

The effectiveness of big data classification control based on principal component analysis

Mostafa Abdulghafoor Mohammed¹, Mustafa Mahmood Akawee¹, Ziyad Hussien Saleh², Raed Abdulkareem Hasan³, Ahmed Hussein Ali^{4,5}, Tole Sutikno⁶

¹Department of Arabic Language, Faculty of Imam Aadhham University College, Baghdad, Iraq

²Department of Petroleum Systems and Control Engineering, College of Petroleum Processes Engineering, Tikrit University, Tikrit, Iraq

³Department of Electrical techniques, Faculty of Al-Hweijja Technical Institute, Noerthern Technical University, Mosel, Iraq

⁴Department of Computer, College of Education, Al-Iraqia University, Baghdad, Iraq

⁵Department Computer Science, Al-Salam University College, Baghdad, Iraq

⁶Department of Electrical Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Article Info

Article history:

Received Jul 14, 2022

Revised Aug 18, 2022

Accepted Oct 3, 2022

Keywords:

Big data

Classification algorithm

Interpretability

Large-scale datasets

Machine learning

Principal component analysis

Spark ecosystem

ABSTRACT

Large-scale datasets are becoming more common, yet they can be challenging to understand and interpret. When dealing with big datasets, principal component analysis (PCA) is used to minimize the dimensionality of the data while maintaining interpretability and avoiding information loss. It accomplishes this by producing new uncorrelated variables that gradually reduce the variance of the system. In the field of data analysis, PCA is a multivariate statistical technique commonly used to obtain rules explaining the separation of groups in a given situation. Classes are predicted using a classification algorithm, a supervised learning technique that indicates which type of data points will be presented. Creating a classification model using classification algorithms is required before any successful classification can be achieved. It is possible to predict the future using a variety of categorized strategies. It is necessary to reduce the dimensionality of data sets using the PCA approach. This article will begin by introducing the basic ideas of PCA and discussing what it can and cannot do. It will then describe some variants of PCA and their applications, and then show how PCA improves the performance using a series of experiments.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mostafa Abdulghafoor Mohammed

Department of Arabic Language, Imam Aadhham University College

Baghdad, Iraq

Email: alqaisy86@gmail.com

1. INTRODUCTION

Large datasets are becoming increasingly popular in a wide range of fields [1]. The methodologies used to analyze large datasets must substantially reduce their dimensionality in an interpretable manner while retaining the vast majority of the information contained within the data. To achieve this, a variety of approaches have been developed, with principal component analysis (PCA) being one of the first and most widely used [2]. Its underlying idea is straightforward: reduce the dimensionality of a dataset while keeping as much “variability” (i.e., statistical information) as is reasonably practicable. According to the latest statistics, online social networks (OSN) and microblogging websites are attracting the greatest number of Internet users of any sort of website. As the volume, pace, and variety of their content increase at an alarming rate, they become an enticing example of big data [3].

Big data has piqued the interest of researchers who are interested in the investigation of people's thoughts as well as the structure and distribution of users on digital media platforms. As a result, the type and number of postings, comments, and messages exchanged on these websites make it practically impossible to monitor the substance of what is said on them. When it comes to obtaining vital information, data mining has evolved into a powerful method. The categorization of the data obtained via this method is one of the most difficult problems to solve in the field. In the field of statistics, dimensionality reduction is a strategy for lowering the total number of random variables under examination to get a collection of primary variables that may be used in further analysis.

PCA is a technique for reducing the dimensionality of linear datasets, which is used in many applications [4]–[6]. Classification is the term used to describe the process of predicting class labels. Classifiers, or models, are created to forecast which classes will be assigned. When completing classification tasks, it is common to observe an excessive number of attributes. It is on the basis of these attributes that the classification is made, but when there are too many such features, it is more difficult for the training set to assimilate and work on those features. Because of this, dimensionality reduction is required to reduce the number of attributes and allow the classification model to function more efficiently. To evaluate the effectiveness of classification models, performance evaluation measures are employed. Accuracy, specificity, sensitivity, and the receiver operating characteristic (ROC) curve are all important performance measures. Three different datasets were used in this work, and four different categorization models were developed for them. The classification models were created using four different classification algorithms: the stochastic gradient descent algorithm, the kernel approximation algorithm, the J48, and the multilayer perceptron. PCA was applied to each dataset separately to reduce the dimensionality of each dataset independently of the others. The results of the performance measurements were compared before and after the use of PCA, and conclusions were drawn from the relevant data.

