International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

# EFFICIENT BIG DATA CLUSTERING USING A HYBRID K-MEANS WITH CLUSTER MERGING STRATEGY

**Ashraf Abd El-Sattar[1], Mohamed Mostafa [2], Farid Ali Mousa [3]**

[1]IT Director, MSMEDA, Giza, Egypt

[2,3] Associate Professor of Information Technology, Faculty of Computers and Artificial Intelligence, Beni-Suef University

E-mail: [1] Ashraf.elgharably@msmeda.org.eg

## ABSTRACT

The use of big data is currently exploding across the board in academia and industry. As essential as the potential of this huge data is, new methods of thinking and training are needed to overcome many of the problems. This vast prospective of the data is undeniable, thus different methods of thinking and new learning methodologies are needed for solving these many issues. Clustering techniques are the most important issues to handle big data efficiently. Many clustering techniques are used to simplify the way we manipulate Big data. Despite the fact that k-means is one of the most often used clustering approaches, many researchers introduced modifications to increase its efficiency. Several frameworks are designed to facilitate the manipulation of big data, Map Reduce is one of the most useful frameworks used. In this paper, we present a modified k-means algorithm based on mergeing nearest clusters that increased its efficiency according to our simulation on benchmark data.

**Keywords*:*** Big Data, Clustering, K-Mean, Machine Learning, Map Reduce.

## 1. INTRODUCTION:

Big data nowadays is a major investment area for its businesses for the wide range of technological breakthroughs that are fueling the tremendous development in data and data collection[1]. There are now computer requirements for handling and processing such a massive number of data, in addition to effective data mining approaches, for processing enormous volumes of data[2]. Analysis of large datasets and the development of accurate descriptions or models for each is the primary problem in the area of big data mining. In the field of machine learning, theories and techniques of learning are examined in depth. In the first and second places. In a place where computer science, statistics, and machine learning all come together (ML) combine to extract hidden information from data in this age. Unsupervised learning is used to find clusters, and the resultant system reflects a data idea from the perspective of machine learning. [3] In addition to recommendation engines, recognition systems sciences, and data processing, it's been used in a variety of other areas as well [4]. Structured class based on the data's properties. Access and computation, data privacy, and domain knowledge are the three primary categories of big data difficulties.[5] Mining huge datasets poses extra computing challenges for clustering analysis[6] . The sheer amount of Big Data presents a slew of concerns, including those related to data access, data privacy, data management, data storage, data security, and data analysis. For this reason, clustering is the

Received: August 13, 2025 

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

unsupervised discovery of a secret data notion. It is the most interdisciplinary area, drawing on concepts from a wide range of disciplines, including optimization theory, computer science, scientific theory, engineering, statistics and arithmetic. Machine learning has had a positive influence on both science and society because of its widespread usage in a wide range of applications, including optimal management. The history of clustering is founded in mathematics, statistics, and numerical analysis, which is why data modelling is so important. The objective of Simplicity is achieved by representing data with fewer clusters, which inevitably results in some fine information being lost (similar to when data compression is lost) [7, 8]. The following is the structure of the document: Big data clustering is the subject of Section 2, which gives some background information. Experiments that were carried out are summarized in Sections 3 and 4. Section 5 is the last chapter of the paper.

## 2. Background

### 2.1 MapReduce-Based Clustering:

Big data solutions such as MapReduce may handle a large amount of data simultaneously using multiple low-end computers. Using the MapReduce framework, which conceals the complexities of parallel execution, enables users to concentrate only on data processing techniques. Mappers and reducers are the two essential components of the MapReduce concept. It is the goal of this programming paradigm to create mappers (or map functions) that may create a series of intermediary key/value pairs. The reducer (or reduce function) may be used as a shuffle or combination function to mix all the intermediate values associated with an analogous intermediate key. In order to do operations on multiple keys and lists of data simultaneously, each map and reduction may be performed independently of the other mappings and reductions in the process[9, 10].

