Journal of Techniques, ISSN: 2708-8383, Vol. 5, No. 1, 2023, Pages 184-192
DOI: https://doi.org/10.51173/jt.v5i1.1027

JOURNAL OF TECHNIQUES

Journal homepage: http://journal.mtu.edu.iq

RESEARCH ARTICLE - MANAGEMENT

Comparison of Feature Selection and Feature Extraction Role in Dimensionality Reduction of Big Data

Haidar Khalid Malik [1*], Nashaat Jasim Al-Anber [1], Fuad Abdo Esmail Al-Mekhlafi [2]

[1] Technical College of Management - Baghdad, Middle Technical University, Baghdad, Iraq

[2] Faculty of Commerce and Economics, Sana'a University, Sana'a, Yemen

[*] Corresponding author E-mail: dac0003@mtu.edu.iq

Article Info.

Article history:

Received
23 October 2022

Accepted
11 December 2022

Publishing
31 March 2023

Abstract

Recently, researchers intensified their efforts on a dataset with a large number of features named Big Data because of the technological revolution and the development in the data science sector. Dimensionality reduction technology has efficient, effective, and influential methods for analyzing this data, which contains many variables. The importance of Dimensionality Reduction technology lies in several fields, including "data processing, patterns recognition, machine learning, and data mining". This paper compares two essential methods of dimensionality reduction, Feature Extraction and Feature Selection Which Machine Learning models frequently employ. We applied many classifiers like (Support vector machines, k-nearest neighbors, Decision tree, and Naive Bayes) to the data of the anthropometric survey of US Army personnel (ANSUR 2) to classify the data and test the relevance of features by predicting a specific feature in USA Army personnel results showing that (k-nearest neighbors) achieved high accuracy (83%) in prediction, then reducing the dimensions by several techniques like (Highly Correlated Filter, Recursive Feature Elimination, and principal components Analysis) results showing that (Recursive Feature Elimination) have the best accuracy by (66%), From these results, it is clear that the efficiency of dimension reduction techniques varies according to the nature of the data. Some techniques are more efficient than others in text data and others are more efficient in dealing with images.

Publisher: Middle Technical University

Keywords: Feature Extraction; Feature Selection; Principal Component Analysis (PCA); Dimensionality Reduction.

1. Introduction

In the current digital world, massive data are created from all sectors like Medical, administrative, industrial, etc. One of the reasons for using Machine learning algorithms is to find meaningful patterns for this data, Which in turn makes it helpful in executive and business decisions [1]. Numerous studies have been conducted on the subject of feature selection and feature extraction for various objectives, like "clustering, classification, and prediction", which is considered the aim of many research works which has datasets that can contain hundreds or thousands of variables (Features), to implement that we use "Dimensionality Reduction" as a pre-processing step to minimizing training time and improving the accuracy of learning features [2]. (FS) methods are widely utilized for selecting the variables (features) with big datasets' most related and valuable characteristics. The main difference between (FS and FE) processes is that feature selection methods are utilized to accomplish the subgroup of the most relevant attributes without duplicating them. FE methods are used to reduce dimensionality by incorporating existing features [3]. ANSUR 2 (The Anthropometric Survey of US Army Personnel) datasets were used in this work, which consists of 109 attributes and 6068 rows. This data was published in 2012 and became available publicly in 2017. [4] this work considers the advantage of feature extraction, where the original variables are preserved but processed into smaller groups to keep as much as possible in the information from original data and feature selection, which deletes input features that do not participate effectively in model performance. [5] a significant task of (FS and FE) reducing the dimensions in datasets with huge variables leads to minimizing model complication and overfitting. The critical dissimilarity between (FS and FE) techniques throw the reduction process that (FS) keeps the original attributes. As for extracting features, it results in complex features that differ from the features of the original data. From this point, we can say the aim of both techniques is summing up three points (1) Decreasing the quantity of the data, (2) Dealing with the correlated variables, (3) amelioration data quality will improve data mining algorithm performance, including learning time and predictive accuracy. [6] both feature extraction and feature selection positively impact learning efficiency, enhancing the effectiveness of computing, reducing memory usage, and improving generalization models. Therefore, they are considered efficient dimensionality reduction methods. Many applications such as "text mining and genetic analysis" prefer feature selection because it keeps the Infrastructure of the original dataset and improves the readability and interpretability of models [7].

In this paper, we presented a comparative study between Features selection techniques and Features Extraction techniques, three-stage applied, first stage ANSUR 2 datasets have measurements of males and females in the USA army. This dataset contains many missing values that should be removed, which is considered pre-processing.

Nomenclature			
PCA	Principal Component Analysis	SNE	t-distributed Stochastic Neighbor Embedding
LDA	Linear Discriminant Analysis	MDS	Multidimensional Scaling
FS	Features selection	Isomap	Isometric mapping
FE	Features extraction	SVD	Singular Value Decomposition
SVM	Support vector machines	MVR	Missing Values Ratio
KNN	k-nearest neighbours	NB	naive Bayes
DT	Decision tree	DR	Dimensionality Reduction

The second stage is classifying the data using (Support Vector Machine, Decision Tree, Random Forest, Neural Network, and K-Nearest Neighborhood) classifiers for Testing how correlated attributes are in the dataset by predicting features (Components*)* in the dataset.