Several big data processing systems have been proposed by the research community. In the research previously done [6]–[8], an attempt was made to examine a wide number of articles to cover the numerous facial recognition techniques that make use of independent component analysis (ICA) in their investigations. The examination of previous work done in face recognition research, work that has been linked to ICA, is a critical component of this survey's findings. There are a number of different ICA techniques developed in the work done by Kanaujia *et al.* [6]. Results from PCA can only be utilized to extract signal features, which means that only signal features may be obtained. ICA research is still quite active today, which gives reason to be optimistic about additional developments in the technique and its algorithmic implementations in the future. Naik *et al.* [7] describe the existing literature on the PCA application for processing surface electromyographic (sEMG) signals acquired from various muscles, notably those in the face and upper and lower limbs, and provide an overview of the PCA application. It has been used in this context as an unsupervised feature extraction strategy to reduce the dimensionality of surface electromyography data to accommodate a variety of myoelectric applications [8]. Research into determining a city's livability is essential both academically and in practical applications. In-depth knowledge of the concept of city livability, as well as a review of the literature, went into the process of developing the indicator system for city livability evaluation. The PCA was performed on the city indices in conjunction with statistical data from 18 cities in Shandong Province to offer an objective assessment of the city's livability, the conclusions of this study are mostly consistent with the facts [9]–[11]. PCA has been modified for a variety of data types, including binary data, ordinal data, compositional data, discrete data, symbolic data, and data with a specific structure, such as time series or datasets with common correlation matrices. Another statistical analysis technique, such as linear regression, has been directly influenced by PCA or PCA-related methodologies (with principal component regression and even simultaneous clustering of both people and variables) [12]–[16]. A role of big data analysis in the healthcare context, a comprehensive analysis of the various techniques involved in mining big data, and an overview of big data applicability in healthcare have been presented [17]. The role of big data analysis in the healthcare context and a comprehensive analysis of the various techniques involved in mining big data, as well as an overview of big data applicability in healthcare, have been presented. On the other hand, dimension reduction using PCA, singular value decomposition (SVD), non-negative matrix factorization (NMF), and sparse coding (SC) in image processing for the mammalian brain has been introduced [18]. Dimensionality reduction techniques, PCA, and NMF, have been used to explore the dependency of the standard PCA, NMF, ICA, and SVD algorithms on the selected number of dimensions [19].

The main contribution of this paper was to demonstrate the usefulness of principle component analysis methods for attribute reduction in massive data classification via machine algorithms by utilizing machine learning techniques. This study investigated whether or not a classifier performed better before or after it was subjected to a PCA-based dimensionality reduction procedure. This paper is outlined as follows: section 2 reviewed the related works regarding the recommended method, while sections 3 analyzed the evaluation results from the several experiments conducted in this study, while section 4 concluded the paper.