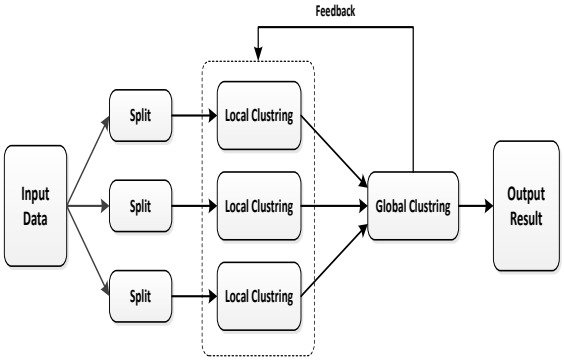


*Figure 1 : MapReduce clustering framework in general [11].*

### 2.2 K-Mean Clustering

The distance between data points and cluster centroids is typically calculated using a simple similarity measure in the K-means clustering algorithm, as the similarity is repeatedly calculated till the number of iterations or changes to the cluster centroids are no longer possible with this approach.[12]

Received: August 13, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

There are a set number of clusters and their initial centroids that are allocated to each data point using any of the distance measurements.[13]

Initializing random centroids using the k-means clustering method yields a varied sum of squared error (SSE), which assesses the quality of grouping, when choosing the starting centroids of the clusters.[14, 15]

**Algorithm 1: K-means clustering**

Inputs: S (feature vectors), K (number of clusters)

Clusters are the result of this process.

Set up the K initial cluster nodes.

assign instances to the nearest cluster center until the termination condition has been met while 2:

Use a city block measurement.

$$D_{i,j} = \sum_{l=1}^{d} |x_{il} - y_{jl}|$$

Based on the assignment, recalculate the cluster centers

Ending while loop

## 3. Related Work

There are recent and various surveys and taxonomies in this field, However, the majority of them concentrate on the Hadoop platform, while some are out-of-date or do not go into great length on utilizing Spark to cluster massive amounts of data.

Big data processing framework, Hadoop, was studied by Rotsnarani and Mrutyunjaya [16]. Hadoop map-reduce aspects are explored in order to address the issues of scalability and complexity when dealing with large amounts of data. Using a map-reduce model, Rujal and Dabhi [17] performed a survey on k-means.

In this article, the technical aspects of utilising Apache Hadoop to parallelize k-means are explored. For several applications of large data clustering, this study found that the K-means method is a feasible option and more popular than any other methodology. A review of the key difficulties in massive data processing using Hadoop map-reduce was carried out by[18]. Hadoop's primary drawback is network latency, according to this study.

An extensive assessment of the spark ecosystem for processing massive amounts of data was undertaken by[19]. Spark's architecture and programming paradigm are explained in this article. It was also reviewed the advantages and disadvantages for large-scale data processing when using the Spark platform, as well as different optimization strategies.

Received: August 13, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Based on Spark, k-means were developed by [20] writers. K-means is an entirely unsupervised learning method that group data without knowing how many clusters there are. Using map-reduce on top of the Spark platform, Sarazin, Lebbah, and Azzag developed a parallel implementation of biclustering [21]. Using the Spark framework, uses Particle Swarm Optimization and Cuckoo-search to improve the selection of k-means cluster centroid[22]. Clustering anomalies in large datasets may be done using Thakur and Dharavath [23] hybrid technique that combines k-means and decision trees. An method known as the decision tree algorithm may be used on each cluster to categorize examples of regular behavior from those that are considered anomalous.

This approach was built by [24] using fuzzy clustering under Spark. Instead of using Euclidian distance, this approach optimizes distance calculations using the L2 norm. Spark is utilized in another article to discover criminal trends in large-scale spatiotemporal datasets using the fuzzy clustering approach[24]. In [25] created a parallel approach for picture segmentation based on Fuzzy logic for processing large amounts of data in agriculture. To begin with, the photos were transformed among RGB and distributed to the cloud's available nodes. In the next step, the cluster centroids of each pixel point were determined.

For the Zika clustering, the scientists employed the Gaussian mixture model on spark MLlib by Lavanya, Sairabanu, and Jain [26]. In order to track the transmission of the virus during an epidemic, clusters were created. According to Chitrakar and Petrovic [27]. By omitting numerous point-centre distance calculations, the modified k-means approach is expected to speed up the analysis of high-dimensional data. Consequently, it may be useful for clustering.

K-means with triangle inequality was used by Fatta & Al Ghamdi [28] to shorten search time and eliminate duplicate computation. The efficiency of k-means may be considerably increased utilizing triangle inequality improvements, according to the authors. Zhang et al. provide a distributed possibility c-means method [29]. Instead of allocating each input point to a single cluster, probabilistic c means are different from previous k-means algorithms in that they assign probability membership values to each cluster for each input point.