The third stage is Reducing the Dimensions by three different techniques (Highly Correlated Filter, Recursive Feature Elimination, and principle component analysis). results of reduction indicated Features selection techniques (Recursive Feature Elimination) performed better than features extraction and kept most of the original dataset information. Technique, their types, and their use. And conclusion section contains a result table of Dimensionality Reduction techniques.

2. Related Work

In this part, we will review some of the articles that deal with the comparison between (the selection of features and extraction of features) Researchers in [8] compared (feature selection and feature extraction) using an eye diseases dataset. The results indicated that feature Extraction achieved better results than feature Selection due to noisy Data because biomedical data is very confusing. The dissimilarity between both (FS and FE) was also addressed, mentioning the techniques used in each type [9]. Various ways to reduce the dimensions of data using microarray data for cancer patients (microarray cancer data), various methods and techniques of (FS and FE) have been described and compared, and the pros and cons of both techniques were clarified [10]. Explaining both techniques (FS and FE), heart patient data was used in India, the main component analysis technique (PCA) was used in the data extraction method, and the Wrapper Filter method was used as a classifying to give the best results, as well as the system performance was improved compared to other recording function such as Euclidean distance and Pearson correlation coefficients [11]. The researchers studied the financial data of (the IBM Bluemix cloud platform) study aimed to reduce the dimensions of the data using (FS and FE) techniques for a large set of financial data, and the results showed that reducing the dimensions of the data led to a significant improvement in execution time without reducing accuracy, and (SVM Classifier) and (logistic regression classifier) were used [12]. Introducing a study to predict cancer disease through a hybrid method founded on (FS and FE) techniques (fast Fourier transform algorithm) to measure the density of the sample women and the average density of the women for both normal and cancer-infected cases was used as a method of extracting new features to achieve high classification accuracy and decreasing training time and to assess the reliability of the new hybrid method, Set of classifiers (naive Bayes, Random Forest, and support vector machine) have been applied to a variety of types of cancer diseases such as (Breast, Colon, and Head), The results presented that there is high precision in the classification and improvement in most cases [13] Introducing a novel approach that uses deep learning to estimate clinical outcomes for cancer patients, where a new algorithm called (AdaBoost algorithm) was applied to classify samples for prediction. It showed good results and was more accurate than other algorithms. Through investigation and discussion, The automatically generated features by neural networks showed an exceptional ability to improve performance and predict the result [3] Applying the classification process to the medical Passover data, the difference between (FS and FE) was also clarified. The process of (FS) improved the knowledge, accuracy, and education of the algorithm as required, which are applications in machine learning the most widespread and used, and some feature selection techniques used Widely used in lung cancer, breast cancer tumours, etc. [14]. Introducing a hybrid system that integrates both (FS and FE) and is based on the interaction of the neural network with the doctor who diagnoses the patient, this system was able to extract the most useful features without losing the physical sensation, except for reducing the dimensions connected to the internet, as well as simplifying the interaction between man and machine in the domain of mining medical data [15]. The results showed a decrease in classification accuracy when using a lower sample rate by studying the effects of the sample rate that gives electrical impulses to amputees to classify hand and finger movements. The SVM classifier was used to classify movements for 26 cases (with 12 movements), (17 movements) (23 movements) [16]. By presenting a hybrid method called “Hybrid Low Rank”, which is a matrix whose work is combined between feature selection and feature extraction, the algorithm was described using the best experimental research method, which consists of three scales and greedy (optimal, suboptimal). this experimental research technique made it possible to calculate the prior and subsequent limits, which shows how close to the optimal solution is [17] study the risk of stroke and sleep apnea, features Extraction was obtained from PSG, which is a comprehensive test for sleep disorders and is used to diagnose cases of sleep apnea, use the SVM classifier to determine the features of sleep disorders, use the features extraction feature with both (ECG signals, oxygen saturation signals, air flow signals, abdominal and chest signals). The dimensions were reduced by the selection of features (features selection) by (the SVM classifier) [18]. The researchers proposed a complete solution for fault diagnosis of an experimental transformer with nested neutral points within five levels by obtaining a signal, extracting and selecting features, and classification, and effective diagnosis is made for 36 lines in the open circuit. This work was done by diagnosing open circuit faults of semiconductor devices by analyzing circuit signals. experimental results have shown that the presented solution is characterized by high efficiency, flexibility, and a super error recognition rate [19]. The researchers presented a hybrid method relying on (FS and Grouped FE). This method was presented for multiple purposes (removing unrelated features and removing repetitive information between features). The results indicated that the novel approach has a competitive classification performance and is quicker. The results also demonstrate that the suggested approach has to be improved in terms of the loss of information for the methods. However, the loss of information in the proposed method does not lead to a significant decrease in classification performance [20]. The selection of features and extraction of features were compared, and a set of classifiers was used with (Features selection), including SVM, KNN, and the best results were achieved with (SVM) and on the other side (Features extraction) classifiers were used, including CNN, DNN, which had a big role and the main component analysis technique (PCA) was used [21].

