Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

A COMPREHENSIVE STUDY OF K-MEANS AND K-NEAREST NEIGHBORS ALGORITHMS IN BIG DATA CONTEXT

Khalid Mahmood ,Ghulam Irtaza, Muhammad Ahsan Raza

[1]Department of Artificial Intelligence,NFC Institute of Engineering and Technology, Multan, Pakistan
[2] Department of Information Sciences, University of Education, Lahore, 54000, Pakistan

[3] Department of Information SciencesUniversity of Education, Lahore, Multan campus 60000, Pakistan

DOI: https://doi.org/

Keywords
Big Data, K-Means, KNN, Clustering, Classification, Machine Learning

Article History
Received on 20 Oct 2025
Accepted on 20 Nov 2025
Published on 27 Dec 2025

Corresponding Author: *
Muhammad Ahsan Raza

Abstract
In the present era of information processing, a huge amount of data is being produced by businesses, the web and social media, government and non-government organizations, and other sources in real time. The analysis of this huge collection of data, popularly known as Big Data, is still a challenging task to produce more accurate and relevant results in an information retrieval system, especially when the velocity is high. Efficient and scalable data processing techniques and platforms are still needed to handle the data efficiently and to retrieve the relevant information more accurately from this large collection of data available in structured and unstructured formats in real-time (such as the data generated in situations of a pandemic like COVID-19). Various machine learning platforms are established, each with its unique features to manipulate the Big Data concerning volume, velocity, and variety of data. The common concern of these platforms is the computational speed and the ability to manipulate the amount of data. There is a need to have a comprehensive analysis of machine learning platforms to facilitate new researchers and application developers to develop new applications to process the data efficiently in a big data environment. K-Means clustering and K-Nearest Neighbors (KNN) classification are studied on some specific datasets. Effective analysis of these algorithms is performed in this research in terms of seven features, including: major contribution, performance measures, effectiveness measures, big data environment, parallel & distributed processing, tools used and evaluation metrics. The results are produced very carefully by thoroughly analyzing the K-Means and KNN schemes. Overall, this study provides an insight into recent trends in the big data environment using K-Means and KNN algorithms.

INTRODUCTION

During the phase of pandemics like COVID-19, Conjunctivitis, and others, social distancing becomes necessary, and businesses switch to online mode from physical interaction. The volume of the data increases with high velocity, and this data is generated through various sources. The data is collected at a large scale, popularly known as Big Data. The processing of such data for analysis of the recent advances in Big Data has been increasingly popular in recent years and a hot area of research in the present age of information processing. Every year, several businesses, such as marketing, medical sciences, image processing, web media, social computing, manufacturing industries, and public sector organizations acquire massive volumes of data. According to the International Data Corporation (IDC), the total produced and copied data grew by roughly nine times in 2011 to 1.8 zettabytes (10^{21}) [1]. This data explosion has also been aided by the cheaper and more sophisticated data storage medium. Efficient and effective technologies, methodologies, and systems are always required to efficiently handle and retrieve information from such a huge abundance of digital data or Big Data. For instance, algorithms for the retrieval of information from Big Data are also becoming increasingly important, specifically during the COVID-19 pandemic. These algorithms are very helpful in the process of decision-making in real-time during pandemics, with the use of these advanced technologies and methods proposed for the fast processing of data.

The traditional and typical data manipulation technologies have been unable to keep up with the development of big data. Researchers all across the world are working harder than ever to make proper hardware and software systems. There are a variety of Big Data processing systems available today, each with its own set of features. These systems process data volume, velocity, and variety in different ways.

The "proper platform" is determined by the task at hand, and selecting it necessitates extensive expertise (Agneeswaran, 2013). The two most popular concerns with the processing systems are:

i. How speedily can the data be processed?
ii. How much volume of data can be processed?

Horizontal scaling and vertical scaling (Singh, 2015) are the two main types of parallel and distributed systems. The task is split between several machines or nodes in horizontal scaling. Scale-out is another name for this approach. Apache Hadoop, Spark, Peer-to-Peer Networks, and others are examples of horizontal scalability platforms. The main benefit of horizontal scaling is that in the big data model, we may use our ordinary commodity equipment. The size and scope of the system can be adjusted as far as we need to, and therefore, the cost of the system in financial terms can be comparatively low. In this circumstance, the software is a huge disadvantage because it must manage a distributed scenario, and programming takes specialized knowledge. The number of software solutions accessible is also limited. Vertical scaling, on the other hand, includes adding multiple processors and more memory into a single computer, GPU (Graphical Processing Unit), FPGA (Field Programmable Gate Arrays), HPC (High-Performance Cluster), and CPU (Central Processing Unit) with multiple cores, these are known as scale-up. More hardware may be managed and installed in a single computer with ease. The system must be powerful enough to accept the additional hardware, which usually requires significant investments. However, after a certain point, it is not viable to add further hardware to vertical scaling.

A comparative study is still not available that provides a comprehensive analysis or comparison of different known systems. This research comparatively evaluates the performance of two data mining algorithms, namely K-Means and K-Nearest Neighbors (KNN), from machine learning and artificial intelligence disciplines. Cloud computing-based AWS (Amazon Web Service) EC2 is used for the scientific study of these Big Data processing environments, which will result in ensuring a comparable cost per hour required to cooperate with two algorithms on each system. A major constraint is that neither of the systems, like AWS, allows us to compare and analyze these systems efficiently. The research presented in this work seeks to determine which machine learning platform will be the most efficient for conducting big data calculations and parallel and distributed computation. We collate these platforms utilizing two well-known data mining methods, K-Means algorithms for clustering of data and the KNN algorithm for the classification of data. Various insights are obtained into the big data environments on which these algorithms are implemented, and a thorough comparison of each machine learning system is conducted based on the feature set. Because many additional data mining algorithms employ these two techniques (i.e., K-Means and KNN) during pre-processing or as a core component, we anticipate that our findings will have an impact on a wide range of applications and algorithms beyond those described in this research article. Conclusively, the major contributions of this research are as follows:

To describe the research based on K-Means & KNN algorithms on various parallel & distributed processing systems in a Big Data environment.

To provide a comparative analysis of these processing systems in terms of running time.

The remaining sections of the dissertation are organized as follows. In the coming section, the background of the study is discussed. In the next section, we go through the literature review of the study. The analysis and discussion about the chosen algorithms are described in the second-to-last section of this paper. The last section of this literature concludes this research work and provides directions for future work.

Study Background

In the current age of information processing, humans and machines are using different media like the web and others through the use of different technologies and generating large-scale data, which is increasing day by day. The data has been collected in huge volumes and can be measured in terms of zettabytes (1021). In (Emani, 2015), it is stated that the data size on Internet storage will exceed the total capacity of the brains of all the creatures who are living in this world by 2025. Another research (Zhanga, 2018) explored that the NSA (National Security Agency) has reported that data of almost 1.8 petabytes in size is being generated on the Internet in a day. The current development and data processing challenge is obviously due to this fast and continuous increase in data size in this era of digital information. In the collection and creation of this large amount of data, different technologies and devices like sensors, storage devices, computational repositories, and communication channels are involved. Some other giant sources, including public and private businesses or organizations, profit-oriented and non-profit companies, scientific research, and industries, are playing a major role in the creation and dissemination of data at such a large scale (Agrawal, 2014). Also reported in (Dykes, 2017) that approximately 2.5 billion gigabytes of mixed

data, both structured (10%) and unstructured (90%), is produced in 24 hours of each day from different sources.

Another research (Rialti, 2019) specifies that an approximate increase in the data volume throughout the globe is up to 44 zettabytes, starting from 4.4 zettabytes from 2013 to 2020. This huge volume of data consists of images, animations, videos, audio, and text in the form of structured, semi-structured, and unstructured data produced from various sources, including social media, machines inter-communication, cyber systems, sensors, and IoT (Internet of Things). The term 'Big Data' is used popularly for such large-scale data. Different organizations including industries, businesses, governments, scientific disciplines, and social cultures, are facing big challenges of change occurring due to Big Data. The processing of Big Data is one of the major trends in all disciplines. In almost all disciplines of research, Big Data is very effective because its raw material has a great influence in this regard (Braganza, 2017). A lot of potential is here in Big Data for the improved and optimized utilization of resources. The extracted information is very helpful for the facilitation of the decision-making process (Gantz, 2011), (Rahmati, 2016).

Problem Statement

A large amount of data, called Big Data, is created or produced daily by the utilization of various online and offline resources within organisations in the current age of information. The processing of such huge amounts of information requires distributed systems and optimized algorithms for processing. Centralized or traditional data processing systems cannot perform efficiently and are not fit for the manipulation of Big Data. Processing precise data efficiently is a creative methodology, but it turns into a further challenging assignment when the volume of the dataset is as huge in size as the size of Big Data. Across the world, various researchers are continuously trying their best to develop systems either at the software or hardware level to fix the issues of Big Data processing complexities to make that information more useful.

A variety of distributed systems for Big Data processing is available. Each system comprises its unique features and functionalities. There is still a need to identify the problems with existing systems and to introduce a comparative performance evaluation of such systems by running some specific machine learning algorithms. This analysis is much more helpful for future researchers to propose new models of distributed Big Data processing.

LITERATURE REVIEW

Data clustering is one of the widely used procedures adopted in different research works and applications for various purposes nowadays. A hot research area in this discipline is text document clustering, belonging to the semantic information retrieval domain, using AI (Artificial Intelligence) and data mining techniques. Many research works have been introduced, but the activities involved have a large range that is impossible to cover; that's why it needs more attention to overcome the problems. Different widely used algorithms like K-Means (Dykes, 2017) and KNN (K Nearest Neighbor) have been introduced in the field of data mining for the grouping of similar data items. In the first algorithm, the similarity in data elements is measured (close to the centroid),

while in the second algorithm, the most similar items are selected near each other as neighbors (Rialti, 2019).

In cluster formation of data (specifically text data), K-Means is one of the most popular techniques in the field of data mining (Rahmati, 2016). The best cluster centroid is selected, which is the closest to the data items in a cluster, to form an accurate cluster. This algorithm is also considered to perform efficiently because the minimum runtime has been measured in various research experiments as compared to other clustering techniques.

A new model for the clustering of textual data has been proposed by (Tom, 2017), in which SOM (Self-Organizing Map) and Naïve Bayes algorithms are used together. The shown improvements in this model are based on the centroid's selection and the initial cluster count. It has also improved the decision-making process of selecting documents having the same distance to the centroids of two different clusters. The research also focuses on the performance in terms of optimal computational cost.

PSO (Particle Swarm Optimization) algorithms are used together with K-means algorithms (Maltby, 2011) for grouping text documents into relevant clusters. A comparative analysis is provided in this research by the formation of clusters of words extracted from related text documents. The results in this study show that hybrid PSO is performing much better than the most well-known K-Means clustering algorithms that test bed designed for experimentation. An algorithm known as GA (Genetic Algorithms) is used along with K-Means to increase its performance to form clusters of similar words. The scheme introduced in (Almeida, 2018) presented that GA can improve the accuracy of clusters as well as reduce the computational cost.

The authors of (Al Nuaimi, 2015) proposed a comparative study of SOM and K-Means algorithms by testing their performance on text documents clustering. They showed in the results that SOM is performing much better than K-Means in document clustering because the latter algorithm is much more sensitive to initial values. Another related research (Benedetti, 2019) also provides a comparison of fuzzy K-Means and ACO (Ant Colony Optimization) algorithms on text document clustering.

Another scheme has been introduced (Hurwitz, 2013) that clusters the documents based on DE (Differential Evolution) and GA algorithms. Mutation and crossover operators are used to optimize the cluster number. In the evaluation part, results show that a hybrid approach is more efficient than the DE and GA algorithms.

The research work proposed by (Hu. H, 2014) performs FWMS (Frequent Word Meaning Sequence-based clustering. In this work, attention is paid to the closeness of words and its meaning. K-Medoid and KNN algorithms are analyzed for text document clustering (Gantz, 2012). Three machine learning algorithms, including Naïve Bayes, C4.5 Decision Tree, and Back Propagated Neural Network, are experimented with to cluster the data of breast cancer (Bellaachia, 2003). In the results after comparison, it is presented that the Naïve Bayes algorithm is performing better than others in terms of computation time.

Artificial Neural Network (Chi, 2007) has been applied for the analysis of survival rates on a couple of datasets regarding breast cancer to cluster the patients in terms of good and bad prognosis. In this literature, it is studied that the algorithm is providing more accurate groups of patients on defined parameters, but the running cost of the algorithms has been skipped in the article.

The research work presented (Bucin'skil, 2007) focuses on the comparative study of some modern data processing schemes for the prediction of disease recurrence. Algorithms used for this purpose are PCA (Principal Component Analysis) and ANN (Artificial Neural Network). The analysis describes that PCA performs much better for the retrieval of information if the data is in bulk.

Data mining algorithms for the formation of clusters are applied to the same dataset of breast cancer to comparatively study the performance of these algorithms (Joshi, 2014). This research work performed the experiments by the implementation of K-Means, FF (Farthest First), HCM (Hierarchical Clustering Method), and EM (Expectation Maximization) clustering algorithms by using various data mining tools like WEKA, Orange, Tavera, and Rapid Miner to get more optimal results. The results section of this article discusses that both K-Means and FF are providing better performance compared to HCM and EM. An enhanced scheme based on KNN and Naïve Bayes algorithms (Saleema, 2014) for clustering the data items available in the SEER cancer dataset was proposed to compute the rate of survival and care quality. The experimentation was performed by using the MATLAB tool. The model provides good results as the sample is increased.

The utilization of machine learning methodologies, especially for clustering of data, including Naïve Bayes, RT (Random Tree), and SVM (Support Vector Machine), has been adapted in this research by the authors for class label prediction (Pandey, 2014). The research is done to identify the students who need advisory and counselling to achieve high-quality education targets. The research provides results in terms of accuracy, but there is a lack of studies about the performance of processing systems where algorithms were implemented.

In this research (Ayat, 2016), the DSS (Decision Support System) based on PNN (Probabilistic Neural Network) was used to group the patients of breast cancer. The performance measures were sensitivity, specificity, and accuracy. A modified version of KNN has been proposed in the research (Parvin, 2008) to group the data concerning the accuracy of terms, but the proposed scheme performs slowly for large input sizes.

A scheme named IBK (Instance-Based KNN) is proposed for the testing of this algorithm on three different breast cancer datasets (Salama, 2012). The performance of the algorithm is also compared with some other related algorithms like MLP (Multi Layered Perception), SMO (Sequential Minimal Optimization), Naïve Bayes, and Decision Tree.

Performance analysis of the KNN algorithm based on accuracy in classification and processing time has been provided in the research work (Madjahed, 2013). The performance results have been validated through the implementation of the algorithm on various training sets. It is analyzed that the algorithm is not good in terms of running time. Various clustering algorithms related to machine learning, including Decision Tree, SVM, KNN, and Random Forest, are implemented in

(Kandhasomy, 2015) to study the classification of patients' diagnosed diabetes mellitus. The results do not provide an analysis of the performance but of the effectiveness. The K-Means and KNN algorithms' performance has been studied in terms of prediction accuracy, but not on run time. The grouping of patients' epilepsy has been performed in the research (Manjusha, 2016) to check the robustness of K-Means and KNN algorithms. The data were generated from EEG signals through sensors. In this case, K-Means is performing better than the KNN algorithm.

These algorithms were selected for study in this research; both are from the unsupervised machine learning domain. A common and foremost drawback of using these algorithms is that the performance is affected in terms of increased running time if the volume of the dataset is increased. In other words, we may say that the volume and computational cost are directly proportional to each other when working with these algorithms (Al Nuaimi, 2015). Research (Benedetti, 2019) declares that clustering such a huge amount of data as Big Data is a more challenging task, specifically in a system having few hardware resources, because a lot of mathematical operations are needed to perform.

The clustering of information or data is used to add semantics in schemes relating to semantic information retrieval systems to achieve desired goals in terms of accuracy, relevancy, and efficiency. A brief overview of some research works is provided in this section related to the research domain. An enhanced algorithm is proposed (Agarwal, 2014), which forms the groups of data items by ranking technique through the use of a K-dimensional vector model. Another proximity-based clustering scheme was introduced by (Al Nuaimi, 2015) for data clustering. A constraint reformulation strategy is adopted in the scheme specifically designed for the clustering of Big Data. As a result, the proposed scheme provides optimal results as compared to previously proposed approaches in all aspects of information retrieval, especially in the formation of clusters, but not considered in terms of computational cost.

An approach (Hu. H, 2014) is proposed for information retrieval based on clusters in which a hybrid indexing mechanism m is adopted to make the scheme more intelligent and efficient. After experimentation, it was observed that the scheme is much more efficient as a result of performance in terms of running time, but poor effectiveness was found in outcomes. After the formation of data clusters, a cluster grading mechanism was introduced (Gantz, 2012), known as a random field selection-based Markov model. The performance is not as good due to the graph complexity, especially when multiple queries are manipulated concurrently.

A genetic algorithm-based and cluster-oriented information search system (Abawajy, 2015) was proposed and developed by the authors. The clustering of URLs available in the dataset is considered to add in the relevant cluster. The selection criteria are based on the ranking information of the URL. The performance evaluation of the system where it is implemented is ignored in this research.

Clustering of data items is a good approach to implement in data extraction and information retrieval approaches to find out more accurate and relevant data. It is also observed that it reduces the effectiveness of results when clustering algorithms are used along with the combination of ranking algorithms. Whereas it may reduce the computational cost and running time, if schemes

are cluster-oriented, it may include irrelevant data items in results, which is the main drawback of such schemes.

The retrieval of information is a very difficult and challenging task when dealing with Big Data, as compared to retrieving information from relational databases. The information can be retrieved from relational databases only by using SQL (Structured Query Language), while we need specialized algorithms for dealing with Big Data. Researchers are continuously doing research in this hot area, and various schemes have been introduced in this regard. Classification or clustering of data is almost a basic requirement for algorithms to propose schemes useful in the retrieval of information from Big Data. Different research has been introduced on the comparative evaluation of various machine learning algorithms useful for semantic information retrieval, but no one has provided a comprehensive analysis of the implementation of these algorithms on distributed processing systems in Big Data environments.

Comprehensive analysis of K-means and KNN schemes

A complete and comprehensive analysis of research focusing on K-means and KNN schemes is given in Table 1.

Table 1: A Comparative Analysis of Various Schemes Using K-means and KNN algorithms

Approach	Major Contribution	Algorithm	Performance Measures	Effectiveness Measures	Big Data Environment	Parallel & Distributed Processing	Tools Used	Evaluation Metrics
(Karegowda, 2012)	Categorization of Diabetic Patients	K-means and KNN	No	Yes	No	No	Weka	Accuracy
(Habibpour, 2014)	Text Document Clustering	K-means and KNN	Yes	Yes	No	No	VC#.Net 2008	Purity and Entropy
(Manjusha, 2016)	Classification of Epilepsy from Electroencephalogram (EEG) Signals	K-means and KNN	Yes	Yes	No	No	NA	Performance index, Sensitivity, Specificity, Average detection and Quality value
(Kuśmirek, 2019)	Comparison of KNN and K-means	K-means and KNN	No	Yes	No	No	NA	Sensitivity, precision
(Mittal, 2019)	Performance Study	K-means and KNN	Yes	Yes	No	No	Weka, RapidMiner	Responsiveness, Relevance, Validity
(Sieranoja, 2020)	Clustering with KNN Graph and K-Means	K-means and KNN	Yes	Yes	No	No	NA	Overlap, Dimensionality, Unbalance cluster size
(Zhao, 2021)	Classification based on negative databases	K-means and KNN	Yes	Yes	Yes	No	C++, Microsoft Visual Studio 2019	Davies–Bouldin Index (DBI), Precision, Recall and F-measure
(Kaiser, 2021)	Detecting Possible Suspects for COVID-19 Infections	K-means, KNN, and Bayesian Distance Tree (BDT)	No	Yes	No	No	Laravel PHP, Firebase	Health condition, proximity detection, contact tracing, COVID-19 self-test and *COVID score*
(Narayana, 2022)	Detect and Identify the Traffic Signals in Rainy Conditions	K-means, KNN and Support Vector Machine (SVM)	No	Yes	No	No	MATLAB, SPSS	Accuracy

(Andreswari, 2022)	Comparative Analysis of Algorithms for Telecom Fraud Detection	K-means and KNN	No	Yes	No	No	Python	Sensitivity (Recall), Specificity, Precision, F1 Measures, Accuracy
(Gómez, 2022)	The Comparison of Lithological Interpretation	K-means and KNN	No	Yes	No	No	MATLAB	Accuracy
(Nascimento, 2022)	The Performance Evaluation of Big Data Processing with Spark and Unicage in a Cluster Environment	Merge Sort	Yes	No	Yes	Yes	Netdata	Execution Time, Processing Rate, and Loading Rate
(Mi Li, 2023)	Asynchronous Selective Batched based Large Scale Data Clustering using GPU	K-Means, Elkan KM, Rapid-KM	Yes	No	Yes	Yes	Python	Execution Time
Proposed	The Implementation of Two well-known Information Retrieval Algorithms on Parallel and Distributed Systems in Big Data Environment	K-means and KNN	Yes	No	Yes	Yes	Java, Python	Runtime

In the latter part, we describe each research listed in the Table. In (Karegowda, 2012), a hybrid model was proposed for diabetic patient classification by using K-Means and KNN algorithms specifically. Only the effectiveness of algorithms in terms of accuracy in results is focused and the performance or efficiency of algorithms is ignored in the research. All the experiments are performed in a non-distributed or parallel computing environment on a small dataset. In the article (Habibpour, 2014), clustering of text documents is studied by applying K-Means and KNN algorithms in a hybrid mode. Both the performance and effectiveness are analyzed in this study, but no big data parallel processing and distributed computing systems are involved in the experimentation and evaluation phase.

The article entitled "Performance Analysis of KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals", (Manjusha, 2016) provides an analysis of the performance of both clustering algorithms. The researchers were engrossed in the comparative analysis of the performance and the effectiveness of selected algorithms in this research. The article (Kuśmirek, 2019) provides a comparative analysis of both algorithms in terms of effectiveness in results by calculating the recall and precision. The major drawback of this research is that the experimentation is not performed on Big Data and in a distributed or parallel computing environment.

In (Mittal, 2019) authors studied the Performance of the prediction of diagnostic accuracy by implementing KNN and K-Means clustering algorithms. They discussed both the performance and effectiveness measurements as a result of this study. The improved relevance and responsiveness are the major outcomes of this research. A thesis (Sieranoja, 2020) presented about the clustering in KNN graph and K-Means algorithms, which provides a detailed study of the comparison of these algorithms. However, the lack of some common interests including the performance of algorithms in Big Data and distributed or parallel processing environments are still missing in both of the documents that can be a useful source for students and future research scholars and the implementation of algorithms in large business environments.

In recent and advanced research (Zhao, 2021), authors classify the data based on negative databases by using K-Means and KNN algorithms. The only scheme provides a detailed analysis of both algorithms in terms of performance and effectiveness measures. The experimentation was also performed on Big Data, but not in a parallel and distributed environment. The article (Kaiser, 2021) is about the mobile application developed for the detection of possible suspects for COVID-19 infection in the working environment of various businesses. The scheme is implemented by using K-Means, KNN, and BDT (Bayesian Distance Tree) algorithms. The model application only determines the condition of health of the employee, proximity disease detection, and the self-test of registered employees for COVID-19, the tracing of the contacts of affected employees with other employees, and the COVID score of the user. The application is developed for mobile devices; therefore, the proposed model is for a single-processor system, and it is not suitable for Big Data processing with such limited computing resources.

Some other advanced researchers (Narayana, 2022) (Andreswari, 2022) (Gómez, 2022) are also implementing K-Means and KNN algorithms in their schemes. These algorithms are implemented along with the SVM (Support Vector Machine) algorithm in (Narayana, 2022), for the detection

and identification of traffic signals in rainy conditions. In (Andreswari, 2022), the algorithms K-Means and KNN both are both used for telecom fraud detection. In this article, a detailed comparative analysis is provided, and the results are presented as a rate of accuracy, recall, precision, and F1 score. In (Gómez, 2022), a comparative study of the KNN and K-Means clustering methods is presented, and the results are discussed in terms of accuracy for lithological interpretation of well logs of the Shushufindi Oilfield, Ecuador. A major and common disadvantage of these schemes is that they focus only on the performance of algorithms regarding effectiveness in results, but not in terms of computational cost, even on a uniprocessor system. These schemes also do not provide an analysis of these algorithms in Big Data and parallel and distributed processing environments as well.

The work presented in (Mi Li, 2023) proposed an Asynchronous Selective Batched K-Means (ASB K-Means) algorithm for the clustering of unstructured Big Data by using a GPU parallel processing system. The performance of the proposed K-Means is also compared with the performance of some existing enhanced K-Means algorithms, including Elkan-KM and Rapid-KM. This work only focuses on the implementation of K-Means on GPU, and the experimentation is performed by using image and textual datasets. In the research work introduced in (Nascimento, 2022) authors evaluate the performance of parallel and distributed processing systems for Big Data processing with Spark and Unicage in a Cluster Environment. The research doesn't involve any machine learning or data mining algorithm for the implementation. The experiments are performed through the implementation of simple sorting and searching operations on Big Data.

Conclusion

In summary, many schemes work only on the effectiveness measures of K-Means and KNN algorithms, and mostly these algorithms are used in information retrieval schemes. It is also analyzed that most schemes are not implementing and evaluating these algorithms in a Big Data environment, except few research works. Similarly, the systems discussed do not provide support for these two algorithms in distributed and parallel processing environments. There is a crucial need to research such modern environments to set pathways for future research directions and implementations.

References

1. Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC iview, 1142(2011), 1-12.
2. Agneeswaran, V. S., Tonpay, P., & Tiwary, J. (2013). Paradigms for realizing machine learning algorithms. Big Data, 1(4), 207-214.
3. Khan, M. A., Khan, S. U. R., Rehman, H. U., Aladhadh, S., & Lin, D. (2025). Robust InceptionV3 with Novel EYENET Weights for Di-EYENET Ocular Surface Imaging Dataset: Integrating Chain Foraging and Cyclone Aging Techniques. International Journal of Computational Intelligence Systems, 18(1), 204.
4. Yang, H., Khan, S. U. R., Bilal, O., Chen, C., & Zhao, M. (2025). CEOE-Net: Chaotic Evolution Algorithm-Based Optimized Ensemble Framework Enhanced with Dual-Attention for Alzheimer's Diagnosis. Computer Modeling in Engineering & Sciences, 145(2), 2401.

5. Meeran, M. T., Raza, A., & Din, M. (2018). Advancement in GSM Network to Access Cloud Services. Pakistan Journal of Engineering, Technology & Science [ISSN: 2224-2333], 7(1).
6. Maqsood, H., & Khan, S. U. R. (2025). MeD-3D: A multimodal deep learning framework for precise recurrence prediction in clear cell renal cell carcinoma (ccRCC). Expert Systems with Applications, 130174.
7. Khan, S. U. R., Asim, M. N., Vollmer, S., & Dengel, A. (2025). Dynamic Weight Adjustment for Knowledge Distillation: Leveraging Vision Transformer for High-Accuracy Lung Cancer Detection and Real-Time Deployment. arXiv preprint arXiv:2510.20438.
8. Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. Journal of big data, 2(1), 1-20.
9. Khan, S. U. R., Zhao, M., & Li, Y. (2025). Detection of MRI brain tumor using residual skip block based modified MobileNet model. Cluster Computing, 28(4), 248.
10. Khan, S. U. R., Asim, M. N., Vollmer, S., & Dengel, A. (2025). FloraSyntropy-Net: Scalable Deep Learning with Novel FloraSyntropy Archive for Large-Scale Plant Disease Diagnosis. arXiv preprint arXiv:2508.17653.
11. Raza, A., & Meeran, M. T. (2019). Routine of Encryption in Cognitive Radio Network. Mehran University Research Journal of Engineering and Technology [p-ISSN: 0254-7821, e-ISSN: 2413-7219], 38(3), 609-618.
12. Bilal, O., Hekmat, A., & Khan, S. U. R. (2025). Automated cervical cancer cell diagnosis via grid search-optimized multi-CNN ensemble networks. Network Modeling Analysis in Health Informatics and Bioinformatics, 14(1), 67.
13. Emani, C. K., Cullot, N., & Nicolle, C. (2015). Understandable big data: a survey. Computer science review, 17, 70-81.
14. Khan, S. U. R. (2025). Multi-level feature fusion network for kidney disease detection. Computers in Biology and Medicine, 191, 110214.
15. Mahmood, F., Abbas, K., Raza, A., Khan,M.A., & Khan, P.W. (2019). Three Dimensional Agricultural Land Modeling using Unmanned Aerial System (UAS). International Journal of Advanced Computer Science and Applications (IJACSA) [p-ISSN : 2158-107X, e-ISSN : 2156-5570], 10(1).
16. Khan, S. U. R., Asif, S., Bilal, O., & Rehman, H. U. (2025). Lead-cnn: lightweight enhanced dimension reduction convolutional neural network for brain tumor classification. International Journal of Machine Learning and Cybernetics, 1-20.
17. Q. Zhanga, L. T. Yang, Z. Chenc and P. Li, "A survey on deep learning for big data," Information Fusion, vol. 42, pp. 146-157, 2018.
18. HUSSAIN, S., Raza, A., MEERAN, M. T., IJAZ, H. M., & JAMALI, S. (2020). Domain Ontology Based Similarity and Analysis in Higher Education. IEEEP New Horizons Journal, 102(1), 11-16.
19. Khan, S. U. R., Asif, S., & Bilal, O. (2025). Ensemble Architecture of Vision Transformer and CNNs for Breast Cancer Tumor Detection From Mammograms. International Journal of Imaging Systems and Technology, 35(3), e70090.
20. Khan, S. U. R., & Khan, Z. (2025). Detection of Abnormal Cardiac Rhythms Using Feature Fusion Technique with Heart Sound Spectrograms. Journal of Bionic Engineering, 1-20.
21. Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. Information systems research, 25(3), 443-448.

22. S. ur R. Khan, Asif. Raza, Muhammad Tanveer Meeran, and U. Bilhaj, "Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers", VFAST trans. softw. eng., vol. 11, no. 4, pp. 80–92, Dec. 2023.
23. Dykes, B. (2017). Big data: Forget volume and variety, focus on velocity. Forbes, https://www. forbes. com/sites/brentdykes/2017/06/28/big-data-forget-volume-andvariety-focus-on-velocity.
24. O. Bilal, Asif Raza, S. ur R. Khan, and Ghazanfar Ali, "A Contemporary Secure Microservices Discovery Architecture with Service Tags for Smart City Infrastructures ", VFAST trans. softw. eng., vol. 12, no. 1, pp. 79–92, Mar. 2024
25. Rialti, R., Marzi, G., Ciappei, C., & Busso, D. (2019). Big data and dynamic capabilities: a bibliometric analysis and systematic literature review. Management Decision.
26. Khan, S. U. R., Asif, S., Zhao, M., Zou, W., Li, Y., & Li, X. (2025). Optimized deep learning model for comprehensive medical image analysis across multiple modalities. Neurocomputing, 619, 129182.
27. Braganza, A., Brooks, L., Nepelski, D., Ali, M., & Moro, R. (2017). Resource management in big data initiatives: Processes and dynamic capabilities. Journal of Business Research, 70, 328-337.
28. Khan, S. R., Asif Raza, Inzamam Shahzad, & Hafiz Muhammad Ijaz. (2024). Deep transfer CNNs models performance evaluation using unbalanced histopathological breast cancer dataset.Lahore Garrison University Research Journal of Computer Science and Information Technology, 8(1).
29. Rahmati, V. (2016). Big data: Now and then. International Journal of Emerging Computing Methods in Engineering (IJECME), 1(2).
30. Raza, Asif , Soomro, M. H., Shahzad, I., & Batool, S. (2024). Abstractive Text Summarization for Urdu Language. Journal of Computing & Biomedical Informatics, 7(02).
31. Al-Khasawneh, Mahmoud Ahmad, Asif Raza, Saif Ur Rehman Khan, and Zia Khan. "Stock Market Trend Prediction Using Deep Learning Approach." Computational Economics (2024): 1-32
32. Maltby, D. (2011, October). Big data analytics. In 74th Annual Meeting of the Association for Information Science and Technology (ASIST) (pp. 1-6).
33. Asif Raza, Salahuddin, Ghazanfar Ali, Muhammad Hanif Soomro, Saima Batool, "Analyzing the Impact of Artificial Intelligence on Shaping Consumer Demand in E-Commerce: A Critical Review", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.17, No.5, pp. 42-61, 2025.
34. S. U. R. Khan, A. Raza, I. Shahzad and G. Ali, "Enhancing Concrete and Pavement Crack Prediction through Hierarchical Feature Integration with VGG16 and Triple Classifier Ensemble," 2024 Horizons of Information Technology and Engineering (HITE), Lahore, Pakistan, 2024, pp. 1-6.
35. Almeida, F. (2018). Big data: concept, potentialities and vulnerabilities. Emerging Science Journal, 2(1), 1-10.
36. Bilal, Omair, Arash Hekmat, Inzamam Shahzad, Asif Raza, and Saif Ur Rehman Khan. "Boosting Machine Learning Accuracy for Cardiac Disease Prediction: The Role of Advanced Feature Engineering and Model Optimization." The Review of Socionetwork Strategies (2025): 1-30.
37. Khan, Zia, Saif Ur Rehman Khan, Omair Bilal, Asif Raza, and Ghazanfar Ali. "Optimizing Cervical Lesion Detection Using Deep Learning with Particle Swarm Optimization." In 2025 6th International Conference on Advancements in Computational Sciences (ICACS), pp. 1-7. IEEE, 2025.

38. Asif Raza, Inzamam Shahzad, Ghazanfar Ali, and Muhammad Hanif Soomro. "Use Transfer Learning VGG16, Inception, and Reset50 to Classify IoT Challenge in Security Domain via Dataset Bench Mark." Journal of Innovative Computing and Emerging Technologies 5, no. 1 (2025).
39. Khan, Saif Ur Rehman, Asif Raza, Inzamam Shahzad, and Shehzad Khan. "Subcellular Structures Classification in Fluorescence Microscopic Images." International Conference on Computing & Emerging Technologies, pp. 271-286. Cham: Springer Nature Switzerland, 2023.
40. Al Nuaimi, A. J. J., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities Journal of Internet Services and Applications.
41. Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). Big data for dummies (Vol. 336). Hoboken, NJ: John Wiley & Sons.
42. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. IEEE access, 2, 652-687.
43. Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012), 1-16.
44. Bellaachia A, Guven E (2003) Predicting breast cancer survivability using data mining techniques. J Soc Ind Appl Math 7(1):37–42
45. Chi, C. L., Street, W. N., & Wolberg, W. H. (2007). Application of artificial neural network-based survival analysis on two breast cancer datasets. In AMIA annual symposium proceedings (Vol. 2007, p. 130). American Medical Informatics Association.
46. Buciński, A., Bączek, T., Krysiński, J., Szoszkiewicz, R., & Załuski, J. (2007). Clinical data analysis using artificial neural networks (ANN) and principal component analysis (PCA) of patients with breast cancer after mastectomy. Reports of Practical Oncology and Radiotherapy, 12(1), 9-17.
47. Joshi, J., Doshi, R., & Patel, J. (2014). Diagnosis of breast cancer using clustering data mining approach. International Journal of Computer Applications, 101(10), 13-17.
48. Saleema, J. S., Bhagawathi, N., Monica, S., Shenoy, P. D., Venugopal, K. R., & Patnaik, L. M. (2014). Cancer prognosis prediction using balanced stratified sampling. arXiv preprint arXiv:1403.2950.
49. Khosravanian, A., & Ayat, S. (2016). Diagnosing breast cancer type by using probabilistic neural network in decision support system. Int J Knowl Eng, 2(1), 73-6.
50. Parvin H, Alizadeh H, Minaei-Bidgoli B (2008) MKNN: modified K-nearest neighbour. Proceedings of World Congress in Engineering and Computer Science, USA
51. Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), 32(569), 2.
52. Medjahed, S. A., Saadi, T. A., & Benyettou, A. (2013). Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. International Journal of Computer Applications, 62(1).
53. Kandhasamy, J. P., & Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.
54. Manjisha, M., & Hari Kumar, R. (2016). Performance analysis of KNN and K-means clustering for robust classification of epilepsy from EEG signals. In Int Conf Wirel Common Signal Process Netw (NISPNET).
55. Benedetti, F., Beneventano, D., Bergamaschi, S., & Simonini, G. (2019). Computing inter-document similarity with context semantic analysis. Information Systems, 80, 136-147.

56. Abawajy, J. (2015). Comprehensive analysis of big data variety landscape. International journal of parallel, emergent and distributed systems, 30(1), 5-14.
57. Karegowda, A. G., Jayaram, M. A., & Manjunath, A. S. (2012). Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients. International Journal of Engineering and Advanced Technology, 1(3), 147-151.
58. Habibpour, R., & Khalilpour, K. (2014). A new hybrid k-means and k-nearest-neighbor algorithms for text document clustering. International Journal of Academic Research, 6(3).
59. Kuśmirek, W., Szmurło, A., Wiewiórka, M., Nowak, R., & Gambin, T. (2019). Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers' performance. BMC bioinformatics, 20(1), 1-10.
60. Mittal, K., Aggarwal, G., & Mahajan, P. (2019). Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. International Journal of Information Technology, 11(3), 535-540.
61. Sieranoja, S. (2020). Clustering with kNN Graph and k-Means (Doctoral dissertation, Itä-Suomen yliopisto).
62. Zhao, D., Hu, X., Xiong, S., Tian, J., Xiang, J., Zhou, J., & Li, H. (2021). K-means clustering and kNN classification based on negative databases. Applied Soft Computing, 110,107732.
63. Kaiser, M. S., Mahmud, M., Noor, M. B. T., Zenia, N. Z., Al Mamun, S., Mahmud, K. A., & Hussain, A. (2021). iWorkSafe: towards healthy workplaces during COVID-19 with an intelligent pHealth App for industrial settings. Ieee Access, 9, 13814-13828.
64. Narayana, M., & Bhavani, N. P. G. (2022, February). Traffic sign identification on rainy conditions using K means algorithm comparison with KNN and SVM. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-6). IEEE.
65. Andreswari, R., Rahmani, D. A., & Rahmawati, R. (2022, January). Comparative Analysis of K-Means and K-Nearest Neighbor Algorithm for Telecom Fraud Detection. In 2022 2nd International Conference on Information Technology and Education (ICIT&E) (pp. 107-111). IEEE.
66. Gómez, F., Flores, Y., & Vadászi, M. (2022). Comparative analysis of the K-nearest-neighbour method and K-means cluster analysis for lithological interpretation of well logs of the Shushufindi Oilfield, Ecuador. Rudarsko-geološko-naftni zbornik, 37(4), 155-165.
67. Wang, W., Liu, J., Xia, F., King, I., & Tong, H. (2017, April). Shifu: Deep learning-based advisor-advisee relationship mining in scholarly big data. In Proceedings of the 26th international conference on World Wide Web companion (pp. 303-310).
68. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.
69. Nascimento, D. M., Ferreira, M., & Pardal, M. L. (2022). Does Big Data Require Complex Systems? A Performance Comparison between Spark and Unicage Shell Scripts. *arXiv preprint arXiv:2212.13647*.
70. Li, M., Frank, E., & Pfahringer, B. (2023). Large-scale K-means clustering using GPUs. *Data Mining and Knowledge Discovery*, *37*(1), 67-109.
71. Zhao, J., & