervised learning. The idea admitted a great progress from accidental forest to KNN backsliding that aggregated to super-learners. Sollich et al. [11]; ensembles applied variety to balance tendency, alteration, and estimator in order to carry out these effects. Therefore, those distinct techniques of dimensional reduction were probably to increase variation within a dimensional reduction ensemble. Van der Laan et al. [12] Throughout all achievements and features, it seemed that the progress of a group of local, international, linear and nonlinear dimensional reduction carriers would supply with better collective merging than any single construction. It was ordinary to differentiate the way that the superlearner ensembles. At least, it gave predictability which any model variable provided.

Dupont et al. [9] With regard to ensembles in dimensionality reduction, so tiny information had been discovered and the existing research papers did not produce ensembles with different base learner techniques. For instance, the Dupont and Ravet tried the variation of the t-distributed Stochastic Neighbor Embedding (t-SNE) criterion in their set. It was clear that it accomplished well and more effective than the tuned t-SNE model. By turning roles, it revealed the best performance among all different dimensional reduction techniques. Thus, no effort had been made to construct an ensemble using various techniques.

Zhao et al. [13] suggested a technique for dimensional reduction to the use of a spectral-space-based classification (SSFC) device to reduce the spectral dimensions. Typically, in a random way the most complicated information was taken using the Convolutional Neural Network (CNN) technique. At first, the obtained characteristics had been got and entered into the Linear Regression classifier in order to perform classification. SSFC was tested with two favorite HIS data sets, and the SVM classifier was applied to classify the images. Yan Xu et al. [14] recommended an accidental dismiss to a piece of picture from side to side with deep learning. There was a directed and unguarded structure of learning in the DNN. PCA was followed for the purpose of dimension reduction and classification. Multiple Instance Learning (MIL) had been used, too. In the beginning, natural and restricted structures should be learnt then features from the image had to be extracted from the existing dataset. The dataset consisted of high-resolution histopathological photographs of 132 patients.

The outcome conveyed that automated learning characteristics were the same like the old-fashioned set of features. Min Chen et al. [15] provided a model to help unsupervised images highlight learning for lung handling via unmarked knowledge by applying a coevolutionary, self-encoder, deep learning algorithm ;that needed a little bit of information to be called for active part learning. Autoencoder separate data information to rebuild as well as diverse input information and unique info information. Coiling autoencoder strengthened the neighborhood coiling relationship with the autoencoder to revise details for the convolution process. Dataset consisted of 4500 lung CT images from 2012 to 2015. Deep learning approaches were also investigated successfully by Yang et al. [16]. The central problems were solved in vacuum knobs examining by highlight extraction, knob detection, false-positive decline. Hence, the threatening order for the enormous volume of the father's chest filter was detected.

Deep learning also served to indicate an accurate diagnose for pneumonic knobs. The two-dimensional CNN, three-dimensional CNN, and Deep Faith Network were applied for clustering. Quantized autoencoder neural network was followed for features' derivation. An automatic technique of feature generation was presented by Rasool et al. [17] to strengthen any prediction and examine so various kinds of cancer. Unmonitored feature learning could be applied for the early detection of cancer. With the help of gene expression data, the sort of cancer could be tested. Concerning the feature learning process, softmax regression was followed as a learning procedure for the classifier. By joining 10-fold cross-validation with an aim to assess the classifier effectiveness. Hence, all the gained results were calculated showing the average accuracy of classification. A strategy for diabetic diagnosis was proposed by Y. Zheng et al. [18]. It supplied an artificial neural network. Experimental outcomes proved that this presented approach was a safe method with the situation of diabetes. It also participated in limiting the computing costs in addition to introducing accuracy. The main strategies of Highlight extraction had been planned and described to be significantly more suitable for the automatic prediction of ophthalmologist diseases than to emphasize detection methods following noisy details.

III. Preliminaries

Several dimensionality reduction techniques were implemented in the following studies:

A. Missing-Values Ratio (MVR)

In (MVR), data revision should precede model construction. There was an observation that some values were missed while examining the data. An attempt was done to discover the reason behind the problem. The solution held two options whether to assign them or to remove the missing values variables at all [4]. The coming steps had to be followed respectively to get such a technique:

- Indicating the type of the missing value.
- Features determined on the ratio of missing value had to be reduced as given:

$$Ratio\ of\ missing\ value = \frac{number\ of\ missing\ values}{total\ number\ of\ rows} \qquad (1)$$

- Rows which had a missing value such as "?, na, NULL, etc…" had to be deleted after dismissing the characteristics that had a high missing value ratio, it would be cancelled from the whole data set.

B. Low-Variance Filter (LVF)

In that method, a function was noticed to be with the same value in the dataset. With a strong observation, it wouldn't boost the formed model via this feature. Therefore, there would be zero variance in such a function.

The followed steps were stated as follows: [19]

- the variance of each feature had to be calculated.

$$\sigma^2 = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \mu) \qquad (2)$$

Where, $x_i$stood for individual values in a dataset. μ stood for the mean of those values. n was the number of values. The term $x_i - \mu$ were named as a deviation from the mean.

- The features having low variance were dropped and compared to the most minimum value.

C. High-Correlation Filter (HCF)

It was acting as an in between technique for the ex – two. It proposed that they had specified tendency. So, similar results were expected in return. As a deduction the performance of such models would be effectively declined (e.g., linear and logistic regression models). [20] Some definite steps had to be taken to apply the (HCF) technique:

- The relation between individual numerical features had to be evaluated.
- One of the functions was completely dismissed if the correlation coefficient achieved the least value.

D. Random Forest (RFC) (for Feature Importance)

It was regarded as one of the most famous machine learning algorithms. This approach made it simple to extract each variable's value on the tree decision leaving a kind of explanation. The algorithms were so genuine because they had high detective efficiency, low overfitting, and easy interpretability.[4]. In a word, whatever each vector was joined to the decision, it was computed at once easily. As a result, a narrower subset of features [21] would be chosen.

E. Principal Component Analysis (PCA)

PCA is a linear dimensional reduction technique depended upon projection techniques. Through its application, a higher-dimensional Euclidean space had been projected into a lower-dimensional Euclidean space [4]. Given a data matrix, X, a target space, Y, and a projection matrix, P, PCA illustrated the following mapping:

$$Y = PX \qquad (3)$$

The rows of P turned to a new basis for X, and by introducing this equation as an optimization problem (maximizing variance, reducing covariance between variables); it could be reconstructed in such a way that the singular value of decomposition was proposed to solve the equation. The covariance matrix, C, can be expressed as:

$$C = \frac{1}{n-1}\ PXX^TP^T \qquad (4)$$

Because of the shift of basis found in the orthogonal bases, and because of the high variance of elements, the truncation of result would convey a mapping to a lower-dimensional space using the new bases. And at the same time a lot of variance was kept from the original dataset.

F. Linear Discriminant Analysis (LDA)

A common monitored dimensionality reduction technique was the linear discriminant analysis (LDA) [20]. LDA reached the optimal linear transformation W, which reduced the distance within the class and lengthened the distance between classes simultaneously. The criterion J (XW) it maximized was:

$$J\ (XW)\ =\ -LDA(XW)\ =\ -\frac{W^{T}S_{B}W}{W^{T}\ S_{W}W} \tag{5}$$

where SB was the between class scatter matrix and SW was the within class scatter defined by:

$$S_B = \sum_c (\mu_c - \bar{x})(\mu_c - \bar{x})^T \tag{6}$$

$$S_W = \sum_c \sum_{x_i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \tag{7}$$

In which $\bar{x}$ was the mean of the data points X, μ_c and was the mean of the data points that belonged to class c.