Fast parallel DBSCAN utilizing Spark was reported by [30] in order to avoid the shuffle procedure in another work. The partial clusters are computed on a per-executor basis. The merging procedure is postponed until the driver has received all of the component clusters. Distributed density based fuzzy clustering is proposed by [31] to locate genes with comparable expression levels.

For stream data processing, The author in [32] developed a clustering algorithm based on swarm intelligence. As a first stage, we use fuzzy c-means to generate the first cluster centres, then we use adaptive particle swarm optimization to refine the clusters.

## 4. Proposed System

The map function is used to split the data sets into numerous mappers, as illustrated in Figure 2. Samples in each mapper are the same. Each mapper's output is combined to get the

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

final result in the reduction step. Modeling the K-Mean technique by merging the clusters using an unique equation that gives each cluster a value according to its support, and then removing clusters with values below the threshold and merging the samples of those deleted from the adjacent cluster, is the concept behind this model. For testing data, the mean value is used to determine the final cluster, which is determined by the reducer.

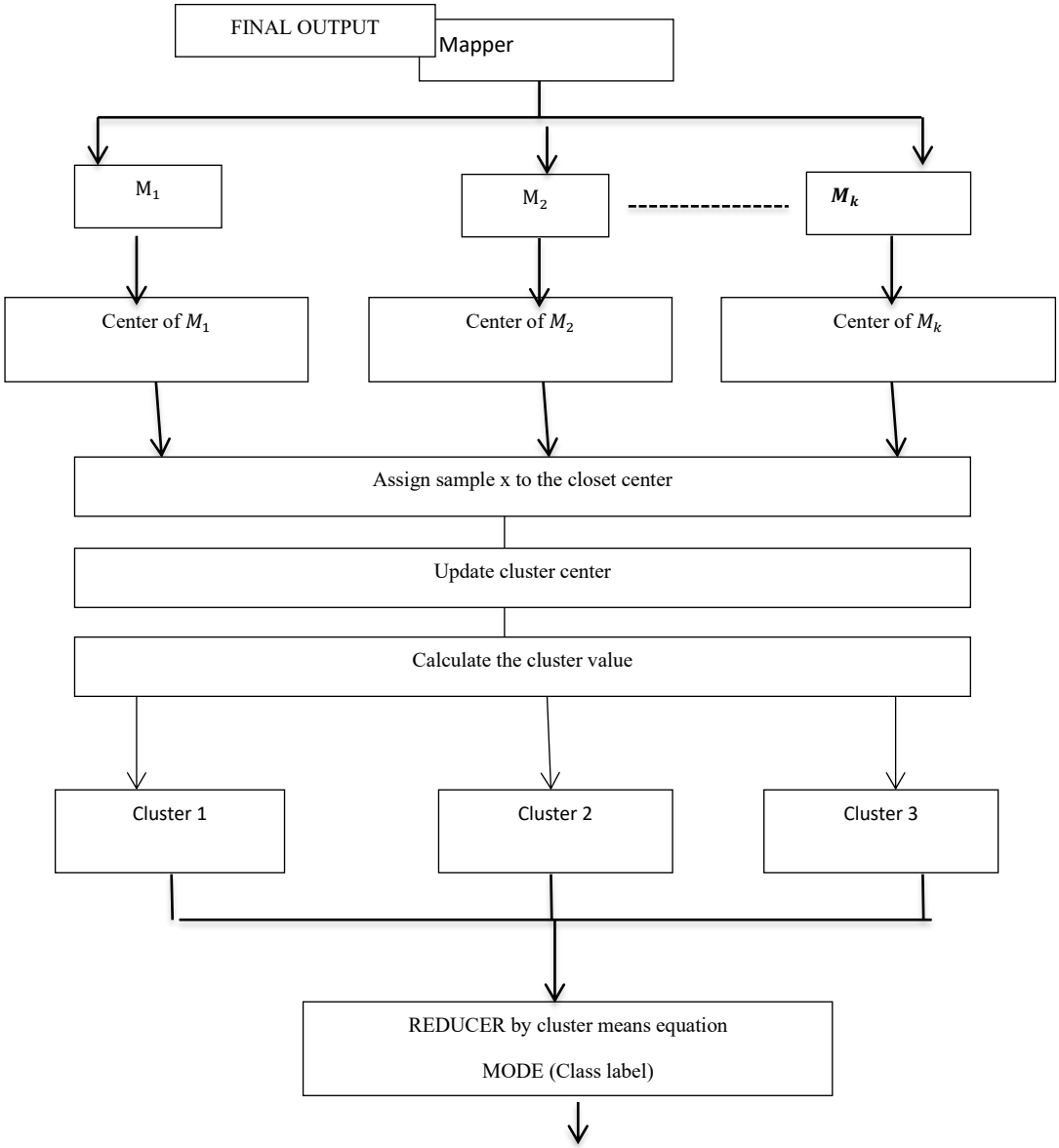


*Figure 2: Proposed Model*

### 4.1 Mapper Phase

Received: August 13, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

***Algorithm 2 :*** **Modified K-means Clustering algorithm on BIG Data –** ***Map Phase***

*A dataset with n data points is the input.*

*It produces a list of the key and the value of all data points that fall inside each cluster's centre.*

*The Training data is available.*

*(Train Data=datastore("My Data.csv","DatastoreType","tabulartext"),*

*K-Mean Map Design with Modified Procedures*

*Open the Train Data file and separate the data into training and testing.*

*In step 2, create a new matrix containing the testing data set*

*My Test Data = csvread (Test Matrix)*

*Step 3, Samples should be read sequentially while the file isn't near the end of its lifespan.*

*The fourth step is to read one mapper at a time from Train Data.*

*K-Mean is called in step 5.*

*1. Map function's input data block is in the form of points.*

*2. Use the Gap Statistics clustering assessment to determine the best number of clusters for each data piece.*

*The k-cluster centres must be established*
*Use distance measurements to assign sample x to the nearest point in order to satisfy a termination condition that has not been met.*
*Refresh the cluster centre to reflect the new example cluster assignments*
*For each cluster calculate the distance between it and all clusters [generate distance matrix]*