2. METHOD

2.1. Spark ecosystem

The apache spark ecosystem is a highly scalable, fast, and in-memory massive data processing engine that was initially developed in the algorithms, machines and people lab (AMPLab). It allows developers to construct distributed applications in Java, Python, Scala, and R. The apache spark streaming, apache spark structured query language (SQL), apache spark GraphX, and apache spark MLlib libraries are the four main components of the distribution. Apache spark SQL is the key scheduling module when it comes to stream processing inside a highly fault-tolerant and batch analytics framework. It implements relational queries to mine multiple database systems and provides a data abstraction model known as DataFrames to do so. Four models are generated from resilient distributed datasets (RDD's) extensive expressive capabilities: spark SQL, spark streaming, machine learning library (MLlib), and GraphX [20], which, when combined with the fundamental pieces of spark, form a simple overall spark framework, as illustrated in Figure 1. RDD's expressive capabilities are demonstrated in Figure 1. The following sections provide an overview of each fundamental component. The RDD data structure, as well as its complete operational interface in spark core, is the key to this technique. Spark SQL is a spark module that works with structured data and is used to process it. In this paper, we will look into spark Streaming, which makes it straightforward to develop a fault-tolerant and highly adaptable streaming application. A key feature of spark's lightweight and low-latency scheduling mechanism is that it effectively enables streaming computation by dividing the streaming computation into a succession of short batch processes [21]. Spark MLlib is a library that allows you to do machine learning tasks. As the name implies, spark MLlib is a spark-based extendable machine-learning library that includes common learning algorithms and techniques such as binary classification, linear regression, clustering, collaborative filtering (gradient descent), and low-level optimization primitives. GraphX by spark is a graph-based visualization tool. Spark is used for graph computing and parallel graph computation. GraphX is a new spark application programming interface (API) that has been developed for the spark platform. Figure 1 shows the spark's overall structure.

2.2. Principal components analysis

The PCA method is the fundamental algorithm for every facial recognition project [22]. When it comes to feature selection and dimension reduction, PCA approaches are quite popular. When dealing with difficult data, the PCA is an easy, yet efficient, non-parametric approach for providing a statistical description of the data by discovering hidden patterns and eliminating noise. It is well understood that this method, which is utilized in many fields, interprets data by relying on second-order statistics and that the results produced are quite reliable. In brief, PCA is an orthogonal linear transformation that decorrelates multivariate data by projecting it onto a new coordinate system known as the principal components (PC). They may be thought of as a collection of spatial directions onto which the projected data assures the maximum amount of variation. The fact that there is a logical way to organize the PCs based on the variation of the predicted data should be noted. The variances of the data projected onto each component are decreasing in succession as the first, second, and third components get smaller. In most cases, the original data can be accurately reproduced with only a small number of starting PCs. It is common practice to use the analytic hierarchy process (AHP) and the Delphi method to calculate the weights of assessment indicators in multi-index assessing models. However, these techniques are more or less subjective, even though some technical instruments are employed. Furthermore, because there are so many different indices of city livability, it is difficult to conduct an adequate analysis of them. In consideration of the aforementioned factors, PCA is employed in this study to reduce the dimensionality of the indicator system to a minimum. It is a statistical approach known as PCA. A new set of indicators, Y, is created by linearly combining the indicators based on the correlation between them. This new set of indicators is known as the "primary component" of the indicator system (1). There are fewer indicators in the primary components, but they retain the same amount of information as did the prior indicator system. PCA makes it easy to evaluate a dimensionally reduced indicator system because it cuts down on the number of dimensions that need to be examined. PCA makes it easy to evaluate a dimension-reduced indicator system because it cuts down on the number of dimensions that need to be looked at.

2.3. Classifying algorithm and performance evaluat

In the field of data mining, "classification" refers to the methods used to predict, categorize, and organize data into useful groups based on the properties of each datum. Each class has its own set of rules and algorithms that must be followed. Some classification systems, such as the J48, rely on decision tree rules to achieve their classification. Figure 2, such as Bayesian networks and NaiveBayes, are based on artificial intelligence and neural networks to achieve their classification objectives. Classification techniques are supported by a wide range of applications and should be applied to a wide range of datasets to maximize their effectiveness. Furthermore, when compared to other classification systems, all classification algorithms will fail to accurately predict data categorization. To select the most appropriate classification method for testing data, the data must be compatible with the rules, algorithms, and other characteristics of the classification strategy being used [23].