G. Backward Feature Elimination (BFE)

To focus more and follow the 'Backward Attribute Removal' method, follow the coming steps: [22]

- The current features had to be obtained in the dataset then applied for the testing model.
- The degree of model performance had to be Calculated
- After computing the output of the model when deleting each function (n times), i.e., one variable was dropped every time and the model on the remaining n-1 variables would be tested.
- Determine the variable whose deletion got the smallest (or no) difference in the model's output, then delete that feature respectively.
- Repeat the ex-procedure many times till it was not easy for the variable to drop.

H. Forward Feature Construction (FFC)

This method was the opposite side we observed above with the Backward Attribute Removal. There was a challenge in having the right characteristics on behalf of deleting the features. Great attempts took place to reach the summit of model's performance [23]:

- The model n was tested many times starting with a single function via trying each function separately.
- The variable always provided with the best output in that indicated starting function.
- From time to time such a process was repeated through adding an element whereas the function that caused the largest progress of output was kept.
- This step was done again and again until no difference in the model efficiency was noticed.

I. Rough Set Theory (RS)

It was defined as a traditional theory created from a main research on the theoretical qualities of information systems [24]. With inaccurate and raucous data, a rough collection approach could be applied to explore any systemic relationships. The main target of this study was to activate the idea of approximation [24][25] statistical techniques were applied to display any hidden data models. Its function was to select features, derive data, and reduce details. The outlined features of reduction's fundamentals were as follows:

- The same equivalence class structure was supplied typically as that reflected by the full feature set which represented by [x]RED = [x]P.
- It is minimum
- It is not perfect

Algorithm 1: Reduct Calculation

```
Input: C, the set of all conditional features
       D, the set of all decisional features
Output: R, a feature subset

1. T := { }, R : = { }
2. repeat
3. T : = R
4. ∀ x ∈ (C – R )
5. if  γ RU{X} ( D ) > γT( D )
6.     T : = R U {x}
7.     R : = T
8. until  γR( D ) = γC( D )
9. return R
```

IV. Methodology

A. Dataset Description

The tested models were explained and confirmed with 2 classification data sets from the UCI machine-learning repository [26] Such types were involved in the experiments and comparative performance. The data sets were chosen according to various numbers of features and examples to introduce different kinds of problems on which the new approach could be examined. In addition to this, algorithm performance could be confirmed via a selection of a set of high-dimensional data. The information of training, calculating, and testing were similar in size. The training component had to be applied to train the used classifier; the validation component was used to compute the performance of the classifier; while, the evaluation component was used to evaluate the last selected characteristics which were revealed by the qualified classifier.

1) Congressional voting records dataset: This data set contained votes by Congressmen of the House of Representatives on the 16 primaries. Votes indicated by the CQA on each of the US. The CQA gathered nine different types of votes: voted for, paired for, and announced for (these three symbolized to yes). But, voted against, paired against, and announced against (these three simplified into no).voted present, to escape conflict of interest, and did not vote or even perform (these three simplified to an undefined provision), as in Table 1.

2) Bands dataset: A rotogravure printing classification query was in the shape of a cylinder unit where the aim was to define a given component. Such group of information was a UCI registry dataset, shown in Table 2.

TABLE I. CONGRESSIONAL VOTING RECORDS DATASET

Data Set Characteristics:	Multivariate	Number of Instances:	435	Area:	Social
Attribute Characteristics:	Categorical	Number of Attributes:	16	Date Donated	1987-04-27
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits	217885

TABLE II. BANDS DATASET

Data Set Characteristics:	Multivariate	Number of Instances:	512	Area:	Physical
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	39	Date Donated	1995-08-01
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits	77629

B. Parameters Settings

Table 3 shows parameter settings used in this study. The tested and investigated models were qualified with five hundred repetitions. ANN's Input sheet based on No. It had thousands of hidden nodes since the helpful process was applied. This method demanded more hidden nodes than traditional algorithms. It had one output layer node introducing the 2-class. Haphazard Search CV algorithm was made to optimize the hyperparameters such like (number of estimators, max depth, ….) (the number of iterations…... … ten iterations were applied). A big notice was that the greatest value of parameters was displayed in RFC. Supportive Vector Classifier plus Radial Basis Function were appeared as a Kernel. Manual tuning of hyper-parameters was served to help in choosing the support vector machine.

C. Performance Evaluation Criteria

The presented comparative models were tested depending on three variables of evaluation. Those parameters measured precision, recall, and accuracy f1 grading for both arrangement as well as research. The assessment parameters were judged as follows:

Confusion matrix:

A hesitation matrix was indicated as a table used to determine a classifier's outcomes according to a chain of data investigation to ensure the real values (i.e., the actual positives and negatives) which were admitted.

- Precision

Precision was calculated by the ratio of results obtained via the system. It could accurately detect positive observations (True Positives) compared to the entire positive notice gained by the system, both right (True Positives) or wrong (False Positives). The accuracy equation was:

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{8}$$

- Recall

The recall was the ratio of the results derived from the system-compared to all real malicious class (Actual Positives) and which in turn correctly expecting positive observations (True Positives). Hence, the recall ratio in the equation was this:

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{9}$$

- Accuracy

Precision was the most observant measure for performance. This was what many people were taught at school regardless accuracy, remember, and F1 ranking.

In short, accuracy was a change of the exacted evaluated classifications (both True Positives + True Negatives) for the whole Research Dataset. The accuracy ratio was stated:

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + false\ negatives + true\ negatives} \tag{10}$$

- F1 Score

The F1 Score was the range of Precision and Recall's weight (or balanced mean). Consequently, to break such an equilibrium between recall and accuracy. This score had to be under focus on both false positives and false negatives. The F1 value ratio in the formula was this:

$$F1\ score = \frac{2*(precision*recall)}{precision + recall} \tag{11}$$

TABLE III. PARAMETER SETTINGS

Model	Parameter	Values
ANN	Input nodes	based on No. Features
	Hidden nodes	1024
	Activation fun for hidden nodes	ReLU Rectified Linear Unit
	Activation fun for output nodes	sigmoid
	Output nodes	1
	No. of Iterations	500
RFC	n_estimators	1700
	max_depth	50
	min_samples_leaf	6
	class_weight	balanced
	random_state	1
SVM	n_estimators	700
	max_depth	110
	min_samples_leaf	6
	class_weight	balanced
	random_state	1

D. Model Selection

Three types of models were selected for grading (ANN – RFC - SVM). Via try/error, the most perfect hyperparameter for all models were picked out.

- As For the model of artificial neural network [27][28], its interior design mainly had Input layer, two hidden layers, and output layer.
- The input Layer based on no features.
- First hidden layer had 1024 neuron and activation mechanism was ReLU Rectified Linear Unit.
- The second hidden layer had 512 neurons, and the activation mechanism is ReLU Rectified Linear Unit.
- The output layer was 1 neuron since its class' classification problem and activation function were Sigmoid function.
- Random Forest Classifier [29].