Received: August 13, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

*Calculate the cluster value*

$$Cluster_{Value}(i) = \frac{\sum_{i=1}^{n} Cluster\ distances}{Max_{distance} - Min_{distance}}$$

*If* $Cluster_{Value} < Threshold$
*Delete Cluster (i)*
*No_of_Clusters = No_of_Clusters-1*
*End While*

*The map function finds the nearest center among optimal k centers for the input point.*

*Step 6, Send Map_Results to the functionDesign (My_Reducer)*

*Step 7: adding (intermediateValuesOut,Map_Result);*

*Loop End*

*Loop End*

*call My_Reducer*

*procedure end*

**4.2 Reduce Phase**

***Algorithm :*** **Modified K-means Clustering algorithm on BIG Data –** ***Reducer Phase***

A mapper's output, where key k is the centre and value is a collection of all data points closest to that centre, is used as an input.

Output: A collection of clusters, each containing a global centre and individual cluster data points as values.

A set of data points is used in the reduction step to arrive at a new centre point.

- Mapper results are merged using the centre of the map as the key and data points are grouped to generate clusters.

Received: August 13, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

For each and every one of these groups,

- Add up all the data in each cluster.

- Each cluster mean equation has a central average.

- A cluster mean as a cluster id with all of the data points in that cluster is generated by the reducer.

Using the Map Reduce architecture for the Modified K-Mean method, we can see that Map and Reduce are two separate processes. Separate 'key-value' pairs are used for each step. After distributing the training data among the mappers, each function in the Mapper function accepts the testing data as well. For training, we utilize the novel equation for each cluster and eliminate the cluster with a value lower than the threshold in the Map procedure. After then, the sample's cluster is determined by the reducer function, which makes use of the mean value. [33]A wide range of distance measurements were proposed;

1. **City block distance**

If you follow a grid-like path, city block distance estimates the distance from one place to the next. It is the total of the two distances, Manhattan and block distance that separates the two objects. In n-dimensional vector space, 'q' and 'r' are determined using a vector-based technique called city block distance. The summing of the boundaries determines the distance L1 or block. The gap between the city block and the sidewalk may be seen below.

$$\boldsymbol{L_1(q,r) = \sum_y |q(y) - r(y)|}$$

There are basically two dimensions in the grids, and the distance between locations is the number of edges that must be traversed in order to go from 'q' to 'r'.