The researchers presented a set of studies that apply (FS and FE) methods to predict temporal and spatial passage. the database that was used included 211 publications and for the period from 1984 to 2018, the benefits and drawbacks of various feature selection and feature extraction applications were clarified [22]. The experiment was conducted on 25 participants, including 18 males and 7 females using deep learning algorithms to create a classification model based on excitement and equivalence using the feature selection technique. To evaluate the performance of the selection of features, a set of features indexed for EEG according to certain criteria was selected and a model for the selection of features was created. the time taken for training was less for data whose dimensions were reduced while maintaining accuracy by 98% through the use of appropriate feature selection. The main goal is to find out the preference of any method over a previously unused data set by using three techniques that were not used on this type of data.

3. Features selection

It decreases the number of input attributes activity when developing a predictive model. Features selection solves processing time and complication problems, losing storage in memory problems. In other words, feature selection reduces the dimensions to a subsidiary set by removing attributes with little information. Hence, features selection chose the ideal subset of features from the big dataset. The new set decreased, still keeping most of the original dataset information [23] during implementation features selection, we should consider removing redundant features and getting out valuable information. The selection mechanism should consider; (1) there is no effect on the accuracy and performance, (2) the output subset should be similar to the original dataset. The best feature selection criterion when achieving data visualization, data understanding, and reducing the storage features selection mechanism consists of two main steps (feature generation). This stage of creating a subset from the enormous amount of data. Second (Feature evaluation) from its name to evaluate the subset generated to suit the requirements [24]. Fig. 1 show the features selection mechanism.

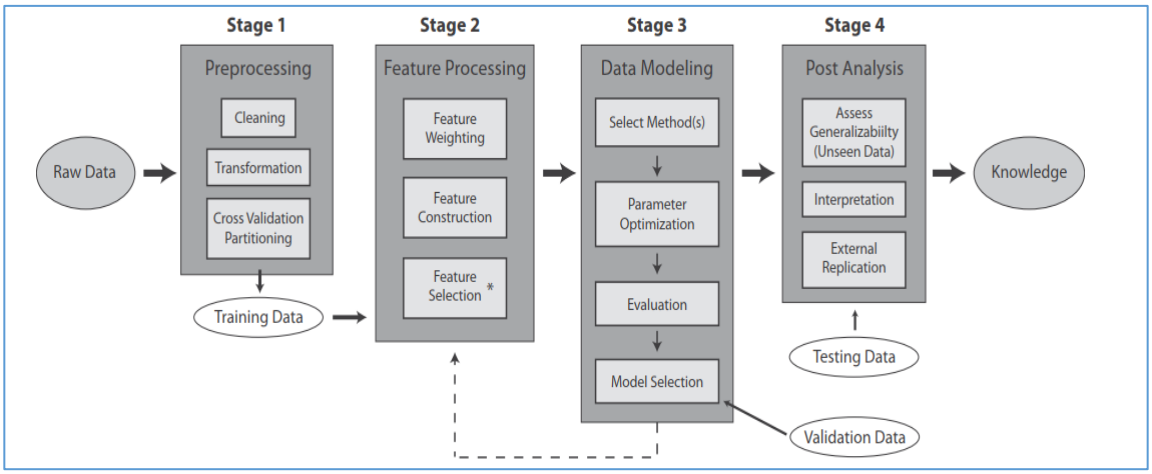

Fig. 1. Features the selection process steps[25]

3.1. Features selection methods

There are three main types in Features selection

3.1.1. Wrappers Methods

In wrapper methods, we use a subset of features and make a model using them by training. The wrappers mechanism chooses the best subset by using a model that records different subsets. A model is trained on each new subset, and its performance is then assessed on a hold-out set. Then select the best subset which achieves better model performance. Wrapper approaches typically offer the best feature set for the selected model type, which is a significant advantage [24].

3.1.2. Filters Methods

They can be viewed as a quicker and easier substitute for wrappers. They merely examine each feature's statistical relationship with the model's aim to determine its usefulness, substituting metrics like correlation or mutual information for the model performance metric. Filters principles do not deal with classifiers and consist of two (multivariate and univariate.). Multivariate methods find the relationship between the features, and Univariate methods deal with each feature independently [25].

3.1.3. Embedded Methods

We'll examine integrating feature selection into the learning process as our final method. This method idea combines the best characteristics (Wrappers and Filters). It has filter method speed and obtaining the best subset and optimal subset for the specific model, much like from a wrapper [25].

3.2. Features selection Models

Feature selection is a widely used approach that has been the subject of decades of technique and application research. Such as "image recognition, image retrieval, text mining, intrusion detection, bioinformatic data analysis, fault diagnosis, data analysis, fault diagnosis", and so on. Theoretically, statistics consider the basics of feature selection techniques and are classified under numerous standards [26].

3.2.1. Supervised feature selection

The relation between the features and the target variable (class label) is the primary criterion for supervised feature selection, frequently targeted to classification problems. It can compute which features in important by correlation measures. For a given dataset A= (Z, Y), with a feature set Z = {Z_1, Z_2,..., Z_n} and class label Y, the supervised model goal is finding an optimum feature subset Nˆ (| Nˆ | = Rˆ) which increasing the classification precision [26]. As we explained earlier, the types of feature selection in supervised (Filters, Wrappers, and Embedded).

