

Large dataset partitioning using ensemble partition-based clustering with majority voting technique

Vunnava Dinesh Babu, Karunakaran Malathi

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

Article Info

Article history:

Received Jun 4, 2022

Revised Oct 10, 2022

Accepted Oct 17, 2022

Keywords:

Clustering

Ensemble clustering

Large dataset

Majority voting technique

Partitioning

ABSTRACT

Large datasets have become useful in data mining for processing, storing, and handling vast amounts of data. However, handling and processing large datasets is time-consuming and memory intensive. As a result, the researchers adopted a partitioning strategy to improve controllability and performance and reduce the time and memory required to handle large datasets. Unfortunately, the numerous clustering techniques available in the literature could confuse experts in choosing the best techniques for a given dataset. Furthermore, no clustering technique can tackle all problems, such as cluster structure, noise, or density. To manage large datasets, existing clustering techniques need scalable solutions. Therefore, this paper proposes an ensemble partition-based clustering with a majority voting technique for large dataset partitioning using the aggregation of k-means, k-medoids, fuzzy c-means, expectation-maximization (EM) and density-based spatial clustering of applications with noise (DBSCAN) techniques. These techniques cluster the large dataset individually in the first stage. The final clusters are discovered in the next stage through a majority voting technique among the five clustering algorithms. These five clustering algorithms assigned data instances to the cluster with the most votes. The experimental findings demonstrate that the ensemble partition-based clustering method surpasses the other five clustering algorithms in terms of execution time and accuracy.

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



Corresponding Author:

Vunnava Dinesh Babu

Department of Computer Science and Engineering, Saveetha School of Engineering

Saveetha Institute of Medical and Technical Sciences

Chennai, India

Email: dineshbabuvunnava@gmail.com

1. INTRODUCTION

Machine-generated datasets have become vaster in volume and internationally dispersed in current years [1]. Big data may be classified according to several factors (such as volume, velocity, variety, veracity, value, and volatility). Big data are large, difficult-to-manage datasets that are collected, stored, examined, and displayed. A collection of pertinent data with special or multiple qualities is called a dataset. Large datasets are made up of enormous volumes of data that are larger than what can be handled by standard database management techniques. Big data has become commonplace due to the ability of contemporary data-intensive technologies to handle structured, unstructured, and semi-structured information sources and formats [2].

The necessity for effective partitioning algorithms that meet processing of documents, memory usage, and execution time criteria has increased as datasets have grown in size. A problem with large data is grouping objects so which information from than data from other groups, the same group's data are more comparable. For example, big data is used to use communications systems, health services, computational biology, finance,

advertising, biotechnology, insurance, urban planning, seismic research, online content categorization, and transportation services [3]. Big data comes in three different kinds. They are structured, unstructured, and semi-structured, respectively.

Structured data refers to any data which is kept, accessed, and used in accordance with a standard format. Any data that has an uncertain form or structure is classified as "unstructured" data. Over time, skill in computer science has been more efficient in creating strategies for working with such data (when the format is properly known in advance) and obtaining its worth. In addition to its huge amount, unstructured data presents a number of processing challenges that should be overcome in order to extract value from it. A heterogeneous data source is one that combines unstructured data types with basic text files, images, and videos.

'Semi-structured' data can contain both kinds of data. Semi-structured data can be structured in a format, although it isn't totally described. Semi-structured data is demonstrated by the XML file's descriptions of data. Massive amounts of data can be gathered and analyzed using big data analytics in order to yield valuable data. Large data is often handled using a variety of data mining approaches. However, the techniques did not succeed in reducing the dimension. As a result, clustering is a crucial data mining approach for large data analysis. Clustering is the process of grouping related data points in order to reduce the dimensionality curse [4].

Clustering is utilized to identify clusters formed for a dataset and to find meaningful groups of objects. The conventional clustering approaches work well with small datasets. Large datasets need to be clustered such that each object or data point is related to every other object or data point in the cluster. Clustering problems could be addressed using a variety of clustering disciplines [5]. One could find hidden similarities and important ideas by autonomously clustering similar objects when a large quantity of information is divided into a few categories. Users are able to comprehend a large amount of data due to this. Based on the intrinsic method utilized to categories data in clusters, clustering algorithms can be categorized into five types. Partitioning algorithms, hierarchical algorithms, density-based algorithms, grid-based algorithms, and model-based algorithms are some of these [6]. There are already a number of algorithms from each type that can be found in the literature and have been effectively applied to actual data mining scenarios [7].