The haphazard Search CV algorithm was applied to reach the hyperparameters to the max like (number of estimators, max depth, ….). It also supplied with the number of repetitions. Ten iterations were used, the best obtained value of parameters would be used in RFC.

- Support Vector Machine [30]

Support vector classifier had to apply the radial basis function as a Kernel. Also, a manual tuning of hyper-parameters had to choose the special support vector machine.

E. Methodology and Discussion

The proposed methodology flowchart is shown in Fig. 5:

1) Dataset had to be valued a well as data preprocessing had to be applied. It was considered to be the most important step I order to create a clear ready data under usage.

2) A preprocessor with instructing data had to perform the following steps:

a) Deleting rows which contained any missing values from dataset.

b) Introducing the approach of "missing values ratio" algorithm.

c) Listing categorical values.

d) DE normalizing data (performed feature scaling).

3) If dataset couldn't be examined, only a division of 66%ran as dataset training samples, and 33% as testing ones.

4) Selection of model had to be applied.

5) Three (NN – RFC - SVM) models were selected for Classification.

6) Try/error was chosen as the best hyperparameter for all models.

7) Training and test dataset had to be proved on the selected model.

8) At last, both of detailed analysis and a comparison between results had to be applied for different models and different datasets.

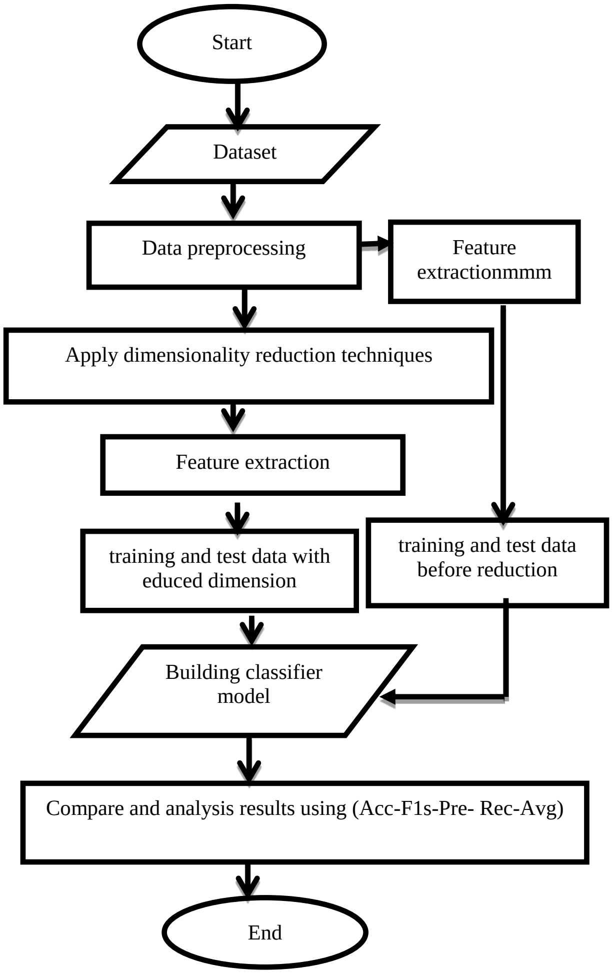

Start
Dataset
Data preprocessing
Feature extractionmmm
Apply dimensionality reduction techniques
Feature extraction
training and test data with educed dimension
training and test data before reduction
Building classifier model
Compare and analysis results using (Acc-F1s-Pre- Rec-Avg)
End

Fig. 5. Proposed Methodology Flowchart.

V. Results

This quoted part showed how to find the best parameters and performance for the nine dimensionality reduction algorithms which were applied on two various datasets. It shows the bar-chart for each dataset after we have selected a comparison between the final results in a table had taken place to get the best values for the parameter in reduction methods.

A. Missing-Values Ratio

Concerning the minimum values for the ratio, the best selected one could attain the best performance. As soon as the threshold value was decreased, the number of characteristics declined with a contrast for the performance of the model which increased. By avoiding the overfitting, as revealed in Figure 6, 7, the result would be a minimum value applied on two different datasets.

B. Low-Variance Filter

With regard to this variance filter, the best minimum values would reach the best performance. As soon as the threshold value was decreased, the number of characteristics declined with a contrast for the performance of the model which increased. By avoiding the overfitting, as revealed in Figure 8, 9, the result would be a minimum value applied on two different datasets.

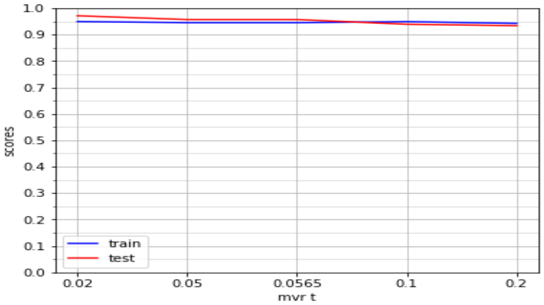


Fig. 6. Result of MVR for “Congressional Voting Records” Dataset.

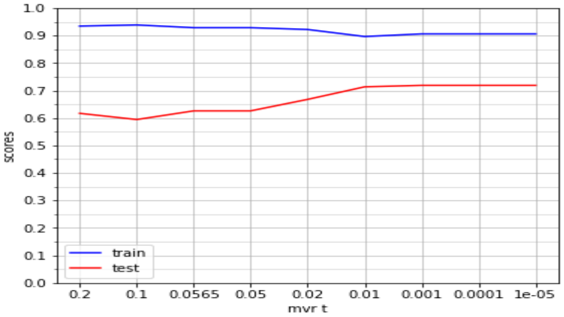


Fig. 7. Result of MVR for “Bands” Dataset

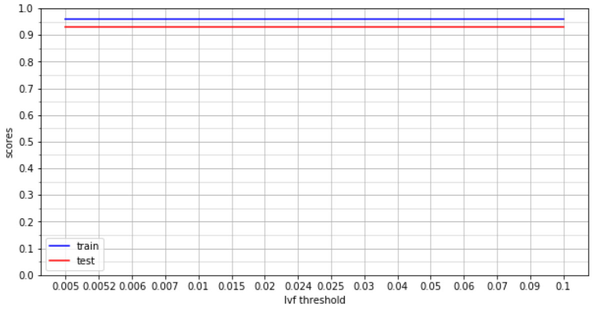


Fig. 8. Result of Low Variance Filter for “Congressional Voting Records” Dataset.

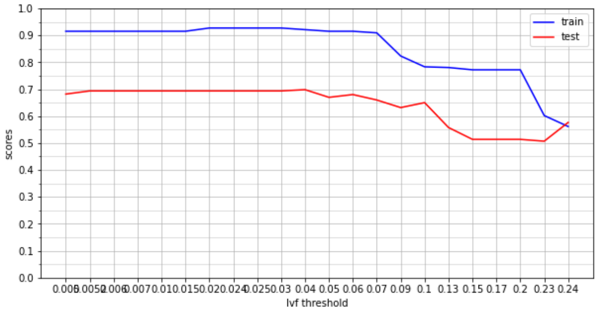


Fig. 9. Result of Low Variance Filter for “Bands” Dataset.

C. High-Correlation Filter

Following the high correlation filter, threshold values would be chosen to have the best attitude as conveyed in the figures. Whenever the threshold value was minimized, the no. features were decreasing whereas the performance of model was increasing. And this didn’t allow the overfitting point to take place. Thus, no features were the same of the original dataset. As cleared in Figure 10, 11. It provided with threshold which was applied on two different datasets.

