Energy Efficient Medical Data Dimensionality Reduction using Optimized Principal Component Analysis

S. Gnana Sophia[1,*], K.K. Thanammal[2] and S.S. Sujatha[2]

[1]Research Scholar, Department of Computer Science and, S.T. Hindu College, Nagercoil, MS University, Abishakapatti, Tirunelveli-627012, Tamilnadu, India.
[2]Associate Professor, Department of Computer Science and Applications, S.T. Hindu College, Nagercoil, MS University, Abishakapatti, Tirunelveli-627012, Tamilnadu, India.

Abstract

INTRODUCTION: The method of minimizing the number of random variables or attributes from the enormous data set is the reduction of dimensionality. The space available for storing the database is therefore minimized by decreasing the scale of the features.
OBJECTIVES: The PCA algorithm is used to achieve dimensional reduction by deep learning to recover image characteristics. This approach is designed to reduce the dimensionality of such datasets, improve interpretability while minimizing the loss of information
METHODS: The dimensionality reduction of the method by using optimized PCA algorithm. The input data set can be reducing the dimension by using PCA algorithm. The tree seed optimization algorithm (TSO) can be utilized to select the optimal data's in PCA algorithms. After completing the TSO-PCA the new data set are created by the reduced dimensions.
RESULTS: The input data and images are used to reduce the dimension based on the TSO-PCA algorithms. The simulations for obtaining the results were carried out using python. The results of the feature dimensionality reduction on DIABETES dataset and Indian pines dataset.
CONCLUSION: The best data for the data collection, the TSO algorithm is used and the PCA algorithm is used to minimize the dimensions. The suggested method is better than the existing method compared to the linear, kernel, random basic function, and polynomial for evaluating the outcome and discussion. In order to improve accuracy in future work, we will continue research and try to find more advanced techniques for this problem.

Keywords: Principal Component Analysis and Tree Seed Optimization algorithm.

Received on 07 July 2021, accepted on 03 August 2021, published on 06 August 2021

doi: 10.4108/eai.6-8-2021.170667

1. Introduction

The method of minimizing the number of random variables or attributes from the enormous data set is the reduction of dimensionality. The space available for storing the database is therefore minimized by decreasing the scale of the features. This paper includes the study of dimension reduction in order to address these concerns. The PCA algorithm is used to achieve dimensional reduction by deep learning to recover image characteristics. This approach is designed to reduce the dimensionality of such datasets, improve interpretability while minimizing the loss of information [1]. In order to uncover information that leads to superior business policy, including latent patterns, unknown correlations, and other analyses, big data analytics analyses broad and

[*]Corresponding author. Email: gnanasphiajournals@gmail.com

different sets of data from different sources [2]. In the future, the usefulness of these approaches for reducing dimensionality will be preserved on data of high dimensionalities, such as images, experiments, etc.

Axes of PCA materials that increase the variance and decrease the dimension finding effective directions [3]. It is an important data relationship that transforms current data on the basis of these relationships and then quantifies the value of these relationships so that we can preserve the most important relationships [4]. The new features that PCA creates are orthogonal, meaning they are interrelated. The transformation also depends on the scale, the dataset normalization and the linear relationship between the characteristics [5]. By projecting them into a subspace which collects most of the variance [6], it is a specialized approach for decreasing the high dimensionality of big data. When the data variables are collinear with each other, PCA produces the best results [7]. The image extraction function is to remove the function language from the image, which is useful in solving the issues [8]. Trends, patterns, and correlations are simple to see by measuring the data returned to simulation, and the method's performance is enhanced [9]. It introduces requirements on an unknown differential operator that would cause one (or more) new operators operating on new spaces to define it as having a low dimension [10]. Large-scale data occurred well before the Big Data era, particularly in bioinformatics, and many techniques were adequately used to analyze both small and large-scale data. [11].The remaining paper contains, section 2 provides the related works, section 3 explains the overall proposed method and also the brief explanation of optimization algorithms. Section 4 provides the result part and section 5 explains the conclusion part.