3.2.2. Unsupervised Feature Selection

The unsupervised (FS) method aims to Find a feature subgroup relying on for each clustering or assessment standard to reinforce the accuracy of clustering and account for natural data classification. Depending on whether they use "cluster algorithms", unsupervised (FS) techniques can be either "unsupervised filter or wrapper techniques" [26]. Unsupervised (FS) is becoming a necessary pre-processing stage because it can decrease computational time greatly due to decreased feature subsets and increase clustering quality since no extra features that could represent noises are involved in unsupervised learning. Three methods of Unsupervised (FS)(Wrappers, Filter, and embedded) [27]. Fig. 2 showing the flowchart of this method.

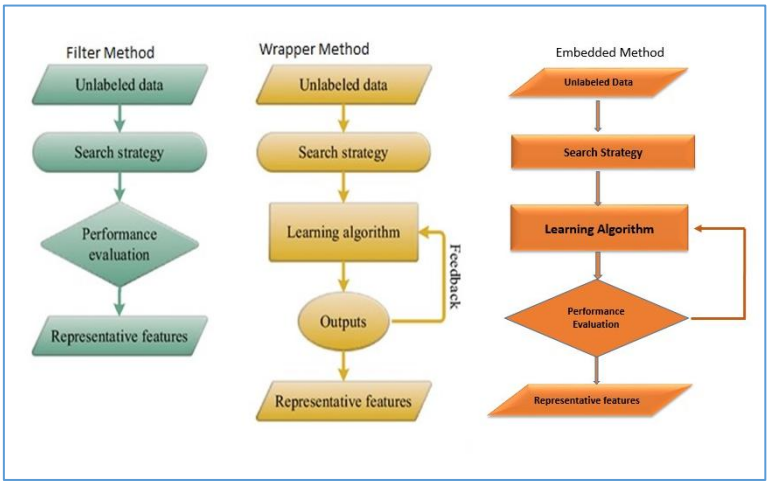

Fig. 2. Unsupervised Features selection approaches

In general, there are many applications in features selection for "dimensionality reduction" like "Missing Value Ratio, Low Variance Filter, High correlation Filter, Random Forest, Backward Feature extraction, Forward features selection, Recursive Feature Elimination (RFE)" [28].

4. Features Extraction

Features Extraction is an approach to separate a new set of features from the original dataset variable. Assuming there are n features X1, X2,….Xn we getting new sew set of features after extraction Y1, Y2,….Ym (m < n), Ri = Gi (S1; S2;…..; An), and Ri is a visualization function. Feature extraction aims to generate a new miniature set of features by some transmutation based on some performance metric [6]. When we want to use fewer resources for processing without losing important feature datasets, the feature extraction method is helpful because it removes the extra features from the original dataset. The number of extra features for a study can be minimized by feature extraction. First, feature extraction transforms features spectacularly to yield more important characteristics. Feature extraction creates new features that rely on the initial input feature set to lower the feature vector's high dimensionality. Algebraic transmutation is used for the transformation process. based on some optimization requirements [29, 30]. By maintaining the initial relative distance among components and accounting for the initial data potential structure, this type of "dimensionality reduction" algorithm tries to keep the most significant dataset information while the information process [31]. Compared to feature selection approaches, (FE) is less vulnerable to "overfitting" and performs well for classifying the data. But occasionally, after the transformation, the data description is lost, and this process is expensive for several datasets [32, 33]. Three matters should be considered in the features extraction mechanism [34].

4.1. Performance Evaluation

It investigates the best method for evaluating extracted features. For example, for a classification task, Predictive accuracy and class labels in the data can be utilized to detect a set of extracted variables. During the clustering step, we have to use metrics like "inter-cluster/intra-cluster" resemblance, Data Variation, etc. With data that do not have class labels.

4.2. Transformation

It refers to the mechanism of mapping original features to new features. Features can be extracted using a variety of mappings. Generally, mapping can be divided into "linear" transformations and "non-linear" transmutation, and according to the dimension's factor, Mapping could be divided into "linear and labelled, linear and non-labelled, non-linear and labelled, non-linear and non-labelled".

4.3. New created Features numbers

It is indicating to surveying approaches that detect the smallest different features. While the main target is generating the smallest group of new features, the question is what new feature numbers guarantee the remaining data is still in nature after transformation.

Features extraction uses some types of Dimensionality Reeducation depending on types of data, like "Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), Isometric Mapping (ISOMAP), Locally Linear Embedding (LLE), Linear Discriminant Analysis (LDA), Latent Semantic Indexing (LSI) and clustering methods", Fig. 3 show feature extraction steps [20].

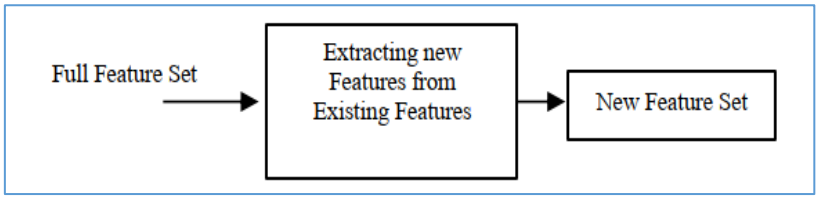

Fig. 3. Features the extraction process [20]

4.4. Feature Extraction Types

Feature Extraction is classified into two main methods (linear and non-linear), each consisting of many types. Choosing process depends on the type of data. New attributes that are generated don't have the same original data values [35].