D. Random Forest

In a random forest, an enormous accurate built chain of trees was created against achieving the highest features. Therefore, the usage statistics of each characteristic was used to obtain the greatest instructive subset of feature s. Figure 12 and 13 illustrated threshold which was applied on two different datasets.

E. Principal Component Analysis

Minimizing No. Features/Dimensions and applying PCA on them in order to choose the best value for PCA. From ex-knowledge, no. PCA was probably from 1: n-1 features. Figure 14 and 15 applied such a minimization on two different datasets.

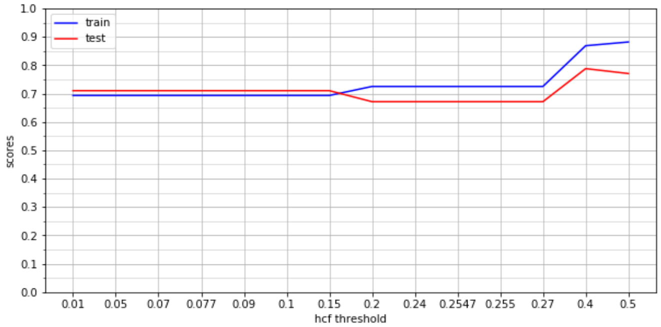


Fig. 10. Result of High-Correlation Filter for “Congressional Voting Records” Dataset.

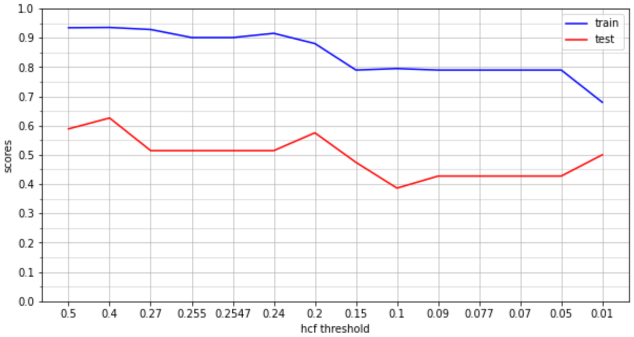


Fig. 11. Result of High-Correlation Filter for “Bands” Dataset.

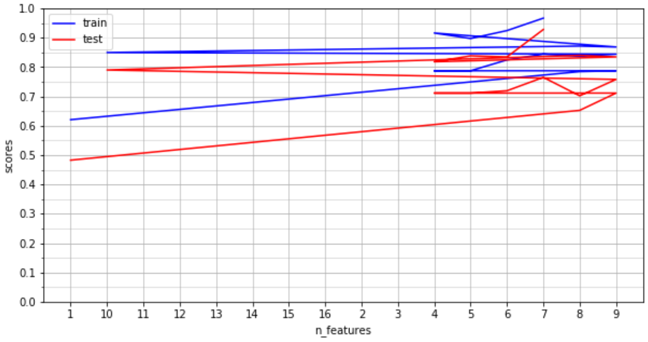


Fig. 12. Result of Random Forest for “Congressional Voting Records” Dataset.

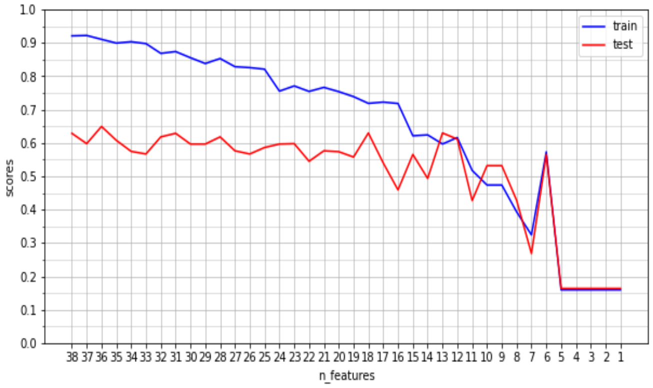


Fig. 13. Result of Random Forest for “Bands” Dataset.

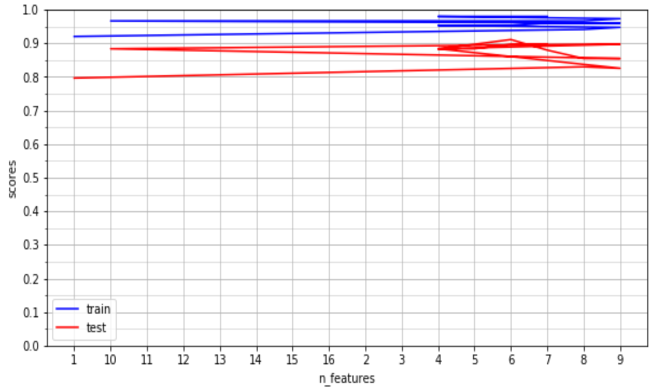


Fig. 14. Result of Principal Component Analysis for “Congressional Voting Records” Dataset.

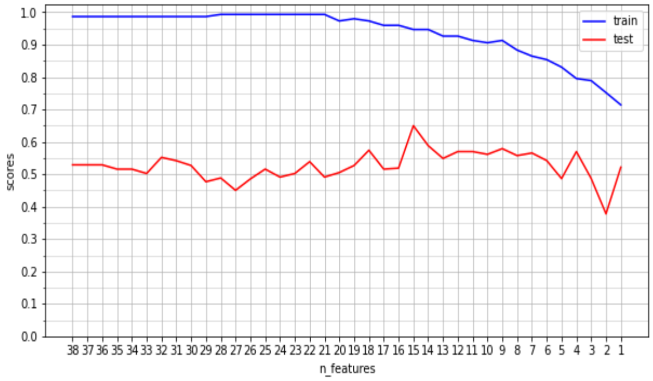


Fig. 15. Result of Principal Component Analysis for “Bands” Dataset.

F. Linear Discriminant Analysis

LDA was carried out to get fitting training data by giving new sized area since its reduction technique built on maximizing the class severability. If there were two classes No. LDA will be 1. The minimum value was applied on two different datasets as described in figure 16 and 17.

G. Backward Feature Elimination

Gradual decline of No Features with a threshold value to reach the best features and threshold values that achieve the best performance. As shown in Figure 18 and 19 applied on 2 different datasets.

H. Forward Feature Construction

The definition of forward feature construction was the opposite of the backward technique. It occurred by raising No. Features with a threshold value to have the best characteristics and the threshold value which achieved the perfect performance. As shown in Figure 20 and 21, it was applied on two different datasets.

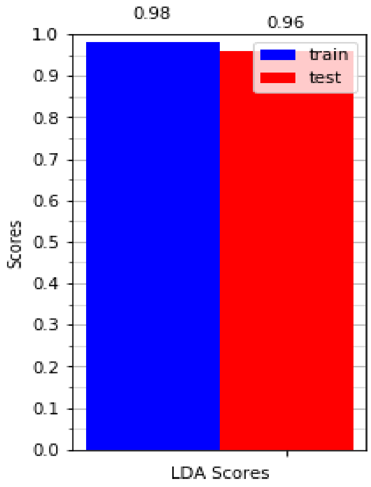


Fig. 16. Result of LDA for “Congressional Voting Records” Dataset.