2. Literature Survey

A lot of researches are analyzed the dimensionality reduction of the data. the various existing methods are given below, In 2015 Swati. A and Ade. R [12] has proposed the PCA goal of looking for a hyperspace whose base vectors are parallel to the positions of the highest variance. And yet, the more descriptive characteristics can be scanned by PCA, it ignores the useful class mark information and was thus not suitable for normal classification tasks. Although the groups inside the training data set were not taken into consideration, the PCA technique was an unsupervised technique. While it was best for restoration, it was inappropriately ideal from the point of view of discrimination.

In 2012 T.L. Grobler [13] has been provided by that can be used to compare the performance of various reduction methods, whether or not the initial dataset has been labeled. They use this and a case study to compare two approaches for minimizing hyper temporal dimensionality, including PCA, to explain how this framework works.

K., Lakshmanna, *et al*, [14] evaluated the presentation of PCA and LDA on multi-dimensional datasets with multiple ML algorithms in 2020, and the experiment was replicated on two other datasets, namely Diabetic Retinopathy (DR). Although high precision is obtained by the extraction of vessels before detecting DR with machine learning, the processing of the marked ground-truth for retinal vessels was time-consuming. Because network-based IDSs (NIDS) were unable to detect encrypted node communication, standard IDSs were not sufficient for the cloud context, and host-based IDSs also did not recognize the secret attacker path (HIDS).

In 2016 Wu, Z. *et al*, [15] has discussed to implement a new parallel and distributed architecture for cloud computing-based massive hyper-spectral image processing. In general, as a case study, they utilize dimensionality elimination to illustrate the suitability and feasibility of the use of cloud computing technology to successfully conduct centralized parallel hyper-spectral data analysis and accelerate hyper-spectral data computing.

In 2017 Fanwu Chu. [16] have proposed to identify irregularities in the operational procedures of the HUs, it was important to differentiate between the various operational conditions and to create the PCA model within various operating conditions. The Recursive PCA (RPCA) and Moving Window PCA (MWPCA) are two powerful adaptive updating approaches that are modified to track the operating phase of HUs depending on various operating environment, respectively.

In 2018, Salo, *et al*, [17] suggested a hybrid approach that incorporates IG and PCA to discard inappropriate features and preserve an optimum subset of parameters whereas a classified system was developed using the SVM, IBK, and MLP-based community classifier. Recent studies have shown in this sense that a successful method to feature selection was a major element in supporting the classification system in the feature extraction.

In 2017, Lazcano, R., *et al*, [18] suggested a study of the inherent parallel processing of the various phases of the PCA methodology in order to exploit the opportunities of parallel processing provided by several key configurations of the MPPA. In addition, the effect of expanding the degree of parallel processing on internal contact was also examined.

3. Proposed methodology

In this section, describes the dimensionality reduction of the method by using optimized PCA algorithm. Here the input data set can be reducing the dimension by using PCA algorithm. The tree seed optimization algorithm (TSO) can be utilized to select the optimal data's in PCA algorithms. After completing the TSO-PCA the new data set are created by the reduced dimensions. The overall diagram of the proposed approach is shown below,