Partition-based algorithms are undoubtedly the most popular and commonly used algorithms because they are easy to learn, implement, and have a low time complexity compared to other methods [8]. Techniques for partition-based clustering separate a data set into k divisions, each of which represents a cluster. A typical data clustering process starts with a group of information items and splits them into k clusters using Euclidean distance and other similarity distance metrics. Partition-based clusters could satisfy a few of following requirements: i) Each cluster should have at least one data item and ii) In non-fuzzy clustering algorithms, each object must only be present in one cluster [9].

Unfortunately, the sheer variety of clustering approaches accessible in the literature may make it difficult for specialists to determine which strategies are optimal for a given dataset. Moreover, no clustering technique can address all issues, including cluster structure, noise, and density. In addition, Scalable solutions are needed for the existing clustering techniques to handle large datasets. As a result, using the aggregation of k -means, k -medoids, fuzzy c -means, expectation-maximization (EM), and density-based spatial clustering of applications with noise (DBSCAN) algorithms, this paper proposed an ensemble partition-based clustering with majority voting technique for large dataset partitioning. In the first stage, these five clustering techniques cluster the huge dataset individually. The final clusters are discovered in the next stage using a majority voting technique among the five clustering algorithms.

The subsequent sections will be described in the following order. Section 2 addresses existing work on large dataset partitioning and clustering. Section 3 describes the proposed ensemble partition-based clustering with a majority voting technique for large dataset partitioning. The proposed work's experimental results are presented in section 4. Finally, in section 5, the paper's conclusion is explored.

2. RELATED WORK

The existing works on large dataset partitioning and clustering approaches are discussed in this section. To improve clustering accuracy, Lu *et al.* [10] proposed an incremental k -means clustering technique based on density. Each basic cluster is made up of the centre points whose density is bigger than or equivalent to the given threshold and points inside the density range after the density of the data points has been calculated. The distance between the two cluster centres is then used as a criterion for combining the basic cluster. The remaining points are then divided into the clusters that are closest to them. On the Hadoop cloud computing platform, the shared database was used to simulate the shared memory space and parallelize the algorithm in order to increase the method's effectiveness and decrease its time complexity. The authors concluded that their approach has a more than 10% clustering accuracy greater than the other two algorithms.

Combining the advantages of specialized machine learning applications on mobile devices, a neural-processor-based k -means clustering algorithm was described by Awad *et al.* [11]. The solution was built to run on a single-instruction machine processor contained in the processor of the mobile device. Utilizing k -means

clustering in a shared technique based on mobile machine learning to effectively manage enormous data clustering over the network is achievable. The scientists observed that putting a neural engine processor on a mobile Smartphone device might accelerate the speed of the clustering approach, leading in a two-fold enhancement in cluttering effectiveness when compared to standard desktop/laptop processors. Moreover, compared to parallel and distributed k-means, the number of iterations necessary to create (k) clusters was reduced by up to two-fold.

Heidari *et al.* [12] utilized a Hadoop platform running MapReduce to try out a new approach for clustering big data with differing densities. This study's main concept is to estimate each point's density using its local density. The issue of linking clusters of various densities is avoided by this technique. The MapReduce paradigm is used to develop and compare their algorithm, which displays the best changing density clustering ability and scalability. To enhance clustering accuracy while reducing complexity, Annathurai *et al.* [13] presented the Sørensen-dice indexing based weighted iterative x-means clustering (SDI-WIXC) technique. The SDI-WIXC method clusters related data points more quickly and accurately. Initially, a number of data points are gathered from a huge dataset. The given dataset is divided into "X" clusters, each with a different weight value. Weighted iterated x-means clustering (WIXC) groups data points according to a similarity measure. The Sørensen-Dice Indexing Procedure is utilized to determine how comparable cluster weight values and data points are. Data points are organised into a certain cluster based on the weight value of the cluster and the weight value of the data point. Furthermore, the WIXC technique enhances cluster assignments by employing a Bayesian probability criterion to repeat subdivisions. This, in turn, aids in the grouping of all data points, hence enhancing clustering accuracy. Experimental evaluations are conducted for a number of factors, such as clustering accuracy, clustering time, and space complexity in proportion to the number of data points. According to the authors, the SDIWIXC approach achieves good clustering accuracy with minimal time and space complexity.