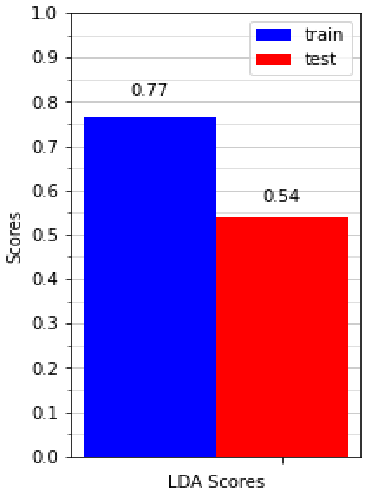


Fig. 17. Result of Linear Discriminant Analysis for “Bands” Dataset.

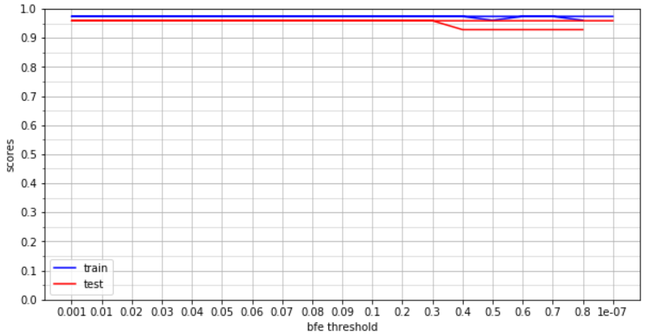


Fig. 18. Result of Backward Feature Elimination for “Congressional Voting Records” Dataset.

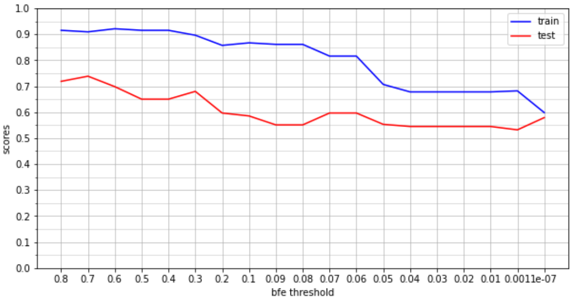


Fig. 19. Result of Backward Feature Elimination for “Bands” Dataset.

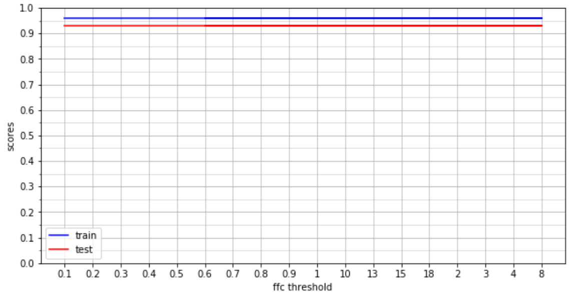


Fig. 20. Result of Forwarding Feature Construction For “Congressional Voting Records” Dataset.

I. *For “Congressional Voting Records” Dataset*

The detailed analysis of tables 4 and 5, the unique reduction techniques were Low-Variance Filter (LVF) in the training dataset and Missing-values Ration (MVR) in the test dataset. In tables six, the greatest reduction approach was Linear Discriminant Analysis (LDA) in both training & testing dataset. Performance, here, was so near compared to the actual overfitting which was slightly reduced (it’s challenging to minimize overfitting without losing preciseness because the dataset was very tiny.

The outcomes of all techniques with average results as (avg) from three models on training & testing dataset (Congressional Voting Records Dataset). The best reduction method was applied on the Missing-Values Ratio symbolized (MVR) as shown in Figure 22.

- The number of features was reduced from 16 to 10
- Reduction percentage = 62.5%
- Performance was improved by 3% (from 94% to 97%)
- Overfitting was decreased by 2%

The result of the test score for the three models for each reduction technique (Congressional Voting Records Dataset) for the Best model was NN as described in Figure 23.

J. *For “Bands” Dataset*

In Tables 7 and 8, the best reduction method was Principal Component Analysis (PCA) in the training dataset and Random Forest (RFC) in the testing dataset. But for Table 9, the best reduction method was the Principal Component Analysis (PCA) in the training dataset and Missing-Values Ratio (MVR) in the testing dataset. At the same time performance was improved and overfitting was slightly decreased.

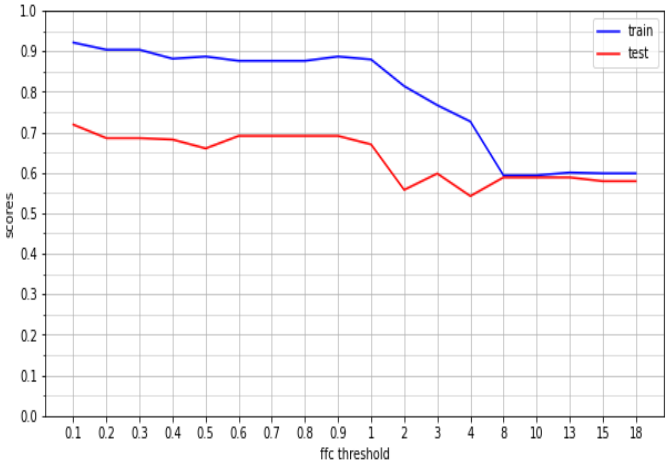


Fig. 21. Result of Forwarding Feature Construction for “Bands” Dataset.

TABLE IV. RESULTS OF ALL TECHNIQUES WITH ANN MODEL ON TRAINING & TESTING DATASET

		ANN									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	1.00	1.00	1.00	1.00	1.00	0.96	0.96	1.00	0.92	0.96
2	Missing-Values Ratio	0.98	0.98	0.99	0.96	0.98	0.99	0.99	1.00	0.98	0.99
3	Low-Variance Filter	0.99	0.99	1.00	0.99	0.99	0.96	0.96	1.00	0.92	0.96
4	High-Correlation Filter	0.67	0.63	0.57	0.69	0.64	0.65	0.63	0.73	0.56	0.64
5	Random Forest	0.90	0.89	0.83	0.95	0.89	0.85	0.80	0.73	0.89	0.82
6	Principal Component Analysis	0.98	0.98	0.97	0.99	0.98	0.95	0.94	0.97	0.91	0.94
7	Linear Discriminant Analysis	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
8	Backward Feature Elimination	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
9	Forward Feature Construction	0.98	0.98	0.99	0.97	0.98	0.95	0.94	0.94	0.94	0.94
10	Rough-set	0.94	0.92	0.92	0.92	0.92	0.89	0.86	0.90	0.82	0.87

TABLE V. RESULT OF ALL TECHNIQUES WITH RFC MODEL ON TRAINING & TEST DATASET

		RFC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.96	0.96	0.97	0.95	0.96	0.94	0.93	0.94	0.91	0.93
2	Missing-Values Ratio	0.95	0.94	0.97	0.92	0.95	0.97	0.95	0.98	0.93	0.96
3	Low-Variance Filter	0.96	0.96	0.97	0.95	0.96	0.94	0.93	0.94	0.91	0.93
4	High-Correlation Filter	0.69	0.73	0.85	0.63	0.73	0.62	0.66	0.88	0.53	0.67
5	Random Forest	0.92	0.92	0.95	0.89	0.92	0.84	0.81	0.85	0.78	0.82
6	Principal Component Analysis	0.98	0.98	0.97	0.99	0.98	0.91	0.89	0.88	0.91	0.90
7	Linear Discriminant Analysis	0.98	0.98	0.99	0.97	0.98	0.96	0.96	1.00	0.92	0.96
8	Backward Feature Elimination	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
9	Forward Feature Construction	0.96	0.96	0.97	0.95	0.96	0.94	0.93	0.94	0.91	0.93
10	Rough-set	0.93	0.91	0.93	0.88	0.91	0.87	0.84	0.90	0.79	0.85