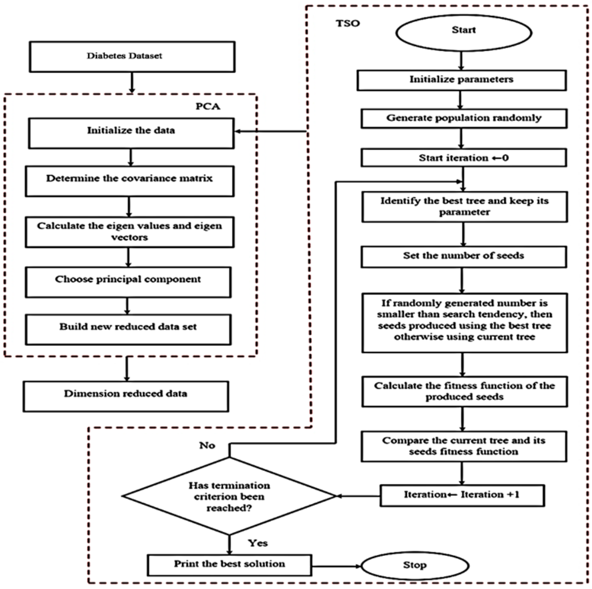

TSO
Start
Initialize parameters
Generate population randomly
Start iteration ←0
Identify the best tree and keep its parameter
Set the number of seeds
If randomly generated number is smaller than search tendency, then seeds produced using the best tree otherwise using current tree
Calculate the fitness function of the produced seeds
Compare the current tree and its seeds fitness function
Iteration← Iteration +1
Diabetes Dataset
PCA
Initialize the data
Determine the covariance matrix
Calculate the eigen values and eigen vectors
Choose principal component
Build new reduced data set
Dimension reduced data
No
Has termination criterion been reached?
Yes
Print the best solution
Stop

Figure 1. Overall diagram of the proposed method

3.1. Principle Component Analysis (PCA) algorithm

The trees can be generated by using the product in the simple TSO algorithm by the different trees, which leads to a decrease in redundancy that decreases population diversity. More specifically, there is a low risk for trees with poor health to achieve the maximal solution. To solve this problem, key component analysis is incorporated in this work. A significant approach to statistical analysis is the main component analysis, which has two main purposes, data reduction, and interpretation. Using PCA, minimally correlated variables perhaps distinguished from reliant variables. The original input solutions are minimally associated with the PCA algorithm and display the original solution detail. In comparison, the new solutions can have higher consistency than the initial solutions and the TSA algorithm is used to find the best solution. Let U is an example of the original knowledge, which can be summarized as follows,

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pn} \end{pmatrix} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{pmatrix} \tag{1}$$

Where *p* represents the number of exemplars.

3.2 Tree seed optimization algorithm (TSO)

The TSA mechanism is inspired by the interaction between trees and their seeds. The possible solutions to optimization problems are represented by trees and seeds, so they scan the possible environments with iterations in parallel, and eventually get the optimum solution. As follows, further specifics are introduced. The initial positions of trees are first uniformly developed in the problem-solution space as follows,

$$i = 1,2,\cdots N \; and \; j = 1,2,\cdots D \tag{2}$$

Where, N represents the number of trees in the population, and D represents the size of the problem. By using the fitness function, objective values of trees are computed after producing trees in the population. One seed is generated by using equations 2 and 3 for each tree until the stopping requirements are met,

$$S_{k,j} = T_{i,j} + \left(B_j - T_{r,j}\right) \times (rand - 0.5) \tag{3}$$

$$S_{k,j} = T_{i,j} + \left(T_{i,j} - T_{r,j}\right) \times (rand - 0.5) \tag{4}$$

Where $S_{k,j}$ represents the j^{th} size of the output of the k^{th} seed for $T_{i,j}$. The j^{th} size of the randomly chosen r^{th} tree is denoted by $T_{r,j}$. In addition, equations 3 and 4 are focused on the optimal threshold search tendency (ST) and take a value at the interval of [0, 1]. If the generated value is less than the ST parameter at random, equation 3 is used; if not, equation 4 is used. The ST variable is a significant variable for seed manufacturing selection. Therefore, the number of seeds depending on the population. The average number of trees in the number of trees is 25% of the number of trees in the number of trees. Initially, the TSO counts the locations of the first trees to the possible solutions to improve problems according to equation 5.

$$T_{i,j} = L_{j,min} + r_{i,j}\left(H_{j,max} - L_{j,min}\right) \tag{5}$$

The lower and upper limits of the L_(j,min) and H_(j,max)search stages, respectively. r_(i,j)describes the randomized generation of the location and each component in the [0, 1] interval. A selection of population and the best solution is given below,

$$B = min\{f\overline{T_i}\} \; i = 1,2,\cdots,N \tag{6}$$

For each tree, the number of trees in the population is expressed as N, and new seed positions are established. The number of seeds generated must be greater than one, focusing on the population level. The optimal solutions are sought using TSO, and linear transformations can be used to quantify the data as follows:

$$\begin{cases} Z_1 = a_1'U = a_{11}U_1 + a_{12}U_2 + \cdots a_{1p}U_p, \\ Z_2 = a_2'U = a_{21}U_1 + a_{22}U_2 + \cdots a_{2p}U_p, \\ \cdots \\ Z_p = a_p' = a_{p1}U_1 + a_{p2}U_2 + \cdots a_{pp}U_p. \end{cases} \tag{7}$$