4.4.1. Linear methods

Using linear methods, original data projection will be linear onto a low-dimensional area. Many techniques are used in this method. The most important are "Principal Component Analysis (PCA), Factor Analysis (FA), Linear discriminant analysis (LDA), and Truncated Singular Value Decomposition (SVD)" under linear methods. These techniques only work effectively with linear data; their activity is less with non-linear data [35].

4.4.2. Non-linear methods

It can be challenging to interpret multi-dimensional data which cannot be expressed in two or three dimensions. Assuming that the relevant data is in a space with fewer dimensions is one method to express it. The data can be seen in an area with fewer dimensions if the significance level of the data is low enough. The following linear methods are connected to some non-linear dimensionality reduction techniques. Non-linear techniques have two categories: first, generate a layout (either from the area with high-dimensional to lower-dimension embedding or vice versa) and merely represent the dimension. Methods that are most widely used in Non-linear "t-distributed Stochastic Neighbor Embedding (t-SNE), Kernel PCA, Multi-dimensional Scaling (MDS), Isometric mapping (Isomap)" [36].

5. proposed work

5.1. Remove missing values in the dataset

Before the reduction process, we should look at the data's nature. The dataset used in this experiment contains some missing values and overburden in the training, modelling, and classification, and we have to eliminate them. The missing Values Ratio (MVR) method considers one of the features selection Tanique to remove these values. This method finds the ratio of missing observations for each attribute, MVR applied on this dataset. The "Ethnicity Features" have been removed due has (76.6%) missed values.

5.2. Classify data

In this stage, we will apply different machine learning classifiers to determine the predictive accuracy of the original dataset. The Tables 1-5 show the prediction results of five classifiers, and Table 6 shows the performance evaluation of each classifier before the reduction process. We divided the dataset into (70%) training and (30%) also tests cross-validation (10).

Table 1. Confusion matrix of Random Forest classifiers

	Army National Guard	**Army Reserve**	**Regular Army**	$\sum$
Army National Guard	1574	0	287	1861
Army Reserve	47	0	54	101
Regular Army	104	0	2016	2120
$\sum$	1725	0	2357	4082

Table 2. Confusion matrix of KNN classifier

	Army National Guard	**Army Reserve**	**Regular Army**	$\sum$
Army National Guard	1705	10	146	1861
Army Reserve	54	15	32	101
Regular Army	114	16	1990	2120
$\sum$	1873	41	2168	4083

Table 3. Confusion matrix of Support Vector Machine

	Army National Guard	**Army Reserve**	**Regular Army**	$\sum$
Army National Guard	1503	0	358	1861
Army Reserve	63	0	38	101

Regular Army	607	0	1513	2120
∑	2173	0	1909	4082

Table 4. Confusion matrix of a Decision tree classifier

	Army National Guard	Army Reserve	Regular Army	∑
Army National Guard	1793	11	57	1861
Army Reserve	53	26	22	101
Regular Army	72	17	2031	2120
∑	1918	54	2110	4082

Table 5. Neural network classifier

	Army National Guard	Army Reserve	Regular Army	∑
Army National Guard	1632	26	203	1861
Army Reserve	59	15	27	101
Regular Army	151	15	1954	2120
∑	1843	56	2184	4082

Table 6. Show the classifier's performance evaluation before the reduction process

Classifier	AUC	CA	F_1	precision	recall	specificity	training set size
Decision tree	77%	77%	77%	77%	77%	83%	70%
SVM	65%	64%	62%	67%	64%	70%	70%
Random Forest	81%	73%	71%	71%	73%	75%	70%
Neural Network	50%	46%	44%	45%	46%	54%	70%
KNN	92%	83%	82%	82%	83%	86%	70%

AUC: The area under ROC is the area under the receiver-operating curve.
CA: Classification accuracy is the proportion of correctly classified
F1: is a weighted harmonic mean of precision and recall
Precision is the proportion of true positives among instances classified as positive.
The recall is the proportion of true positives among all positive instances in the data.
Specificity is the proportion of true negatives among all negative instances.

5.3. Dimensions reduction stage

In this stage, we used different Dimensionality reduction techniques to decrease features in the dataset, and we mentioned the dataset previously, "ANSUR 2", which have (109 features and 6068 observation). We will find the effectiveness and accuracy of each method on this dataset.

5.3.1. Highly Correlated Filter

It is a 'feature selection technique' based on detecting the most correlated features and removing them from the dataset, and the result is clarified in the Table 2. The prediction accuracy (65.2%) when features correlation is (0.88) As shown in Table 2.

5.3.2. Recursive Feature Elimination (RFE)

It is a feature selection method. This technique mechanism is based on removing the weak features in the dataset and frequently removing many features in each loop. Fig. 4 show the accuracy of prediction for this method.