Channel modelling has a lot of advantages in the big data era, especially when it comes to utilising algorithmic methods implemented for big data applications. In this context, He *et al.* [14] discusses the challenges and potential in clustering-enabled wireless channel modelling. Firstly, certain famous clustering algorithms are discussed, which have the potential to enable clustered channel modelling. The purpose of cluster-based channel modelling is then explained. Next, the most common cluster concepts utilised in channel models are presented, and current clustering and monitoring algorithms are examined and contrasted. Finally, numerous intriguing channel clustering research challenges are discussed.

For intrusion detection systems (IDS), Peng *et al.* [15] proposed a clustering method using mini batch k-means and principal component analysis. To increase the effectiveness of clustering, the strings are first digitised using a preprocessing approach, and then the data set is normalised. Second, the principal component analysis method is utilised to reduce the dimension of the collected data set before the small batch K-means technique is utilized to enhance clustering effectiveness. To stop the algorithm from reaching the optimum solution, the authors used K-means++ to initiate the cluster nodes. To create the clustering result easier to recognise, they used the CalsskiHarabasz indication. The findings of the experiment and computation time evaluation demonstrated that their technique is successful and efficient when compared to other techniques. Their clustering technique, in particular, could be employed for IDS in a huge data context.

Ilango *et al.* [16] presented a massive data clustering strategy based on artificial bee colonies. The major goal of the ABC technique is to lessen execution time and optimise the optimum cluster for different sizes of the dataset. To deal with this, we're moving to a distributed environment to save time and improve accuracy. The ABC algorithm is utilized to solve numerical optimizing issues, particularly clustering, by simulating the behaviour of real bees. The size of the dataset is changed for the algorithm, and the proper timings are mapped. The outcome is observed for a variety of fitness and probability values derived from the employed and onlooker phases of the ABC method, from which also calibrations of classification percentage error are carried out. The ABC Algorithm is executed utilizing mapper and reducer programming in the Hadoop context. The ABC approach minimises the execution time and classification error for identifying ideal clusters, according to an experimental outcome. In terms of time efficiency, the authors concluded that the ABC scheme outperforms differential evolution (DE) and particle swarm optimization (PSO).

Nursing work in the operating room is characterized by long hours, high complexity, and hard lifting, all of which significantly impact the operation's quality. Clinical nursing and nursing administration face a number of real problems that can be resolved with the aid of data mining-based operating room nursing recommendations. The broadly used data mining method of clustering is selected as the study's object of research by Wu *et al.* [17], and the effect of this algorithm on operating room nursing suggestions is examined. In China, there is currently very little data mining technology development in the nursing profession. By investigating the real use in the area of nursing, the authors hope to bring fresh ideas for the area of nursing research.

Manogaran *et al.* [18] used a Bayesian hidden Markov model (HMM) with Gaussian mixture (GM) clustering technique to forecast the DNA copy number variation across the genome. The pruned precise linear time approach, binary segmentation technique, and segment neighbourhood technique are contrasted with the Bayesian

HMM with GM clustering technique. The authors concluded that their suggested change detection technique produced successful outcomes. Rajyalakshmi *et al.* [19] focused on a number of clustering techniques for sizable datasets like big data, including grid clustering, density-based clustering, hierarchical techniques, and dividing techniques. The most crucial parameter, time complexity, is utilised to differentiate all algorithms. Murugesan *et al.* [20] described their analysis of homomorphic approach for data protection in fog computing. Ramasamy *et al.* [21] used SVM models to enhance the sentiment assessment of the Twitter dataset. Through the use of a task scheduling algorithm, the makespan and resource usage in the fog computing setting are optimised [22]. In [23] research contemplation, Jayanthi *et al.* engaged in a thorough discussion on data mining in education utilizing a new stacking classification and prediction algorithm intensively ambient assisted living for elderly [24].