Where a_idescribes the coefficient vector, it is convenient to describe the linear regression system as described:

$$Var(Z_i) = a_i' \textstyle\sum a_i, \quad i = 1,2,\cdots p \tag{8}$$

$$Cov(Z_i,Z_j) = a_i' \textstyle\sum a_j, \quad i,j = 1,2,\cdots p \tag{9}$$

It is important to choose the number m of principal components. The m is defined as follows,

$$\frac{\lambda_1 + \lambda_2 + \cdots \lambda_m}{\sum_{i=1}^{s} \lambda_i} \geq \delta \tag{10}$$

Where δ represents the rate of contribution. The proper values of the covariance matrix are $\lambda_1, \lambda_2$and λ_m. TSO $X = \{x_1^t, x_2^t, \cdots x_n^t\}$ a covariance matrix is V, the main population can be generated as follows,

$$\begin{cases} F_1^t = a_1'X = a_{11}x_1^t + a_{12}x_2^t + \cdots a_{1n}x_n^t, \\ F_2^t = a_2^tX = a_{21}x_1^t + a_{22}x_2^t + \cdots a_{2n}x_n^t, \\ \cdots \\ F_m^t = a_m'X = a_{m1}x_1^t + a_{m2}x_2^t + \cdots a_{mn}x_n^t \end{cases} \tag{11}$$

Where $(a_1', a_2', \cdots, a_m')$denotes V feature vectors. The newly generated population is uncorrelated according to the main component analysis theory. The individuals with bad fitness are replaced by the main population operation.

Steps involved to calculate eigenvalues and eigenvectors: To calculate eigenvalues, there are 3 simple features: First,

- To calculate the eigenvalues/eigenvectors of a square matrix (n x n, the matrix's coverage).
- To calculate the n-space of n vectors present in them.
- Calculating the distance of eigenvectors with subsequent eigenvalue representing the vector's power.

This makes sense because the knowledge they serve constitutes both of them. Because new features decrease the dimensionality of Diabetes data and Indian pine data, data covariance matrix eigenvectors are measured to find patterns (eigenvectors) with their importance (eigenvalues). The covariance matrix, eigenvectors will represent new features and pick some of them according to their influence or effect on their value.

After calculating the eigen values, it can be arranged highest to lowest eigen values. This arrangement the highest eigen values are taken as to build the new features and reduce the dimension of the data set. The new data set can be given as,

$$newDataset^T = featureVector^T.Dataset^T \tag{12}$$

Where T considered as the transpose of the data set and finally the dimension reduce data can be given.

4. Result and discussion

This section describes the technique proposed to reduce the data dimension. Here the input data and images are used to reduce the dimension based on the TSO-PCA algorithms. The simulations for obtaining the results were carried out using python. The results of the feature dimensionality reduction on DIABETES dataset and Indian pines dataset.

4.1 Diabetes Dataset

Many important and descriptive databases such as the Diabetes Dataset are included in databases, and its first four dimensions are a string that describes the number of numbers and its last dimension flower species. There are nine columns in the Diabetes Dataset, such as Pregnancies, Glucose, blood pressure, skin thickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome.

Table 1. Diabetes Data Set

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1

1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

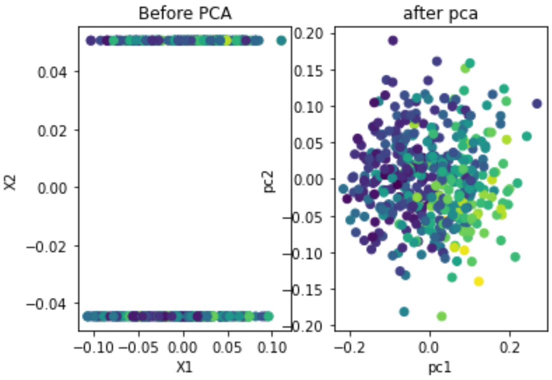


Figure 2. Representation of before and after TSO-PCA performance for DIABETES dataset