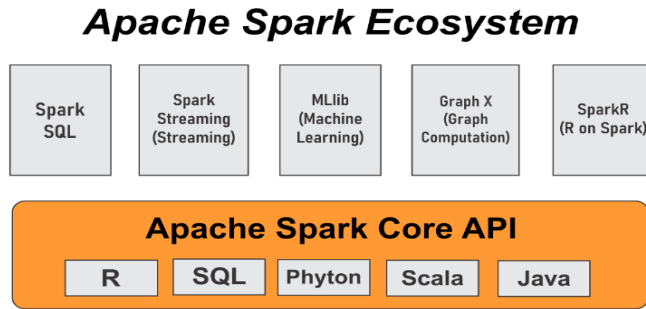


Figure 1. Spark overall structure

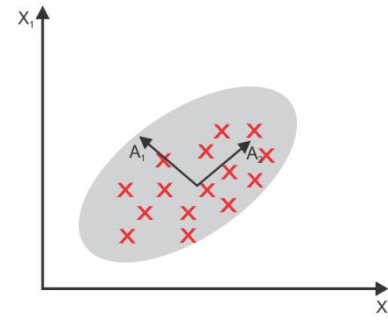


Figure 2. Principal components

The stochastic gradient descent algorithm is a strategy for iteratively optimizing an objective function with adequate smoothness attributes. It is known as iterative optimization (e.g., differentiable or sub-differentiable). It is referred to as a stochastic approximation to gradient descent optimization because it substitutes the genuine gradient (which is produced from the entire data set) with an estimate of the gradient as calculated from a randomly selected subset of the data. Especially in high-dimensional optimization problems, this reduces the computational cost by allowing for faster iterations in exchange for a lower convergence rate. The Kernel Approximation Algorithm The support-vector machine is the most well-known member of the kernel machine class of pattern analysis tools, and it is also the simplest to understand and implement (SVM). The primary goal of pattern analysis is to uncover and study various types of associations (for example, clusters, rankings, PC, correlations, and classifications) in datasets, which is accomplished via the use of statistical techniques [24]–[26]. Numerous techniques that deal with these problems need the explicit conversion of raw input into feature vector representations using a feature map that is provided by the user. Unlike this, kernel techniques only need a kernel chosen by the user, which can be shown as a similarity function over pairs of raw data points.

J48: the classifier in C4.5 has been tuned for performance. The J48 approach is based on the concept of a decision tree. The J48 technique is a data classification strategy that is commonly employed. J48 rules and algorithms make use of decision tree techniques, which contain a central node and a number of subordinate branches. Each branch or leaf makes a decision that has an associated consequence that is separate from the others. When using the J48 classification strategy, some datasets may have very large tree models with many branches. This can lead to different results than when only a few branches of the decision tree are used, as the following example shows.

Multilayer perceptron: The multilayer perceptron classifier is based on Artificial Intelligence and Neural Networks and does not require any further certification. A multi-layer perceptron (MLP) is a perceptron with at least three layers of information. The input layer is the first layer, the second layer is the hidden layer, and the last layer is the output layer, in that order. An MLP is a feedforward neural network with one or more hidden layers that operate in the forward direction. In the MLP model, each node in each layer is connected to all of the nodes in the other levels by a network of links. It is a way to classify data that is used in neural networks, deep learning, and other applications that deal with data processing.

3. RESULTS AND DISCUSSION

This research made use of large categorization datasets provided by the University of California, Irvine (UCI) data repository, from which six datasets were selected. The important aspects of these datasets are summarized in Table 1, which includes the number of records, attributes, and classifications for each data collection for each dataset. The datasets were selected from a variety of fields where there is a dearth of available data. The data in Table 1 is organized by dataset size and includes information on the number of features and data types included in each dataset selected. The datasets that were selected are identified as DS1 to DS6. It is estimated that there are between 102944 and 11000000 instances of each feature in these datasets and that the number of features ranges between 18 and 116. Every dataset has its own set of characteristics and attributes that set it apart from the other two datasets in the collection. All of the datasets are either entirely numerical or numerical plus textual. The accuracy, sensitivity, specificity, and ROC curve values were all used to evaluate the performance of the classification models, equations (1, 2, 3, 4, and 5). In the next section, you will find the formulas for measuring accuracy, sensitivity, and specificity. Table 1 shows the details of the dataset description. Table 2 shows classification evaluation without PCA, and Table 3 shows classification evaluation with PCA.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (3)$$