3. PROPOSED METHOD

The enormous number of clustering methods available in the literature may perplex specialists attempting to select the best method for a large dataset. Furthermore, no single clustering algorithm can tackle all difficulties, such as cluster structure, noise, or density. To address this issue, this work suggested an ensemble partition-based clustering with majority voting technique for large dataset to address this issue using the aggregation of k-means, k-medoids, fuzzy c-means, EM, and DBSCAN techniques partitioning.

These five clustering techniques cluster the huge dataset individually in the first stage. The final clusters are identified in the next stage, which involves a vote procedure among the five clustering techniques. It means that these five clustering techniques assigned each data instance to the cluster with the most votes. In comparison to other clustering algorithms, it increases clustering accuracy while shortening clustering time. Figure 1 displays the ensemble techniques flow diagram. The ensemble partition-based clustering with the majority voting technique is shown in Algorithm 1.

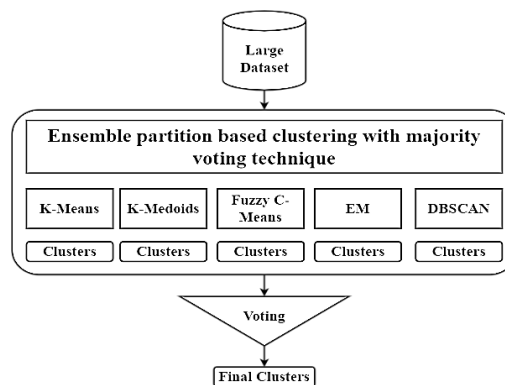


Figure 1. Flow diagram of the ensemble partition-based clustering with majority voting technique

Algorithm 1. Ensemble partition-based clustering with a majority voting technique

- Input : Large Dataset (D)
- Output : Place each data instance in the cluster with the most votes
- Step 1 : Load D
- Step 2 : K-Means Clustering should be applied for D
- Step 3 : K-Medoids Clustering should be applied for D
- Step 4 : Fuzzy c-means Clustering should be applied for D
- Step 5 : EM Clustering should be applied for D
- Step 6 : DBSCAN Clustering should be applied for D
- Step 7 : For each data instance DI from D
- Step 8 : R1 = Extract the outcome of K-Means cluster of DI
- Step 9 : R2 = Extract the outcome of K-Medoids cluster of DI
- Step 10 : R3 = Extract the result of Fuzzy c-means cluster of DI
- Step 11 : R4 = Extract the result of EM cluster of DI
- Step 12 : R5 = Extract the result of DBSCAN cluster of DI
- Step 13 : R[] □ Put R1, R2, R3, R4, R5
- Step 14 : MVC = Extract most frequent value from R // Majority voting technique
- Step 15 : End For

4. RESULTS AND DISCUSSION

4.1. Dataset description

A Higgs Boson dataset evaluated the proposed Ensemble partition-based clustering with a majority voting technique. The Higgs Boson dataset obtained from the Kaggle data repository [25]. This dataset has 250,000 events, with 33 features including an ID column, a weight column, and a label column. We used 10% of Higgs Boson dataset to evaluate Ensemble partition-based clustering with majority voting technique for simplicity. It contains 25,000 events, with 33 features including an ID column, a weight column, and a label column.

These 33 features are EventId, DER_mass_MMC, DER_mass_transverse_met_lep, DER_mass_vis, DER_pt_h, DER_deltaeta_jet_jet, DER_mass_jet_jet, DER_prodelta_jet_jet, DER_deltar_tau_lep, DER_pt_tot, DER_sum_pt, DER_pt_ratio_lep_tau, DER_met_phi_centrality, DER_lep_eta_centrality, PRI_tau_pt, PRI_tau_eta, PRI_tau_phi, PRI_lep_pt, PRI_lep_eta, PRI_lep_phi, PRI_met, PRI_met_phi, PRI_met_sumet, PRI_jet_num, PRI_jet_leading_pt, PRI_jet_leading_eta, PRI_jet_leading_phi, PRI_jet_subleading_pt, PRI_jet_subleading_eta, PRI_jet_subleading_phi, PRI_jet_all_pt, Weight and Label.