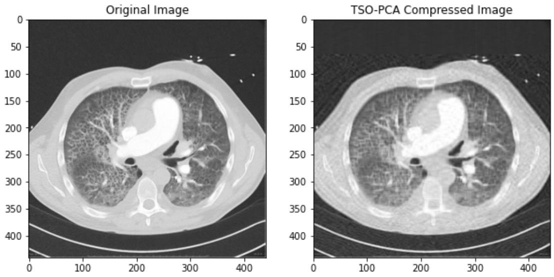


Figure 3. Original Image and TSO-PCA Compressed Image

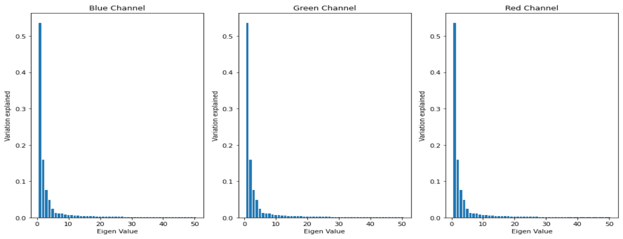
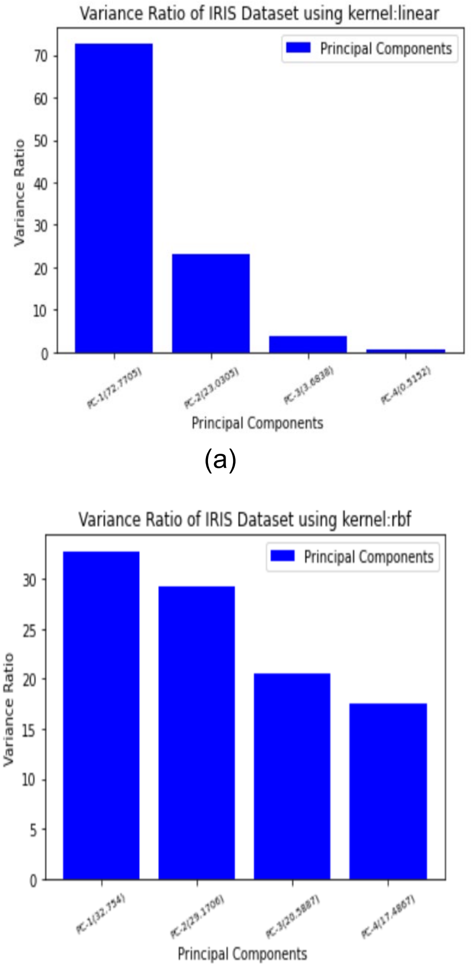


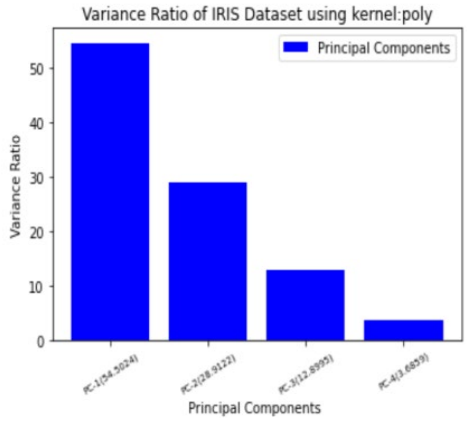
Figure 4. Let's plot bar charts to check the explained variance ratio by each eigenvalue separately for each of the 3 channels.

Figure (4) represents the performance of DIABETES data set based on the TSO-PCA algorithms. By projecting each data point on only the first few key components to obtain lower-dimensional data while retaining as much of the variance of the data as possible, PCA is widely used for dimension reduction. By using TSO-PCA, the dimension can be reduced by high to low and also arranged. Figure (4) represents the performance based on the TSO-PCA algorithms. It is utilized to reduce the high dimensional images in to the low dimension. The graphical representation of variance ratio of the existing method is given below,



(a)

(b)

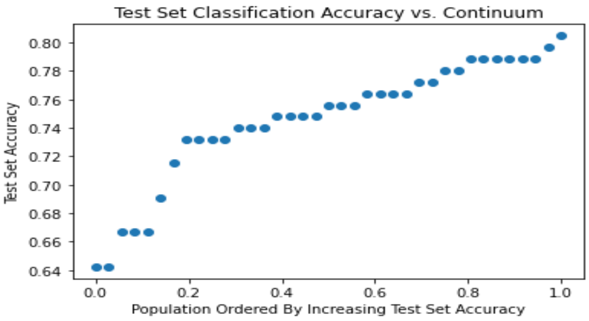


(c)

Figure 5. Variance ratio of the existing method (a) kernel: linear (b) KPCA on DIABETES dataset using kernel: linear (c) kernel: RBF.