TABLE VI. RESULT OF ALL TECHNIQUES WITH SVC MODEL ON TRAINING & TEST DATASET

		SVC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.98	0.98	0.99	0.97	0.98	0.94	0.93	0.94	0.91	0.93
2	Missing-Values Ratio	0.98	0.97	0.98	0.96	0.97	0.97	0.97	1.00	0.93	0.97
3	Low-Variance Filter	0.98	0.98	0.99	0.97	0.98	0.94	0.93	0.94	0.91	0.93
4	High-Correlation Filter	0.69	0.73	0.85	0.63	0.73	0.62	0.66	0.88	0.53	0.67
5	Random Forest	0.95	0.95	0.97	0.94	0.95	0.82	0.78	0.76	0.81	0.79
6	Principal Component Analysis	0.98	0.98	0.99	0.97	0.98	0.94	0.93	0.94	0.91	0.93
7	Linear Discriminant Analysis	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
8	Backward Feature Elimination	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
9	Forward Feature Construction	0.97	0.97	0.97	0.97	0.97	0.94	0.93	0.94	0.91	0.93
10	Rough-set	0.96	0.95	0.97	0.94	0.96	0.89	0.86	0.92	0.81	0.87

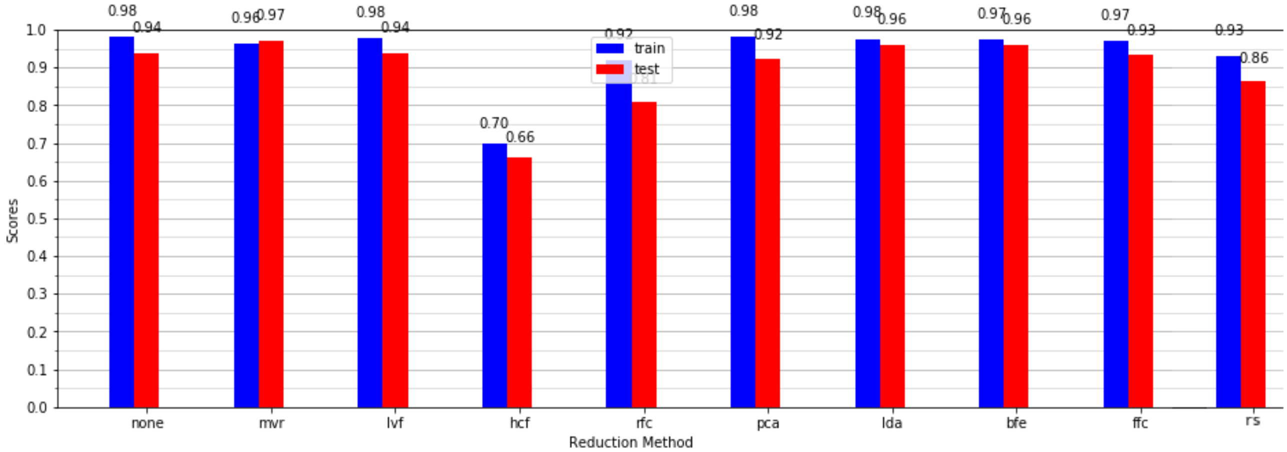


Fig. 22. Result of All Techniques with Average Results of (AVG) from Three Models on Training & Testing Dataset.

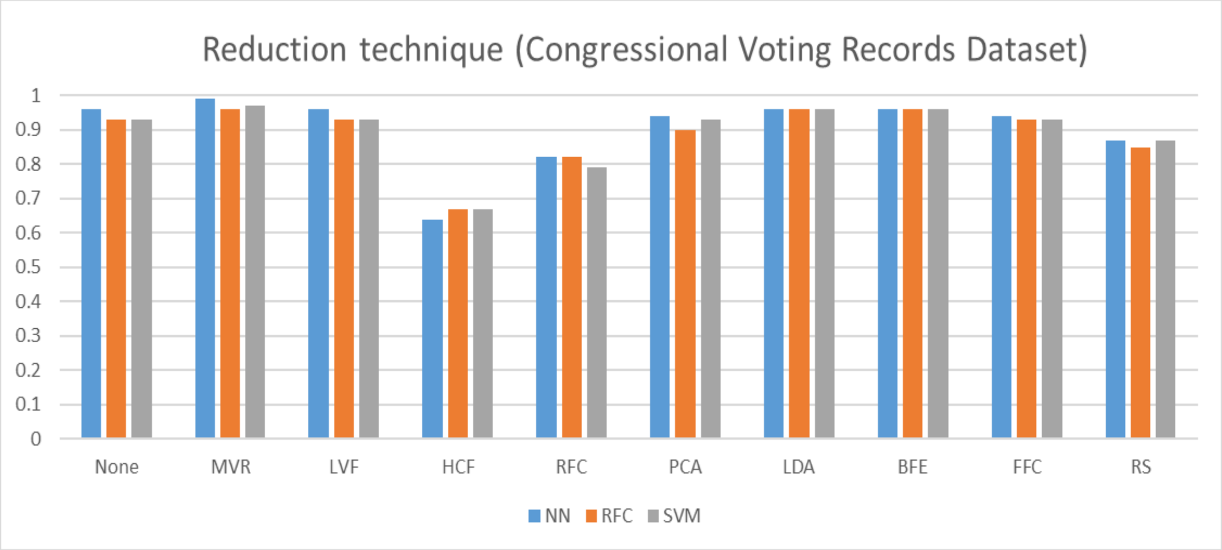


Fig. 23. Test Score for Three Models for each Reduction Technique (Congressional Voting Records Dataset).

TABLE VII. RESULT OF ALL TECHNIQUES WITH THE ANN MODEL ON TRAINING & TESTING DATASET

		ANN									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.74	0.72	0.94	0.58	0.75	0.64	0.56	0.66	0.49	0.59
2	Missing-Values Ratio	0.65	0.70	0.97	0.55	0.72	0.51	0.62	0.97	0.46	0.64
3	Low-Variance Filter	0.64	065	0.90	0.50	0.67	0.56	0.51	066	0.41	0.53
4	High-Correlation Filter	0.67	0.58	0.61	0.54	0.60	0.54	0.26	0.24	0.29	0.33
5	Random Forest	0.63	0.62	0.81	0.50	0.64	0.68	0.58	0.66	0.53	0.61
6	Principal Component Analysis	0.76	0.73	0.91	0.61	0.75	0.67	0.53	0.55	0.52	0.57
7	Linear Discriminant Analysis	0.76	0.69	0.74	0.65	0.71	0.67	0.50	0.48	0.52	0.54
8	Backward Feature Elimination	0.63	0.58	0.70	0.49	0.60	0.67	0.55	0.59	0.52	0.58
9	Forward Feature Construction	0.60	0.59	0.80	0.47	0.61	0.52	0.47	0.62	0.38	0.50
10	Rough-set	0.62	0.41	0.37	0.48	0.47	0.55	0.36	0.37	0.35	0.41