Except for EventId, PRI jet num, and Label, all variables in this dataset are floating point. The "raw" quantities regarding the bunch collision as calculated by the detector are variables prefixed with PRI (for PRImitives). The quantities calculated from the primitive features selected by the ATLAS physicists are prefixed with DER (for DERived).

4.2. Performance measurement

4.2.1. Accuracy

Clustering assessment using class's mode is used to evaluate the performance of Ensemble partition-based clustering performance with majority voting technique. Clustering is done in two steps in this mode to calculate accuracy. The first phase ignores the class attribute and constructs a dataset's clustering. Then, the clustering for a dataset with a class attribute is induced in the next phase. If the findings of both clusters for a data instance are the same, we can assume that the data instance clustered correctly. Otherwise, they're clustered wrongly. The (1) shows the accuracy computation formula for clustering.

$$Accuracy = \frac{\text{Number of Correctly Clustered Instances}}{\text{Total number of instances}} * 100 \quad (1)$$

Table 1 compares the accuracy of clustering techniques for the 10% of the Higgs Boson dataset. Figure 2 shows the accuracy comparison for the 10% of the Higgs Boson dataset. Again, the proposed ensemble clustering accuracy is great compared to other clustering techniques.

Table 1. Clustering techniques accuracy comparison for the 10% of Higgs Boson dataset

Algorithm	Accuracy (in %)
K-Means	70.836
K-Medoids	80.02
Fuzzy c-means	77.844
EM	81.448
DBSCAN	69.092
Proposed Ensemble	82.572

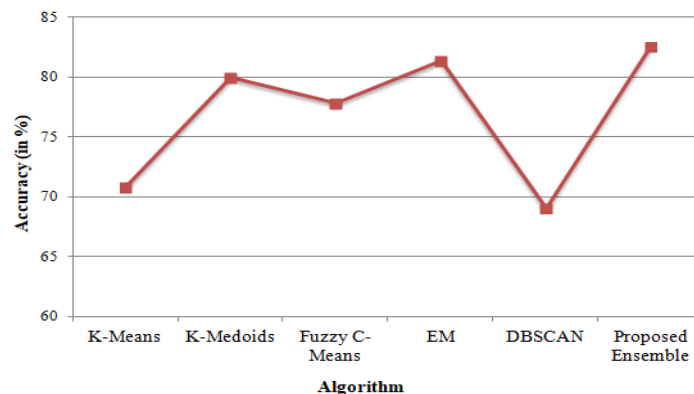


Figure 2. Comparison of accuracy for 10% of Higgs Boson dataset

4.2.2. Execution time

The time it takes to cluster a dataset is called execution time. Figure 3 compares the execution times of the 10% of the Higgs Boson dataset. Figure 3 shows the execution time comparison for the 10% of the Higgs Boson dataset. The proposed ensemble clustering technique takes less time to cluster than existing clustering techniques. The clustering outcomes are assessed using the following metrics: precision (Pre), rand index (RI), normalized mutual information (NMI), and recall (Rec). These evaluation metrics are computed using the classes to cluster assignment (CCA) table are shown.