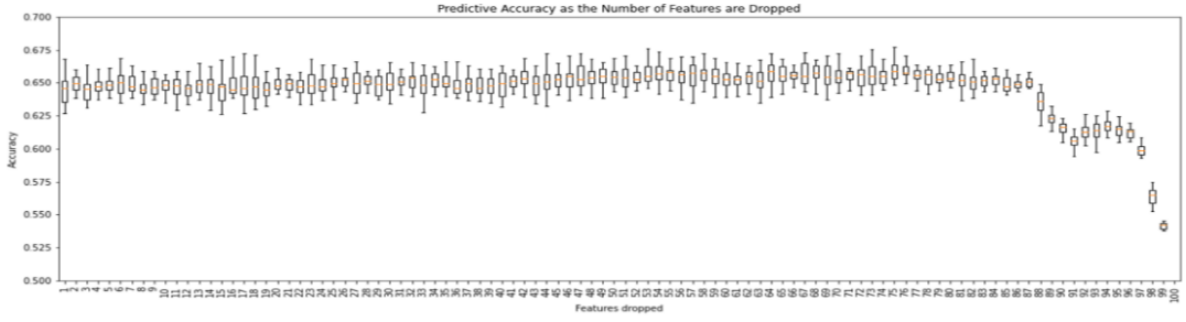

Fig. 4. REF

5.3.3. Principal Component Analysis (PCA)

It is A widely known technique, and it is feature extraction, used to reduce the dimensions of a huge dataset by converting a large dataset feature to a smaller one with keeping the most information of original data. Fig. 5 accuracy of prediction with PCA technique.

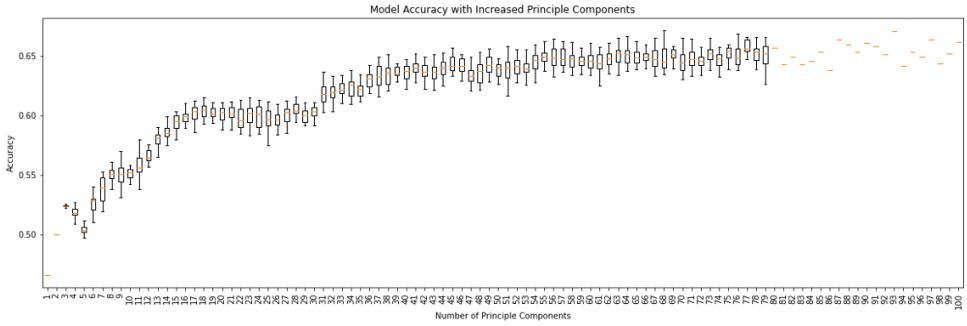

Fig. 5. REF

6. Result and Discussion

It made a practical experiment on ANSUR 2 datasets used in this work which has (109 attributes and 6068 observations). We used five classifiers for classification (Decision Tree, SVM, Random Forest, Neural Network, and KNN) with cross-validation (10) and dataset separated in (70%) training and (30%) test. We used Python programing language. We measured classification accuracy as a performance evaluation. It is calculated by dividing the correct prediction number by the total of prediction according to this equation as in equation (1)

$$\text{Accuracy}= \frac{\text{Number of correct predictions}}{\text{Total Number of prediction}}\text{s...} \quad (1)$$

Table 7 illustrates the Classifiers' results with the Highly Correlated Filter technique after reducing (17) attributes from the data set, Table 8 shows the Classifiers' results with the Recursive Feature Elimination technique after reducing (75) attributes, and Table 9 shows classifiers result after reducing data by PCA. Table 10 shows the result of dimensionality reduction techniques used in this experiment, and we have applied all techniques (Features selection and feature extraction). Prediction accuracy is one of the most important criteria for evaluating the work of algorithms, and the results show the prediction accuracy of each method. 'Missing Value Ratio' was used to remove (8) features of missing values as a pre-processing. Due to these features being dropped from the dataset, the REF Algorithm achieved a prediction score highly with the KNN classifier and an Improvement percentage (1.7%).

Table 7. Classifiers results with Highly Correlated Filter technique

classifier	AUC	CA	F_1	precision	recall	specificity	training set size
Decision tree	78%	77%	77%	77%	77%	82%	70%
SVM	68%	60%	58%	57%	60%	64%	70%
Random Forest	75%	68%	66%	65%	68%	70%	70%
Neural Network	77%	67%	66%	65%	67%	72%	70%
KNN	92%	83%	82%	82%	83%	86%	70%

Table 8. Classifiers results with Recursive Feature Elimination technique

classifier	AUC	CA	F_1	precision	recall	specificity	training set size
Decision tree	81%	80%	80%	80%	80%	85%	70%
SVM	61%	55%	53%	51%	55%	59%	70%
Random Forest	83%	76%	74%	76%	76%	77%	70%
Neural Network	73%	64%	63%	62%	64%	69%	70%
KNN	92%	83%	84%	83%	86%	86%	70%

Table 9. Classifiers results with Principal Component Analysis technique

classifier	AUC	CA	F_1	precision	recall	specificity	training set size
Decision tree	78%	76%	76%	76%	76%	82%	70%
SVM	77%	65%	63%	62%	65%	70%	70%

Random Forest	85%	77%	74%	75%	77%	78%	70%
Neural Network	88%	77%	76%	75%	77%	80%	70%
KNN	92%	83%	82%	82%	83%	86%	70%