2. **Cosine similarity**

An interior product area's cosine similarity measure compares the cosine of the angle between two vectors of the interior product area. When Cosine's similarity metric is used to calculate the normalising point product of two texts, the attribute will be utilised as a vector. Consumers try to determine the cosine of the angle between two things by comparing their cosine similarity. No characteristics are exchanged since the sample is at an angle of 90 degrees between the two objects, which results in the value of zero. It is possible to describe this in mathematical terms as follows:

$$\boldsymbol{cos\ (\theta) = \frac{X.Y}{|X| * |Y|}}$$

3. **Euclidean distance**

Using a ruler, you may measure the Euclidean distance between two points. A graph may be used to determine similarity since the axes on a graph are common properties for data. The resultant map's data pieces are said to be diagrammed in the preferred space.

Received: August 13, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Mathematically, it is possible to determine the relationship between two items.

$$\boldsymbol{E\ (X,Y) = \sqrt{\sum_{i=0}^{n}(\ X_i - \ Y_i)^2}}$$

**4. Jaccard Coefficient**

Scale and amounts of overlap are used to assess the similarity between two sets. The cardinality of its intersections and the cardinality of its union are used to define it for two sets. Calculated as follows:

$$\boldsymbol{J(A,B) = \frac{|A \cap B|}{|A \cup B|}}$$

**5. Minkowski Distance**

To show how far Minkowski is from Euclidean, have a look at this illustration.

$$\boldsymbol{dist = (\sum_{k=1}^{n} |p_k - q_k|^r)^{\frac{1}{r}}}$$

Data objects P and Q are represented in the data set by n and r, respectively, and k is the number of components.

**6. Mahalanobis Distance**

These weights are based on the correlation between the variables. The Euclidean distance may be obtained if the covariance is zero and the variables are uniform. The distance between the two Mahalanobis is shown in the diagram below:

$$\boldsymbol{dist(p,q) = (p-q)\sum^{-1}(p-q)^T}$$

Such that, $\sum^{-1}$ represents the matrix of the covariance among the variables.

Classifiers might be evaluated using a variety of measures. The Positive Predictive Index is the first statistic.

In order to determine how "accurate" your classifier is in predicting the class label of feature vectors, you need to evaluate the measurements. The diagnostic test is the most important part of assessment. Specificity, precision, and accuracy (also known as recognition rate), are some of the classifier assessment metrics. The following formulas are used to compute these metrics [34]:

$$\text{Accuracy } = \frac{(TP + TN)}{(P + N)}$$

$$Error\ Rate = \frac{FP+FN}{P+N} \qquad -$$

Received: August 13, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$Sensitivity\ (Recall) = \frac{TP}{P} \quad \ldots$$

$$Specificity = \frac{TN}{N}$$

$$Precision = \frac{TP}{TP+FP}$$

The number of positive feature vectors is P, while the number of negative feature vectors is N, given two classes, for example, the positive feature vector indicates that the teeth are diseased, while the negative feature vector indicates that the teeth are not infected.. This is done for each feature vector by comparing the classifier's predicted class label to the feature vector's actual class label. In order to compute several assessment measures, we must understand four more words known as "building blocks".

TP are the positive feature vectors that the classifier properly identifies as positive. This is the amount of positives that you can count on. While TN are the negative feature vectors that the classifier properly identifies as negative. Let TN be the total number of negatives. Positive feature vectors that were mistakenly classified as negative (e.g., feature vectors of class teeth infected = no that the classifier predicted teeth infected = yes) are referred to as false positives. The number of false positives is FP. [34]

For example, the classifier predicted teeth infected = no for vectors of class teeth infected = yes, resulting in false negative (FN) results. The number of false negatives is shown as FN. [34]

The accuracy of a classifier on a given test set is the percentage of test set feature vectors that are correctly classified by the classifier. Error rate is the misclassification rate of a classifier. Sensitivity and specificity measures are used in accessing how well the classifier can recognize the positive feature vectors and how well it can recognize the negative feature vectors respectively. The *precision* and *recall* measures are also widely used in classification. Precision can be thought of as a measure of *exactness* (i.e., what percentage of feature vectors labeled as positive are actually such), whereas recall is a measure of *completeness* (what percentage of positive feature vectors are labeled as such). The proportion of successfully categorized feature vectors in a classifier's test set indicates the classifier's accuracy. The rate at which a classifier makes a mistake is known as the error rate. A classifier's sensitivity and specificity are measured in terms of how effectively it can identify positive and negative feature vectors, respectively. Classification also uses metrics of accuracy and recall. According to this definition, precision may be defined as the degree to which feature vectors labelled as positive are truly true. Conversely, recall can be defined as the degree to which feature vectors labelled as negative are actually true (what percentage of positive feature vectors are labelled as such). You may recognize the term recall since it is similar to sensitivity (or the true positive rate). [34]