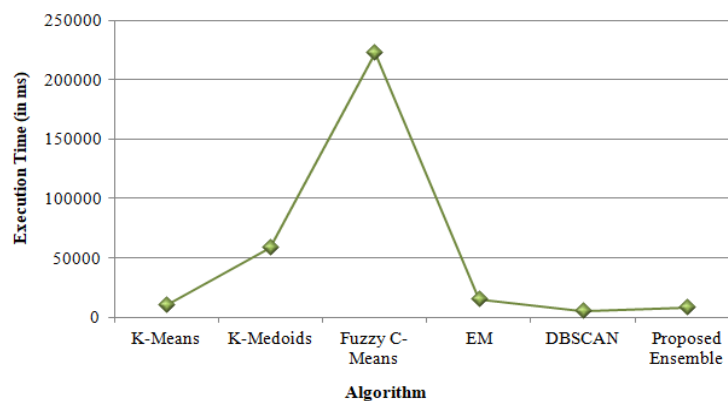


Figure 3. Comparison of execution time for 10% of Higgs Boson dataset

5. CONCLUSION

Data partitioning is a fundamental method of scientific research and data mining that describes a dataset based on similarities between its components. The most common and widely used partitioning technique is partition-based clustering. Due to the digitization of every sector in this information era, data analysts have access to a massive amount of data. Because of the rapid expansion of these datasets, ten-year-old computing systems, programming paradigms, and clustering techniques are insufficient for extracting knowledge from them. To cluster such large datasets, this paper proposed an ensemble partition-based clustering with majority voting technique for large dataset partitioning using the aggregation of k-means, k-medoids, fuzzy c-means, EM and DBSCAN techniques. We employed two key metrics to evaluate the performance of the ensemble clustering technique: accuracy and execution time. The result concluded that the proposed ensemble clustering technique provides the highest accuracy and took the least time to cluster than the other five clustering techniques.




REFERENCES

- [1] A. Kumar, A. Kumar, A. K. Bashir, M. Rashid, V. D. A. Kumar, and R. Kharel, "Distance based pattern driven mining for outlier detection in high dimensional big dataset," *ACM Transactions on Management Information Systems*, vol. 13, no. 1, pp. 1–17, 2022, doi: 10.1145/3469891.
- [2] A. Erraissi, A. Belangour, and A. Tragha, "Meta-modeling of data sources and ingestion big data layers," *SSRN Electronic Journal*, 2018, doi: 10.2139/ssrn.3185342.
- [3] P. V. Desai, "A survey on big data applications and challenges," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018, doi: 10.1109/icicct.2018.8472999.
- [4] J. Amutha, S. Sharma, and S. K. Sharma, "Strategies based on various aspects of clustering in wireless sensor networks using classical, optimization and machine learning techniques: Review, taxonomy, research findings, challenges and future directions," *Computer Science Review*, vol. 40, p. 100376, 2021, doi: 10.1016/j.cosrev.2021.100376.
- [5] C. Sreedhar, N. Kasiviswanath, and P. C. Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-12C) in Hadoop," *Journal of Big Data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0087-2.
- [6] J. Oyelade *et al.*, "Data clustering: algorithms and its applications," *2019 19th International Conference on Computational Science and Its Applications (ICCSA)*. IEEE, 2019, doi: 10.1109/iccsa.2019.000-1.
- [7] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *IEEE Access*, vol. 7, pp. 31883–31902, 2019, doi: 10.1109/access.2019.2903568.
- [8] Z. Ansari, A. Afzal, and T. H. Sardar, "Data categorization using Hadoop Mapreduce-based parallel k-means clustering," *Journal of The Institution of Engineers (India): Series B*, vol. 100, no. 2, pp. 95–103, 2019, doi: 10.1007/s40031-019-00388-x.
- [9] N. Niroomand, C. Bach, and M. Elser, "Vehicle dimensions based passenger car classification using fuzzy and non-fuzzy clustering methods," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2675, no. 10, pp. 184–194, 2021, doi: 10.1177/03611981211010795.
- [10] W. Lu, "Improved k-means clustering algorithm for big data mining under Hadoop parallel framework," *Journal of Grid Computing*, vol. 18, no. 2, pp. 239–250, 2019, doi: 10.1007/s10723-019-09503-0.