Table 10. Show the dimensionality reduction techniques' performance

Technique	Reduction type	Accuracy	Note	Improvement percentage
Missing Value Ratio	Features selection	26.0% missing value percentage	8 features are dropped (deleted)	
Highly Correlated Filter	Features selection	65.2%	17 features are dropped (deleted)	0.9 %
Recursive Feature Elimination (RFE)	Features selection	66%	75 features are dropped (deleted)	1.7%
PCA (principal components Analysis)	Features Extraction	65.03%	55 features are dropped (deleted)	0.75%

7. Conclusion

Nowadays, due to a large number of dealing with large data, there has become an urgent need for dimension-reduction techniques to facilitate dealing with large data. Reducing dimensions may lead to losing a large part of the information. For this reason, several standards must be followed, such as preserving the original data infrastructure and representing the reduced data by a high percentage of the original data. Therefore, we have to use a suitable technique for reduction. In this article, we mentioned Dimensionality Reduction and its role in dealing with big data (which contains many numbers of attributes). DR consists of two main methods, 'Features selection & Features extraction'. Both have the main task, which leads to reduced dataset dimensions. Each technique has different algorithms that are Probably applied to different datasets. ANSUR 2 datasets used in this work which has (109 attributes and 6068 observations). This dataset is available at [6] and was published in 2012 in USA 2017 became available on the website and consists of body measurements. A practical experiment implemented throw three stages; (1) classifying dataset., (2) Reducing Dimensions, and (3) Classifying Reduced data. Recursive Feature Elimination (RFE) is the most accurate with KNN Classifier, as shown in Table 3.

Before reducing the dimensions, we note that the (KNN) classifier showed a high prediction rate of (83%) and after the reducing stage, it showed the same result; therefore, KNN classifier is the most suitable one for this dataset.

RFE technique achieved (66%) accuracy percentage and dropped 55 features from the dataset, and (RFE with KNN) achieved what is required. The results in a Table 8 show that.

Acknowledgment

The authors would like to thank the Middle Technical University for the financial support of this project.

References

[1] M. Al-Ayyoub, Y. Jararweh, A. Rabab'ah, and M. Aldwairi, "Feature extraction and selection for Arabic tweets authorship authentication," J. Ambient Intell. Humaniz. Comput., vol. 8, no. 3, pp. 383–393, 2017, doi: 10.1007/s12652-017-0452-1.
[2] C. A. Buckner et al., "We are IntechOpen, the world ' s leading publisher of Open Access books Built by scientists, for scientists TOP 1 %," Intech, vol. 11, no. tourism, p. 13, 2016, [Online]. Available: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics
[3] M. Ziaye, S. Khalid, and Y. Mehmood, "Survey of Feature Selection/Extraction Methods used in Biomedical Imaging," Int. J. Comput. Sci. Inf. Secur., vol. 16, no. 5, pp. 169–177, 2018.
[4] "ANSUR II | The OPEN Design Lab." https://www.openlab.psu.edu/ansur2/ (accessed Sep. 26, 2022).
[5] Z. Chen et al., "IFeature: A Python package and web server for features extraction and selection from protein and peptide sequences," Bioinformatics, vol. 34, no. 14, pp. 2499–2502, 2018, doi: 10.1093/bioinformatics/bty140.
[6] H. Motoda and H. Liu, "Feature selection, extraction, and construction," Commun. IICM, vol. 5, pp. 67–72, 2002.
[7] J. Li et al., "Feature selection: A data perspective," ACM Comput. Surv., vol. 50, no. 6, 2017, doi: 10.1145/3136625.
[8] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," Proc. 2014 Sci. Inf. Conf. SAI 2014, pp. 372–378, 2014, doi: 10.1109/SAI.2014.6918213.
[9] I. Journal and I. Factor, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," Comput. Math. Methods Med., vol. 2015, no. 1, pp. 2–4, 2015, [Online]. Available: http://dx.doi.org/10.1155/2015/
[10] R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," 1st Int. Conf. Emerg. Trends Eng. Technol. Sci. ICETETS 2016 - Proc., 2016, doi: 10.1109/ICETETS.2016.7603000.
[11] P. M. M. Manohara, G. Attigeri, and R. M. Pai, "Analysis of feature selection and extraction algorithm for loan data: A big data approach," 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017, vol. 2017-Janua, pp. 2147–2151, 2017, doi: 10.1109/ICACCI.2017.8126163.
[12] A. A. Raweh, M. Nassef, and A. Badr, "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation," IEEE Access, vol. 6, pp. 15212–15223, 2018, doi: 10.1109/ACCESS.2018.2812734.
[13] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating Feature Selection and Feature Extraction Methods with Deep Learning to Predict Clinical Outcome of Breast Cancer," IEEE Access, vol. 6, pp. 28936–28944, 2018, doi: 10.1109/ACCESS.2018.2837654.
[14] I. Perova and Y. Bodyanskiy, "Adaptive human machine interaction approach for feature selection-extraction task in medical data mining,"