TABLE VIII. RESULT OF ALL TECHNIQUES WITH RFC MODEL ON TRAINING & TESTING DATASET

		RFC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.89	0.86	0.94	0.79	0.87	0.69	0.55	0.55	0.55	0.59
2	Missing-Values Ratio	0.78	0.76	0.82	0.71	0.77	0.66	0.66	0.80	0.56	0.67
3	Low-Variance Filter	0.73	0.70	0.89	0.58	0.72	0.63	0.55	0.66	0.47	0.58
4	High-Correlation Filter	0.74	0.66	0.70	0.63	0.68	0.56	0.24	0.21	0.30	0.33
5	Random Forest	0.67	0.64	0.83	0.53	0.67	0.67	0.58	0.66	0.51	0.60
6	Principal Component Analysis	0.84	0.81	0.93	0.72	0.83	0.67	0.53	0.55	0.52	0.57
7	Linear Discriminant Analysis	0.76	0.68	0.73	0.65	0.70	0.67	0.48	0.45	0.52	0.53
8	Backward Feature Elimination	0.63	0.58	0.70	0.49	0.60	0.67	0.55	0.59	0.52	0.58
9	Forward Feature Construction	0.64	0.58	0.69	0.50	0.60	0.68	0.56	0.59	0.53	0.59
10	Rough-set	0.69	0.70	0.63	0.79	0.70	0.56	0.60	0.55	0.66	0.59

TABLE IX. RESULT OF ALL TECHNIQUES WITH SVC MODEL ON TRAINING & TEST DATASET

		SVC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.94	0.92	0.93	0.90	0.92	0.75	0.63	0.62	0.64	0.66
2	Missing-Values Ratio	0.92	0.90	0.90	0.90	0.91	0.74	0.71	0.76	0.67	0.72
3	Low-Variance Filter	0.81	0.77	0.87	0.69	0.78	0.71	0.62	0.69	0.57	0.65
4	High-Correlation Filter	0.91	0.87	0.87	0.87	0.88	0.71	0.52	0.45	0.45	0.58
5	Random Forest	0.63	0.57	0.69	0.49	0.60	0.70	0.60	0.66	0.56	0.63
6	Principal Component Analysis	0.96	0.94	0.94	0.94	0.95	0.74	0.62	0.62	0.62	0.65
7	Linear Discriminant Analysis	0.81	0.75	0.77	0.73	0.73	0.68	0.49	0.45	0.54	0.54
8	Backward Feature Elimination	0.63	0.58	0.70	0.49	0.60	0.67	0.55	0.59	0.52	0.58
9	Forward Feature Construction	0.63	0.57	0.69	0.49	0.59	0.68	0.56	0.59	0.53	0.59
10	Rough-set	0.86	0.87	0.84	0.91	0.87	0.68	0.50	0.47	0.53	0.55

The result of all techniques with average results of (avg) from three models on training and testing dataset (Bands Dataset), applying the best reduction method, was Missing-Values Ratio (MVR) as reflected in Figure 24.

- The number of features reduced from 38 to 12
- Reduction percentage = 68%
- Performance was improved by 7% (from 66% to 72%)
- Overfitting was reduced

The result of the test score for the 3 models for each reduction technique (Bands Dataset) applying the best model was RFC as shown in Figure 25 reduction techniques. Those techniques were called as follows: missing-values ratio, low variance filter, high correlation filter, random forest, key component analysis, linear discriminant analysis, removal of backward function, construction of forwarding features, and rough set theory. It was observed how dimensionality reduction could be useful in minimizing overfitting as well as getting the perfect performance. The models' average performance was tested in two various datasets for three different models, as cleared in Table 10.

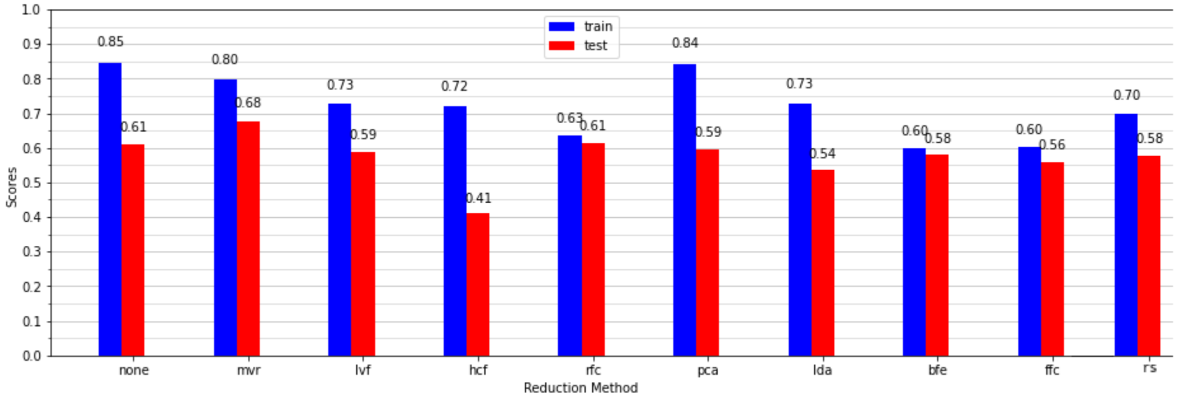


Fig. 24. Result of All Techniques with Average Results of (AVG) from Three Models on Training & Testing.

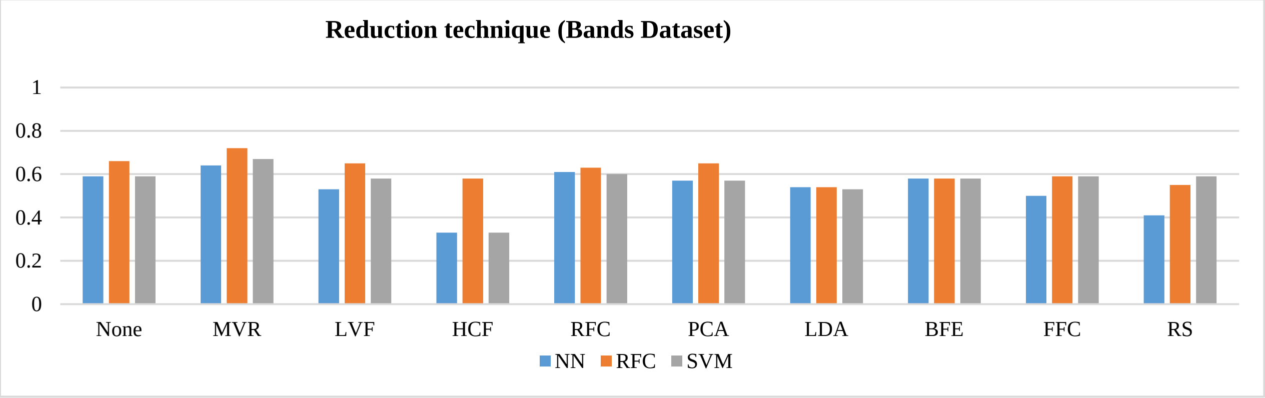


Fig. 25. The Test Score of Three Models for each Reduction Technique (Bands Dataset) Described a Number of Nine-Dimensional.