- [11] F. H. Awad and M. M. Hamad, "Improved k-means clustering algorithm for big data based on distributed smartphoneneural engine processor," *Electronics*, vol. 11, no. 6, p. 883, 2022, doi: 10.3390/electronics11060883.
- [12] S. Heidari, M. Alborzi, R. Radfar, M. A. Afsharkazemi, and A. R. Ghatari, "Big data clustering with varied density based on MapReduce," *Journal of Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0236-x.
- [13] K. Saravanan and A. Tamarasi, "Sørensen-dice similarity indexing based weighted iterative clustering for big data analytics," *The International Arab Journal of Information Technology*, vol. 19, no. 1, 2022, doi: 10.34028/iajit/19/1/2.
- [14] R. He *et al.*, "Clustering enabled wireless channel modeling using big data algorithms," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 177–183, 2018, doi: 10.1109/mcom.2018.1700701.
- [15] K. Peng, V. C. M. Leung, and Q. Huang, "Clustering approach based on mini batch kmeans for intrusion detection system over big data," *IEEE Access*, vol. 6, pp. 11897–11906, 2018, doi: 10.1109/access.2018.2810267.
- [16] S. S. Ilango, S. Vimal, M. Kaliappan, and P. Subbulakshmi, "Optimization using artificial bee colony based clustering approach for big data," *Cluster Computing*, vol. 22, no. S5, pp. 12169–12177, 2018, doi: 10.1007/s10586-017-1571-3.
- [17] X. Wu, C. Wang, F. Cai, and Y. Wu, "Application of the improved clustering algorithm in operating room nursing recommendation under the background of medical big data," *Journal of healthcare engineering*, vol. 2022, p. 4299280, Mar. 2022, doi: 10.1155/2022/4299280.
- [18] G. Manogaran, V. Vijayakumar, R. Varatharajan, P. M. Kumar, R. Sundarasekar, and C.-H. Hsu, "Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering," *Wireless Personal Communications*, vol. 102, no. 3, pp. 2099–2116, 2017, doi: 10.1007/s11277-017-5044-z.
- [19] P. Rajyalakshmi, M. K. Kumar, G. U. Maheswari, and P. Naresh, "Implementation of clustering algorithms for real time large datasets," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 11, pp. 2303–2304, 2019, doi: 10.35940/ijitee.c2570.0981119.
- [20] A. Murugesan, B. Saminathan, F. Al-Turjman, and R. L. Kumar, "Analysis on homomorphic technique for data security in fog computing," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 9, 2020, doi: 10.1002/ett.3990.
- [21] L. K. Ramasamy, S. Kadry, Y. Nam, and M. N. Meqdad, "Performance analysis of sentiments in Twitter dataset using SVM models," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2275–2284, 2021, doi: 10.11591/ijece.v11i3.pp2275-2284.
- [22] R. Vijayalakshmi, V. Vasudevan, S. Kadry, and R. L. Kumar, "Optimization of makespan and resource utilization in the fog computing environment through task scheduling algorithm," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 18, no. 01, p. 1941025, 2020, doi: 10.1142/s021969131941025x.
- [23] M. A. Jayanthi, R. L. Kumar, A. Surendran, and K. Prathap, "Research contemplate on educational data mining," *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. IEEE, 2016, doi: 10.1109/icaca.2016.7887933.
- [24] J. Padikkapparambil, C. Ncube, F. Khan, L. K. Ramasamy, and Y. R. Gashu, "Novel stacking classification and prediction algorithm based ambient assisted living for elderly," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–19, 2022, doi: 10.1155/2022/5728880.
- [25] Kaggle, "Higgs Boson machine learning challenge | Kaggle," Accessed: Apr. 08, 2022. [Online]. Available: www.kaggle.com, 2014. <https://www.kaggle.com/competitions/higgs-boson/data>

BIOGRAPHIES OF AUTHORS



Vunnava Dinesh Babu    is working as a Research scholar in the Department of computer science and engineering at Saveetha school of engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India. He did his masters in 2014 at St. Marys group of institutions, Guntur, Andhra Pradesh in the stream of computer science. He is having total 8 years of teaching experience at undergraduate and postgraduate level. He can be contacted at dineshbabuvunnava@gmail.com.



Karunakaran Malathi    associated with Saveetha Institute of Medical and Technical Science as Associate Professor in the Department of Computer Science and Engineering. She did her doctorate in Saveetha Institute of Medical and Technical Sciences, Chennai. She has an experience of 14 years of teaching and 1.5 years of industry experience as a Database Administrator. She had presented many papers at various national and international conferences and published more than 60 journal papers in various International and National Scopus and SCI type of journal. She can be contacted at malathi.learning@gmail.com.