Received: August 13, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Our study evaluates the classifier's accuracy as seen in the below Equation by testing it on the two classes of data. The accuracy of the classifier used to train and test the data for the four classes is measured as follows [35]:

$$\text{Accuracy} = \frac{No.of\ correct\ results}{No.of\ tests}$$

## 5. Experiment Results

In this part, we go into depth about the tests we ran and the data sets we utilized to get them.

### 5.1 Datasets

In our experiments, we make use of the following two types of datasets:

- Covtype

  Lodgepole Pine, Spruce/Fir, Ponderosa Pine, Aspen, Cottonwood/Willow, Douglas-fir, and Krummholz are among 581012 items in this collection.

  To test our hypotheses, we utilized the data in two different ways: with two classifications and with seven. 54 different properties may be assigned to one object. [35] Provides a more in-depth analysis of this collection of data.

- Poker Hand.

  Poker Hand Dataset is provided as the second dataset utilized in this research. The following is a list of all 1025010 items and 10 ordinal classes (0-9) in the dataset:

  0: No cards are held; no poker hand is recognized.

  1: A pair; a pair of cards with the same rank within a span of five cards

  2: Among the five cards, there are two pairings of equal among the five cards, there are two pairings of equal rank.

  3: Three of a kind; three cards of the same rank in the same suit

  4. Straight; no pauses in consecutively ordered five-card sequence

  5: Five cards of the same suit are called a "flush."

  6: A pair and a rank apart. a set of three matching cards

  7: A four-of-a-kind hand; a hand having four cards of the same rank.

  8: Straight flush; straight + flush

  9: Royal flush consists of the aces, king, queen, jack, and ten plus a flush.

  For additional information on the gathering and explanation of this data set, see [36].

  Each object has eleven properties.

  First, we utilize the data with two classes and then we use it with ten classes.

### 5.2 Results

*Table1. Proposed Model Accuracy and time taken*

**International Journal of Applied Mathematics**
**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| Dataset | Basic K-Means | | Proposed Method | |
|---|---|---|---|---|
| | Accuracy | Time Taken | Accuracy | Time Taken |
| CovType | 56.72 % | 840.38 S | 67.3 % | 925.16 S |
| CovType 2 | 62.1 % | 784.86 S | 79.6 % | 650.06 S |
| Poker | 62.67 % | 701.66 S | 78.1 % | 850.11 S |
| Poker2 | 63.2 % | 725.34 S | 81.2 % | 700.31 S |

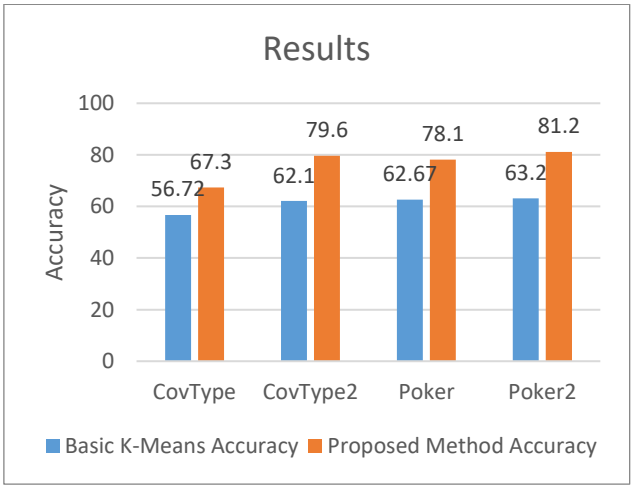


*Figure 3. Experiment Results*

In both data sets, the suggested model outperformed the simple K-mean approach, as can be seen from the results shown above. In the CovType data set the modified k mean algorithn gave us 76.3 % accuracy while the basic k-mean algorithm gave us 56.72 %, In CovType 2 data set the proposed model gave us 79.6 % and the basic k-mean algorithm gave us 62.1. In the Poker data set the proposed model gave us 78.1 % while the k-mean gave us 62.67%; finally the Poker2 data set gave us 81.2 % while k-mean gave us 63.2%; Also we can notice that the time taken for the modified k-mean is nearly from the time taken for K-mean algorithm.