Table 1. Datasets description

Data No.	Name	No of record	No of attributes	No of classes
DS1	Dota2	102944	116	2
DS2	Covtype	581012	54	7
DS3	Covtype-2	581012	54	2
DS4	SUSY	5,000,000	18	2
DS5	Botnet Attacks	7,062,606	115	10
DS6	Higgs	11,000,000	28	2

Table 2. Classification evaluation without PCA

Dataset	Classifier	Accuracy	Sensitivity	Specificity
DS1	Stochastic gradient descent algorithm	0.8944	0.7988	0.8937
	Kernel approximation algorithm	0.8886	0.8829	0.8083
	J48	0.8631	0.843	0.7783
	Multilayer perceptron	0.9104	0.7832	0.8541
DS2	Stochastic gradient descent algorithm	0.7862	0.9344	0.9388
	Kernel approximation algorithm	0.7633	0.8269	0.9313
	J48	0.7942	0.9731	0.9298
	Multilayer perceptron	0.7586	0.8786	0.9231
DS3	Stochastic gradient descent algorithm	0.8244	0.909	0.8616
	Kernel approximation algorithm	0.7271	0.9625	0.8957
	J48	0.7352	0.8703	0.8308
	Multilayer perceptron	0.7688	0.8737	0.8774
DS4	Stochastic gradient descent algorithm	0.8833	0.8686	0.8493
	Kernel approximation algorithm	0.8956	0.9168	0.8429
	J48	0.8902	0.8018	0.9281
	Multilayer perceptron	0.9345	0.7819	0.8687
DS5	Stochastic gradient descent algorithm	0.8677	0.9256	0.9269
	Kernel approximation algorithm	0.9652	0.8879	0.8398
	J48	0.8279	0.8928	0.8542
	Multilayer perceptron	0.8914	0.8834	0.9465
DS6	Stochastic gradient descent algorithm	0.8338	0.8266	0.887
	Kernel approximation algorithm	0.973	0.8196	0.8751
	J48	0.9827	0.7998	0.7816
	Multilayer perceptron	0.7972	0.8441	0.9520

Table 3. Classification evaluation with PCA

Dataset	Classifier	Accuracy	Sensitivity	Specificity
DS1	Stochastic gradient descent algorithm	0.9729	0.9445	0.8799
	Kernel approximation algorithm	0.9689	0.9944	0.9083
	J48	0.9072	0.9179	0.9252
	Multilayer perceptron	0.9108	0.8999	0.8994
DS2	Stochastic gradient descent algorithm	0.8895	0.9113	0.8486
	Kernel approximation algorithm	0.9058	0.9345	0.9062
	J48	0.8984	0.9757	0.8916
	Multilayer perceptron	0.9498	0.9465	0.9194
DS3	Stochastic gradient descent algorithm	0.9006	0.8937	0.8606
	Kernel approximation algorithm	0.9674	0.919	0.9059
	J48	0.9126	0.8917	0.9008
	Multilayer perceptron	0.9999	0.9903	0.8911
DS4	Stochastic gradient descent algorithm	0.9458	0.9678	0.8918
	Kernel approximation algorithm	0.8991	0.9244	0.9161
	J48	0.9747	0.9405	0.8916
	Multilayer perceptron	0.9355	0.9291	0.8623
DS5	Stochastic gradient descent algorithm	0.9262	0.9842	0.8698
	Kernel approximation algorithm	0.9395	0.9337	0.858
	J48	0.9176	0.9754	0.8964
	Multilayer perceptron	0.9672	0.986	0.8957
DS6	Stochastic gradient descent algorithm	0.9327	0.9778	0.8908
	Kernel approximation algorithm	0.9946	0.9032	0.8560
	J48	0.9112	0.9300	0.9135
	Multilayer perceptron	0.9915	0.9412	0.9340

For classifying data, the receiver operating characteristics (ROC) curve is an important assessment statistic to use in determining the efficacy of a classification model. The curve is plotted with the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis. The formulae for calculating TPR and FPR are presented in the following sections.

$$\text{TPR} = \text{TP} / (\text{TP} + \text{FN}) \quad (4)$$

$$\text{FPR} = \text{FP} / (\text{TN} + \text{FP}) \quad (5)$$

True positive is represented by the letters TP, true negative by the letters TN, false positive by the letters FP, and false negative by the letters FN. True positives are obtained when the model correctly predicts the positive classes, and true negatives are attained when the model correctly predicts the negative classes. In the same way, false positives occur when the system predicts the positive classes incorrectly and false negatives occur when the system wrongly predicts the negative classes. The best prediction model should provide a point with 100 percent sensitivity in the top left corner of the ROC space, indicating that no false negatives occur.

Machine learning repository and each dataset was submitted to four different grading methods to determine its classification. The material was then presented linearly. According to the findings, when PCA was applied to the dataset, the classifiers performed much better. When dealing with linear data, PCA can be used to reduce the dimensionality of the data. Additionally, PCA may be used to improve the performance of classifiers in a range of classification challenges, including supervised learning. There are, however, certain limitations to this. Certain conditions require a significant difference between performance measurements taken before and after PCA. In Table 3, we show classification evaluation with PCA, but in other cases, the difference is minimal. It is not possible to ensure that the use of PCA will result in a substantial improvement in the performance of the classifying model for every dataset. The output varies based on the dataset used to calculate it.

4. CONCLUSION

This study aimed to demonstrate the usefulness of principle component analysis methods for attribute reduction in massive data classification via machine algorithms by utilizing machine learning techniques. It was investigated in this study whether or not a classifier performed better before and after it was subjected to a PCA-based dimensionality reduction procedure. Six different datasets were downloaded from the UCI machine learning repository, and each dataset was submitted to four other grading methods to determine its classification. The material was presented linearly. According to the findings, the classifiers performed much better when PCA was applied to the dataset. When dealing with linear data, PCA can be used to reduce the dimensionality of the data. In addition, PCA may be used to improve the performance of classifiers in a range of classification challenges, including supervised learning. There are, however, certain limitations to this. Certain conditions require a significant difference between performance measurements taken before and after PCA, but the difference is minimal in other cases. It is impossible to ensure that the use of PCA will substantially improve the performance of the classifying model for every dataset. The output varies based on the dataset used to calculate it.




REFERENCES

- [1] A. L. Oliveira, "Biotechnology, big data and artificial intelligence," *Biotechnology Journal*, vol. 14, no. 8, p. 1800613, Aug. 2019, doi: 10.1002/biot.201800613.
- [2] U. Demšar, P. Harris, C. Brunson, A. S. Fotheringham, and S. McLoone, "Principal component analysis on spatial data: An overview," *Annals of the Association of American Geographers*, vol. 103, no. 1, pp. 106–128, Jan. 2013, doi: 10.1080/00045608.2012.689236.
- [3] W. A. Hammood, R. A. Arshah, S. M. Asmara and O. A. Hammood, "User Authentication Model based on Mobile Phone IMEI Number: A Proposed Method Application for Online Banking System," *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, 2021, pp. 411-416, doi: 10.1109/ICSECS52883.2021.00081.
- [4] T. Zhang and B. Yang, "Big data dimension reduction using PCA," in *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, Nov. 2016, pp. 152–157, doi: 10.1109/SmartCloud.2016.33.
- [5] R. Naik, D. P. Singh, and J. Chaudhary, "A Survey on comparative analysis of different ICA based face recognition technologies," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Mar. 2018, pp. 1913–1918, doi: 10.1109/ICECA.2018.8474860.
- [6] M. Kanaujia and G. Srivastava, "ECG signal decomposition using PCA and ICA," in *2015 National Conference on Recent Advances in Electronics & Computer Engineering (RAECE)*, Feb. 2015, pp. 301–305, doi: 10.1109/RAECE.2015.7510211.
- [7] G. R. Naik, S. E. Selvan, M. Gobbo, A. Acharyya, and H. T. Nguyen, "Principal component analysis applied to surface electromyography: A comprehensive review," *IEEE Access*, vol. 4, pp. 4025–4037, 2016, doi: 10.1109/ACCESS.2016.2593013.
- [8] L. Yin and Y. Yin, "Research on assessment of city livability based on principle component analysis-taking shandong province for example," in *2009 International Conference on Management and Service Science*, Sep. 2009, pp. 1–4, doi: 10.1109/ICMSS.2009.5301952.