Int. J. Comput., vol. 17, no. 2, pp. 113–119, 2018, doi: 10.47839/ijc.17.2.997.
[15] A. Phinyomark, R. N. Khushaba, and E. Scheme, "Feature extraction and selection for myoelectric control based on wearable EMG sensors," Sensors (Switzerland), vol. 18, no. 5, pp. 1–17, 2018, doi: 10.3390/s18051615.
[16] B. He, S. Shah, C. Maung, G. Arnold, G. Wan, and H. Schweitzer, "Heuristic search algorithm for dimensionality reduction optimally combining feature selection and feature extraction," 33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019, pp. 2280–2287, 2019, doi: 10.1609/aaai.v33i01.33012280.
[17] X. Li, S. H. Ling, and S. Su, "A hybrid feature selection and extraction methods for sleep apnea detection using bio-signals," Sensors (Switzerland), vol. 20, no. 15, pp. 1–14, 2020, doi: 10.3390/s20154323.
[18] S. Ye, J. Jiang, Z. Zhou, C. Liu, and Y. Liu, "A Fast and Intelligent Open-Circuit Fault Diagnosis Method for a Five-Level NNPP Converter Based on an Improved Feature Extraction and Selection Model," IEEE Access, vol. 8, pp. 52852–52862, 2020, doi: 10.1109/ACCESS.2020.2981247.
[19] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang, and H. Wan, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," Expert Syst. Appl., vol. 150, p. 113277, 2020, doi: 10.1016/j.eswa.2020.113277.
[20] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 56–70, 2020, doi: 10.38094/jastt1224.
[21] S. Wang et al., "Research and Experiment of Radar Signal Support Vector Clustering Sorting Based on Feature Extraction and Feature Selection," IEEE Access, vol. 8, pp. 93322–93334, 2020, doi: 10.1109/ACCESS.2020.2993270.
[22] T. L. Kei Suzuki, "Constructing an Emotion Estimation Model Based on EEG/HRV Indexes Using Feature Extraction and Feature Selection Algorithms," 2021.
[23] Priyanka and D. Kumar, "Feature Extraction and Selection of kidney Ultrasound Images Using GLCM and PCA," Procedia Comput. Sci., vol. 167, no. 2019, pp. 1722–1731, 2020, doi: 10.1016/j.procs.2020.03.382.
[24] U. R. Aparna and S. Paul, "Feature selection and extraction in data mining," Proc. 2016 Online Int. Conf. Green Eng. Technol. IC-GET 2016, 2017, doi: 10.1109/GET.2016.7916845.
[25] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection : Introduction and review," J. Biomed. Inform., vol. 85, no. July, pp. 189–203, 2018, doi: 10.1016/j.jbi.2018.07.014.
[26] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70–79, 2018, doi: 10.1016/j.neucom.2017.11.077.
[27] S. Solorio-Fernández, J. Ariel Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A systematic evaluation of filter Unsupervised Feature Selection methods," Expert Syst. Appl., vol. 162, 2020, doi: 10.1016/j.eswa.2020.113745.
[28] Z. Liu, R. Wang, N. Japkowicz, Y. Cai, D. Tang, and X. Cai, "Mobile app traffic flow feature extraction and selection for improving classification robustness," J. Netw. Comput. Appl., vol. 125, pp. 190–208, 2019, doi: 10.1016/j.jnca.2018.10.018.
[29] M. K. Elhadad, K. M. Badran, and G. I. Salama, "A novel approach for ontology-based dimensionality reduction for web text document classification," Int. J. Softw. Innov., vol. 5, no. 4, pp. 44–58, 2017.
[30] D. A. Zebari, H. Haron, S. R. M. Zeebaree, and D. Q. Zeebaree, "Enhance the Mammogram Images for Both Segmentation and Feature Extraction Using Wavelet Transform," 2019 Int. Conf. Adv. Sci. Eng. ICOASE 2019, pp. 100–105, 2019, doi: 10.1109/ICOASE.2019.8723779.
[31] N. Abd-Alsabour, "On the Role of Dimensionality Reduction," J. Comput., vol. 13, no. 5, pp. 571–579, 2018, doi: 10.17706/jcp.13.5.571-579.
[32] R. Aziz, C. K. Verma, and N. Srivastava, "Dimension reduction methods for microarray data: a review," AIMS Bioeng., vol. 4, no. 2, pp. 179–197, 2017.
[33] A. S. Eesa, A. M. Abdulazeez, and Z. Orman, "A DIDS Based on The Combination of Cuttlefish Algorithm and Decision Tree," Sci. J. Univ. Zakho, vol. 5, no. 4, pp. 313–318, 2017.
[34] B. Ghojogh et al., "Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review," 2019, [Online]. Available: http://arxiv.org/abs/1905.02845
[35] I. De-La-bandera, D. Palacios, J. Mendoza, and R. Barco, "Feature extraction for dimensionality reduction in cellular networks performance analysis," Sensors (Switzerland), vol. 20, no. 23, pp. 1–10, 2020, doi: 10.3390/s20236944.
[36] K. Gajamannage, R. Paffenroth, and E. M. Bollt, "A non-linear dimensionality reduction framework using smooth geodesics," Pattern Recognit., vol. 87, no. Xx, pp. 226–236, 2019, doi: 10.1016/j.patcog.2018.10.020.