The figure (5) represents the variance ratio of the existing method based on the linear, kernel, and random basis function. In figure (a) explained the variance ratio of the principal components using kernel PCA with Linear kernel and the result for 4 Principal Components according to their variance ratio. Since the initial two principal components have high variance. So, we selected the first two principal components. In figure (b) the KPCA on DIABETES dataset using kernel: linear shows that the two-principal component of the linear kernel.

In figure (c) provides the variance ratio of the principal components using kernel PCA with RBF kernel and the result is shown in the bar graph for the 4 Principal Components according to their variance ratio. Since the initial two principal components.



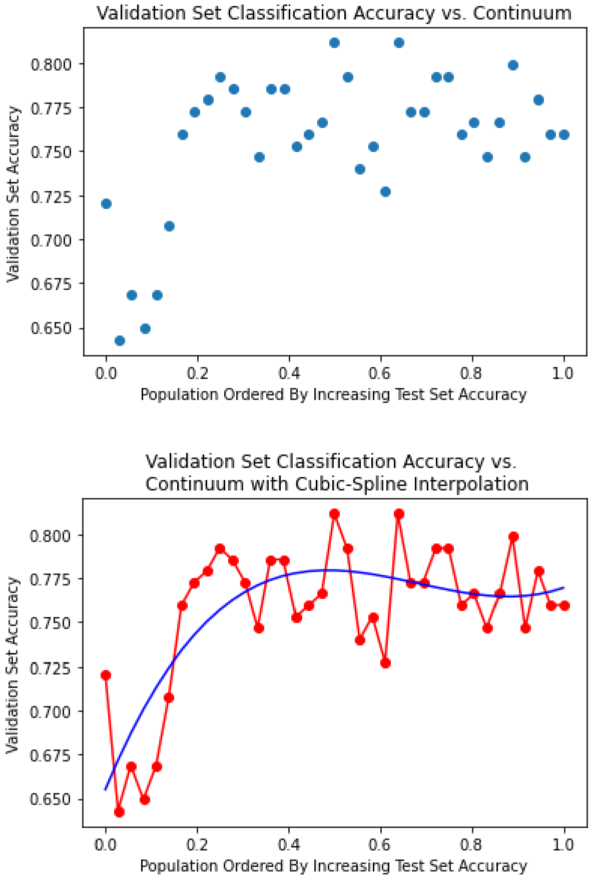


Figure 6. Validation set Classification Accuracy.

5. Conclusion

The purpose of the analysis was to analyze the integration of PCA dimensionality reduction using an optimized PCA algorithm for DIABETES and Indian pines data collection. The PCA method is capable of retaining valuable details in the DIABETES dataset and Indian pines dataset, as exemplified from the findings collected, while effectively reducing the measurements of the features in the used dataset, as well as providing a realistic data visualization model. Here, to choose the best data for the data collection, the TSO algorithm is used and the PCA algorithm is used to minimize the dimensions. The suggested method is better than the existing method compared to the linear, kernel, random basic function, and polynomial for evaluating the outcome and discussion. In order to improve accuracy in future work, we will continue research and try to find more advanced techniques for this problem.

References

[1] Abdulhammed R. Faezipour M. Musafer H. & Abuzneid A. Efficient Network Intrusion Detection Using PCA-Based Dimensionality Reduction of Features. In: 2019

International Symposium on Networks, Computers and Communications. 18-20 June 2019; Istanbul, Turkey.

[2] Acharjya and Ahmed, Acharjya D, Ahmed KP, A survey on big data analytics: challenges, open research issues and tools Int. J. Adv. Computer. Sci. Appl. 2016; 7 (2016), pp. 511-518.