- [9] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, pp. 1–16, Apr. 2016, doi: 10.1098/rsta.2015.0202.
- [10] H. Kumar, P. J. Soh, and M. A. Ismail, "Big data streaming platforms: A review," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, no. 2, pp. 95–100, 2022, doi: 10.52866/ijcsm.2022.02.01.010.
- [11] A. Rghioui, J. Lloret, and A. Oumnad, "Big data classification and internet of things in healthcare," *International Journal of E-Health and Medical Communications*, vol. 11, no. 2, pp. 20–37, Apr. 2020, doi: 10.4018/IJEHMC.2020040102.
- [12] C.-C. Hung, E. Song, and Y. Lan, "Dimensionality reduction and sparse representation," in *Image Texture Analysis, Foundation, Models and Algorithms*, Cham: Springer International Publishing, 2019, pp. 103–127.
- [13] S. Mallapragada, M. Wong, and C.-C. Hung, "Dimensionality reduction of hyperspectral images for classification," *Ninth International Conference on Information*, Tokyo, Japan, 2018.
- [14] R. A. Anugrah *et al.*, "Application of coconut paper motor speed control technology for increasing coconut liquid organic fertilizer productivity," *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*, vol. 2, no. 2, pp. 74–83, 2021.
- [15] D. Cheng, Y. Chen, X. Zhou, D. Gmach, and D. Milojicic, "Adaptive scheduling of parallel jobs in spark streaming," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, May 2017, pp. 1–9, doi: 10.1109/INFOCOM.2017.8057206.
- [16] Y. Zhu, C. Zhu, and X. Li, "Improved principal component analysis and linear regression classification for face recognition," *Signal Processing*, vol. 145, pp. 175–182, 2018, doi: 10.1016/j.sigpro.2017.11.018.
- [17] R. A. Hasan, H. W. Abdulwahid, and A. S. Abdalzahra, "Using Ideal Time Horizon for Energy Cost Determination," *Iraqi Journal For Computer Science and Mathematics*, vol. 2, no. 1, pp. 9–13, 2021, doi: 10.52866/ijcsm.2021.02.01.002.
- [18] S. I. Jasim, M. M. Akawee, and R. A. Hasan, "A spectrum sensing approaches in cognitive radio network by using cloud computing environment," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 750–757, doi: doi.org/10.11591/eei.v11i2.3162.
- [19] H. Abubakar, A. Muhammad, and S. Bello, "Ants colony optimization algorithm in the Hopfield neural network for agricultural soil fertility reverse analysis," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 1, pp. 32–42, 2022, doi: 10.52866/ijcsm.2022.01.01.004.
- [20] H. R. Ibraheem, Z. F. Hussain, S. M. Ali, M. Aljanabi, M. A. Mohammed, and T. Sutikno, "A new model for large dataset dimensionality reduction based on teaching learning-based optimization and logistic regression," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 18, no. 3, pp. 1688–1694, Jun. 2020, doi: 10.12928/telkomnika.v18i3.13764.
- [21] I. Suwarno *et al.*, "Community empowerment in landslide management in Sonyo hamlet," *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*, vol. 1, no. 12, pp. 501–507, 2021.
- [22] Z. A. Abdalkareem, M. A. Al-Betar, A. Amir, P. Ehkan, A. I. Hammouri, and O. H. Salman, "Discrete flower pollination algorithm for patient admission scheduling problem," *Computers in biology and medicine*, vol. 141, p. 105007, Feb. 2022, doi: 10.1016/j.compbiomed.2021.105007.
- [23] A. H. Ali, M. N. Abbod, M. K. Khaleel, M. A. Mohammed, and T. Sutikno, "Large scale data analysis using MLlib," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 19, no. 5, pp. 1735–1746, Oct. 2021, doi: 10.12928/telkomnika.v19i5.21059.
- [24] R. A. I. Alhayali, M. Aljanabi, A. H. Ali, M. A. Mohammed, and T. Sutikno, "Optimized machine learning algorithm for intrusion detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 1, pp. 590–599, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp590-599.
- [25] Z. A. Abdalkareem, O. Z. Akif, F. A. Abdulatif, A. Amiza, and P. Ehkan, "Graphical password based mouse behavior technique," *Journal of Physics: Conference Series*, vol. 1755, no. 1, p. 012021, Feb. 2021, doi: 10.1088/1742-6596/1755/1/012021.
- [26] B. I. Farhan and A. D. Jasim, "A survey of intrusion detection using deep learning in internet of things," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, no. 1, pp. 83–93, Jan. 2022, doi: 10.52866/ijcsm.2022.01.01.009.