TABLE X. PERFORMANCE EVALUATION OF DIFFERENT MODELS ON TEST DATASET

		Congressional Voting Records Database			Bands Database		
		ANN	RFC	SVC	ANN	RFC	SVC
		Avg	Avg	Avg	Avg	Avg	Avg
1	Before Reduction	0.96	0.93	0.93	0.59	0.59	0.66
2	Missing-Values Ratio	0.99	0.96	0.97	0.64	0.67	0.72
3	Low-Variance Filter	0.96	0.93	0.93	0.53	0.58	0.65
4	High-Correlation Filter	0.64	0.67	0.67	0.33	0.33	0.58
5	Random Forest	0.82	0.82	0.79	0.61	0.60	0.63
6	Principal Component Analysis	0.94	0.90	0.93	0.57	0.57	0.65
7	Linear Discriminant Analysis	0.96	0.96	0.96	0.54	0.53	0.54
8	Backward Feature Elimination	0.96	0.96	0.96	0.58	0.58	0.58
9	Forward Feature Construction	0.94	0.93	0.93	0.50	0.59	0.59
10	Rough-set	0.87	0.85	0.87	0.41	0.59	0.55

VI. CONCLUSION

This paper discussed nine-dimensional reduction techniques, and their effect on overfitting problem. These techniques are namely, missing-values ratio, low variance filter, high correlation filter, random forest, key component analysis, linear discriminant analysis, removal of backward function, construction of forwarding features, and rough set theory respectively. These techniques are valuable in reducing overfitting as well as obtaining a quite accepted performance. The used techniques were compared in both training and testing performance on two different datasets with three different models (ANN, SVM, and RFC). Performance was so close to the original for the RFC model. Missing-values ratio was closer to the removal of backward feature. The datasets got reduced to almost half of their original size; that allows machine-learning models to work faster on the datasets, which was another advantage of dimensionality reduction.

Some improvements on used models will be added by using metaheuristic optimization algorithms to find the best solution.

REFERENCES

[1] C. L. Blake, "CJ Merz UCI repository of machine learning databases," Ph.D. dissertations, Dept. Inform. Comput. Sci., Univ. California, Irvine, CA, USA, 1998.

[2] O. Deniz, A. Pedraza, N. Vallez, J. Salido, and G. Bueno, "Robustness to adversarial examples can be improved with overfitting". International Journal of Machine Learning and Cybernetics, 1-10, 2020.

[3] A. Zheng and A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, Newton, MA, USA:O'Reilly Media, 2018.

[4] X. Huang, L. Wu, and Y. Ye, "A Review on Dimensionality Reduction Techniques," International Journal of Pattern Recognition and Artificial Intelligence, vol. 33, no. 10, p. 1950017, 2019.

[5] A. Juvonen, T. Sipola, T. Hämäläinen, Online anomaly detection using dimensionality reduction techniques for http log analysis, Comput. Netw. 91 (2015) 46–56.

[6] M. Verleysen, D. François, The Curse of Dimensionality in Data Mining and Time Series Prediction, in: International Work-Conference on Artificial Neural Networks, Springer, 2005, pp. 758–770.

[7] T. Lesort , N. Díaz-Rodríguez , J.-F. Goudou , D. Filliat , State representation learning for control: an overview, Neural Netw. 108 (2018) 379–392 .

[8] C. Meng, O.A. Zeleznik, G.G. Thallinger, B. Kuster, A.M. Gholami, A.C. Culhane, Dimension reduction techniques for the integrative analysis of multi-omics data, Brief. Bioinform. 17 (4) (2016) 628–641.

[9] J.P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: survey, insights, and generalizations, J. Mach. Learn. Res. 16 (1) (2015) 2859–2900.

[10] L. Xie, Z. Li, J. Zeng, U. Kruger , Block adaptive kernel principal component analysis for nonlinear process monitoring, AlChE J. 62 (12) (2016) 4334–4345.

[11] A. Akkalkotkar, K.S. Brown, An algorithm for separation of mixed sparse and gaussian sources, PloS one 12 (4) (2017) e0175775.

[12] S. Deegalla , H. Boström , K. Walgama , Choice of Dimensionality Reduction Methods for Feature and Classifier Fusion with Nearest Neighbor Classifiers, in: 15th International Conference on Information Fusion (FUSION), IEEE, 2012, pp. 875–881.

[13] S. Ahmadkhani, P. Adibi, Face recognition using supervised probabilistic principal component analysis mixture model in dimensionality reduction without loss framework, IET Comput. Vision 10 (3) (2016) 193–201.

[14] I.T. Jolliffe , J. Cadima , Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202.

[15] NB. Erichson, P. Zheng, K. Manohar, S.L. Brunton, J.N. Kutz, A.Y. Aravkin, Sparse principal component analysis via variable projection. arXiv preprint arXiv:1804.00341.

[16] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. A. Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," Computers & Electrical Engineering, vol. 70, pp. 871-882, 2018.

[17] S. Chormunge and S. Jena, "Correlation-based feature selection with clustering for high dimensional data," Journal of Electrical Systems and Information Technology, vol. 5, no. 3, pp. 542-549, 2018.

[18] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 2264-2268: IEEE.

[19] B. Liu, Y. Li, L. Li, and Y. Yu, "An approximate reduction algorithm based on conditional entropy." In International Conference on Information Computing and Applications (pp. 319-325). Springer, Berlin, Heidelberg, 2010.

[20] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. A. Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," Computers & Electrical Engineering, vol. 70, pp. 871-882, 2018.

[21] S. Hu , Y. Gu , H. Jiang , Study of Classification Model for College Students' M-learn- ing Strategies Based on Pca-lvq Neural Network, in:

8th International Conference on Biomedical Engineering and Informatics (BMEI), IEEE, 2015, pp. 742–746.

[22] A.R. Santos, M.A. Santos , J. Baumbach , J.A.M. Culloch , G.C. Oliveira , A. Silva , A. Miyoshi , V. Azevedo, A Singular Value Decomposition Approach for Improved Taxonomic Classification of Biological Sequences, in: BMC Genomics, volume 12, BioMed Central, 2011, p. S11.

[23] A. Swati, and R. Ade, "Dimensionality reduction: an effective technique for feature selection". Int. J. Comput. Appl. 117(3), 18–23, 2015.

[24] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," Information Fusion, vol. 59, pp. 44-58, 2020.

[25] P. Mills, Singular value decomposition (svd) tutorial: Applications, examples, exercises, 2017, https://blog.statsbot.co/singular-value-decomposition- tutorial-52c695315254, (Accessed on 09/04/2019).

[26] I. Brigadir, D. Greene, J. P. Cross, and P. Cunningham, "Dimensionality Reduction and Visualisation Tools for Voting Record." In 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Ireland, 20-21 September 2016. CEUR Workshop Proceedings

[27] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics." Cancer Genomics-Proteomics, 15(1), 41-51, 2018.

[28] J. Rahmanishamsi , A. Dolati , M.R. Aghabozorgi , A copula based algorithm and its application to time series clustering, J. Classif. 35 (2) (2018) 230–249.

[29] Q. Hu, L. Zhang, Y. Zhou & W. Pedrycz, "Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets." IEEE Transactions on Fuzzy Systems

[30] (1), 226-238, 2018.

[31] A. Zeng, T. Li, D. Liu, J. Zhanga and H. Chen, " A fuzzy rough set approach for incremental feature selection on hybrid information systems" Fuzzy Sets and Systems, 258, 39-60, 2015.