[3] De Feis I, Dimensionality Reduction. Reference Module in Life Sciences. 2018;

[4] Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput, DS, Srivastava G, & Baker T, Analysis of Dimensionality Reduction Techniques on Big Data. IEEE Access, 2020; 8, 54776–54788.

[5] Du TY, Dimensionality Reduction Techniques for Visualizing Morphometric Data: Comparing Principal Component Analysis to Nonlinear Methods. Evolutionary Biology. 2018;

[6] Li M, Wang H, Yang L, Liang Y, Shang Z, & Wan H, Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction. Expert Systems with Applications. 2020; 113277.

[7] Oo MCM, & Thein T, An efficient predictive analytics system for high dimensional big data. Journal of King Saud University - Computer and Information Sciences. 2019;

[8] Ma J, & Yuan Y, Dimension Reduction of Image Deep Feature using PCA. Journal of Visual Communication and Image Representation. 2019; 102578.

[9] Genender-Feltheimer A, Visualizing High Dimensional and Big Data. Procedia Computer Science. 2018; 140, 112–121.

[10] Vieu P, On dimension reduction models for functional data. Statistics & Probability Letters. 2018; 136, 134–138.

[11] De Feis I, Dimensionality Reduction. Reference Module in Life Sciences.2018;

[12] Swati A, Ade R, Dimensionality reduction an effective technique for feature selection. Int. J. Computer. Appln. 117(3), 18–23 (2015).

[13] Grobler TL, Sequential and non-sequential hyper temporal classification and change detection of MODIS timeseries, Ph.D. thesis, University of Pretoria, 2012.

[14] Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput, DS, Srivastava G, & Baker T, Analysis of Dimensionality Reduction Techniques on Big Data. IEEE Access. 2020; 8, 54776–54788.

[15] Wu Z, Li Y, Plaza A, Li J, Xiao F, & Wei Z, Parallel and Distributed Dimensionality Reduction of Hyper spectral Data on Cloud Computing Architectures. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(6), 2270–2278.

[16] Fanwu Chu. (2017). An improved PCA algorithm for anomaly detection of hydropower units. 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA).

[17] Salo, Fadi; Nassif, Ali Bou, Essex, Aleksander. Dimensionality Reduction with IG-PCA and Ensemble Classifier for Network Intrusion Detection. Computer Networks, S1389128618303037

[18] Lazcano R, Madroñal D, Salvador R, Desnos K, Pelcat M, Guerra R, Sanz C, Porting a PCA-based hyperspectral image dimensionality reduction algorithm for brain cancer detection on a many core architecture. Journal of Systems Architecture. 2027; (77) 101–111.

[19] Kumar SN, Ahilan A, Fred AL, and Kumar HA, ROI extraction in CT lung images of COVID-19 using Fast Fuzzy C means clustering. In Biomedical Engineering Tools for Management for Patients with COVID-19. 2021; (pp. 103-119).

[20] Kumar SN, Ahilan A, Haridhas AK, and Sebastian J, Gaussian Hermite polynomial based lossless medical image compression. Multimedia Systems. 2021; 27(1): pp.15-31.

[21] Dhaya R, Kanthavel R, and Ahilan A, Developing an energy-efficient ubiquitous agriculture mobile sensor network-based threshold built-in MAC routing protocol (TBMP). Soft Computing. 2021; pp.1-10.

[22] Ramji DR, Palagan CA, Nithya A, Appathurai A, and Alex EJ, Soft computing based color image demosaicing for medical Image processing. Multimedia Tools and Applications. 2020 79(15): pp.10047-10063.

[23] Appathurai A, Sundarasekar R, Raja C, Alex EJ, Palagan, CA, and Nithya A, An efficient optimal neural network-based moving vehicle detection in traffic video surveillance system. Circuits, Systems, and Signal Processing, 39(2), pp.734-756.

[24] Appathurai A, Carol JJ, Raja C, Kumar SN, Daniel AV, Malar AJG, Fred AL, and Krishnamoorthy S, A study on ECG signal characterization and practical implementation of some ECG characterization techniques. Measurement. 2019; 147(p.106384):