BIOGRAPHIES OF AUTHORS






Mostafa Abdulghafoor Mohammed    currently works at the Al-Imam Al-aadham University college. Mostafa does research in information technology, computer communications (networks), cloud computing and communication engineering. His current project is 'offloading in mobile cloud computing'. He has finish his master from Computer Science Department at BAMU University, India, and he finish Ph.D in Computer Science and IT at University Polytechnic of Bucharest, Romania. He can be contacted at email: alqaisy86@gmail.com.






Mustafa Mahmood Akawee    currently works at the al-imam Adham. Mostafa does research in information systems (business informatics), computer communications (networks) and communication engineering. His current project is 'offloading in mobile cloud computing'. He has finished his master from Computer Science Department at Kharkiv University, Ukraine. He can be contacted at email: it.diyala2@gmail.com.






Ziyad Hussien Saleh    currently works at the College of Petroleum Processes Engineering, Tikrit University. Ziyad does research in information technology, control, and communication engineering. He finished Ph.D in Computer and Control Engineering University of technology. He can be contacted at email: ziad_1966@tu.edu.iq.






Raed Abdulkareem Hasan    currently works at the Al-Hawija Technical Institute, Northern Technical University. Raed does research in information systems (business informatics), computer communications (networks) and communication engineering. His current project is 'offloading in mobile cloud computing'. He has finished his master from Computer Science Department at BAMU University, India. He can be contacted at email: raed.isc.sa@gmail.com.



Ahmed Hussien Ali    was born in Baghdad, Iraq, in 1988. He received the B.Sc. degree in Computer Science from the University of Al-Mustansiriyah, Iraq, in 2010, and the M.Sc. degree from BAMU University, India. He got Ph.D. from ICCI, Informatics Institute for Postgraduate Studies, Baghdad, Iraq and Faculty member, Department Computer Science, College of Education, Al-Iraqia University, Adhamyia, Baghdad, Iraq. He can be contacted at email: msc.ahmed.h.ali@gmail.com.



Tole Sutikno    is currently employed as a lecturer in the Electrical Engineering Department at Universitas Ahmad Dahlan (UAD), which is located in Yogyakarta, Indonesia. In 1999, 2004, and 2016, he graduated with a Bachelor of Engineering from Universitas Diponegoro, a Master of Engineering from Universitas Gadjah Mada, and a Doctor of Philosophy in Electrical Engineering from Universiti Teknologi Malaysia. All three degrees are in the field of electrical engineering. Since the year 2008, he has held the position of Associate Professor at the University of Ahmad Dahlan in Yogyakarta, Indonesia. He is currently the Head of the Embedded Systems and Power Electronics Research Group in addition to holding the position of Editor-in-Chief of TELKOMNIKA. His research interests include the areas of digital design, industrial applications, industrial electronics, industrial informatics, power electronics, motor drives, renewable energy, FPGA applications, embedded systems, artificial intelligence, intelligent control, digital libraries, and intelligent control. He can be contacted at email: tole@te.uad.ac.id.