## CONCLUSION:

In order to have a better understanding of machine learning and its use in large data processing, this article examines the fundamentals and current research in this topic.

Received: August 13, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Supervised, unsupervised, semi-supervised, reinforcement, multi-tasking, ensemble, neural network, and instance-based methods of machine learning have all been classified and described in further detail. There are also some more sophisticated ways of learning shown. Representation learning, deep learning, transfer learning, distributed and parallel learning, kernel-based learning and active learning are all examples of this. -based methods of instruction.

In this study we introduce a new clustering method by modification of K-mean clustering algorithm with Map-Reduce paradigm to cluster big data. The proposed method is preferred when number of clusters is large because it used by merging the nearest clusters. Experiments conducted on two benchmark data sets reveal that the proposed fusion technique results in a considerable improvement over the k-mean clustering algorithm.

## References

1. Wiener, M., C. Saunders, and M. Marabelli, Big-data business models: A critical literature review and multiperspective research framework. Journal of Information Technology, 2020. 35(1): p. 66-91.
2. Mohamed, A., et al., The state of the art and taxonomy of big data analytics: view from new big data framework. Artificial Intelligence Review, 2020. 53(2): p. 989-1037.
3. Ang, K.L.-M. and J.K.P. Seng, Big data and machine learning with hyperspectral information in agriculture. IEEE Access, 2021. 9: p. 36699-36718.
4. Naeem, M., et al., Trends and future perspective challenges in big data, in Advances in Intelligent Data Analysis and Applications. 2022, Springer. p. 309-325.
5. Yousefi, S., F. Derakhshan, and H. Karimipour, Applications of big data analytics and machine learning in the internet of things, in Handbook of big data privacy. 2020, Springer. p. 77-108.
6. Hass, G., P. Simon, and R. Kashef. Business applications for current developments in big data clustering: an overview. in 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). 2020. IEEE.
7. Aliguliyev, Ramiz, and Tural Badalov. "EXPLORING BIG DATA CLUSTERING: APPROACHES, ALGORITHMS, AND PLATFORMS." *Reliability: Theory & Applications* 20.SI 7 (83) (2025): 177-187.
8. Gan, G., C. Ma, and J. Wu, Data clustering: theory, algorithms, and applications. 2020: SIAM.
9. Mao, Y., et al., A MapReduce-based K-means clustering algorithm. The Journal of Supercomputing, 2021: p. 1-22.
10. Jadhav, Abhijit Dnyaneshwar, et al. "Distributed two phase intrusion detection system using machine learning techniques and underlying big data storage and processing architecture-HDFS." Inteligencia Artificial 28.76 (2025): 124-148.
11. Chen, M., S.A. Ludwig, and K. Li, 16 Clustering in Big Data.
12. Jacksi, K., et al. Clustering documents based on semantic similarity using HAC and K-mean algorithms. in 2020 International Conference on Advanced Science and Engineering (ICOASE). 2020. IEEE.
13. Nasrullah, Harun, Arif Bramantoro, and Ahmad A. Alzahrani. "K-Means Optimization with Kernel Gaussian Radial Basis Function in C." *Contemporary Mathematics* (2025): 6754-6779.

Received: August 13, 2025 

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

14. Ahmed, S.R.A., et al., Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set. Periodicals of Engineering and Natural Sciences, 2019. 7(2): p. 448-457.

15. Mongkolluksamee, Sophon, and Subhorn Khonthapagdee. "Privacy-Preserving Breast Density Classification in Mammograms Using Fuzzy C-Means and Homomorphic Encryption." *2025 17th International Conference on Knowledge and Smart Technology (KST)*. IEEE, 2025..

16. Zamzami, Moh Rifqi, Moh Riswandha Imawan, and Imam Ghozali. "A Comparative Study On Hadoop Ecosystem: Hive And HBase–A Literature Review." *JSSTEK-Jurnal Studi Sains dan Teknik* 2, no. 1 (2024): 97-112.

17. Ji, Chunmei, Jing Zhou, and Fengbo Kong. "K-means Clustering with AES in Hadoop MapReduce." *Informatica* 48, no. 20 (2024): 81-92.

18. Sanghi, A., et al., HYDRA: a dynamic big data regenerator. Proceedings of the VLDB Endowment, 2018. 11(12): p. 1974-1977.

19. Saeed, M.M., Z. Al Aghbari, and M. Alsharidah, Big data clustering techniques based on spark: a literature review. PeerJ Computer Science, 2020. 6: p. e321.

20. Shan, J., L. Jingguo, and L. Xianhu, 3D Point Cloud Clustering Based on Improved Ant Colony Algorithm. Bulletin of Surveying and Mapping, 2018(3): p. 38.

21. Forest, F., et al. A generic and scalable pipeline for large-scale analytics of continuous aircraft engine data. in 2018 ieee international conference on big data (big data). 2018. IEEE.

22. Liu, Z., et al., A hybrid genetic-particle swarm algorithm based on multilevel neighbourhood structure for flexible job shop scheduling problem. Computers & Operations Research, 2021. 135: p. 105431.

23. Thakur, S. and R. Dharavath, KMDT: a hybrid cluster approach for anomaly detection using big data, in Information and Decision Sciences. 2018, Springer. p. 169-176.

24. Win, K.N., et al., PCPD: a parallel crime pattern discovery system for large-scale spatiotemporal data based on fuzzy clustering. International Journal of Fuzzy Systems, 2019. 21(6): p. 1961-1974.

25. Chaudhary, Yashi, and Heman Pathak. "Role of Machine Learning for Big Data." *International Conference on Smart Systems and Advanced Computing (SysCom 2022)*. Springer Nature, 2025.

26. Huy, Nguyen Quang, Vu Thu Diep, and Phan Duy Hung. "Apache Spark Implementation of the Constrained K-Means Clustering Algorithm." *International Conference on Computational Science and Its Applications*. Cham: Springer Nature Switzerland, 2025.

27. Shrestha Chitrakar, A. and S. Petrović. Efficient k-means using triangle inequality on spark for cyber security analytics. in Proceedings of the ACM international workshop on security and privacy analytics. 2019.

28. Al Ghamdi, S. and G. Di Fatta, Efficient clustering techniques on hadoop and spark. International Journal of Big Data Intelligence, 2019. 6(3/4): p. 269-290.

29. Zhang, S., et al., A cooperative spectrum sensing method based on information geometry and fuzzy c-means clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 2019. 2019(1): p. 1-12.

Received: August 13, 2025 

**Volume 38 No. 12s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

30. Arriesgado, Alimuddin, and Marc Caesar Talampas. "GeoSeg: Path Loss Estimation Method for Outdoor LoRa LPWAN." *2025 International Symposium on Multimedia and Communications Technology (ISMAC)*. IEEE, 2025.

31. Nguyen, Danh-Nguyen, Thi-Thuy Mac, and Hong-Hai Hoang. "Supply chain collaboration and green supply chain management: A bibliometric review, 2000–2023." *Journal of Transport and Supply Chain Management* 19 (2025): 1170.

32. Desanamukula, V.S. and K.N. Rao. Dealing Big Data using Improved Fuzzy C Means Based Improved Redundant Particle Swarm Optimization with Map Reduction. in 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM). 2022. IEEE.

33. Mousa, F. and T. Mohamed, Detection of COVID-19 from chest CT images using selected fourier transform features. Journal of Theoretical and Applied Information Technology, 2021: p. 2515-2524.

34. Mahmoud, A., et al., TGT: A Novel Adversarial Guided Oversampling Technique for Handling Imbalanced Datasets. Egyptian Informatics Journal, 2021. 22(4): p. 433-438.

35. Junaid, KA Mohamed, D. Paulraj, and T. Sethukarasi. "A comprehensive ensemble classification techniques detecting and managing concept drift in dynamic imbalanced data streams." *Wireless Networks* 31.1 (2025): 19-30..

36. Cheng, Fang. "A comparative study of the performance of Spark-based k-means algorithm based on Euclidean distance and Manhattan distance." *2024 3rd International Conference on Big Data, Information and Computer Network (BDICN)*. IEEE, 2024.

Received: August 13, 2025 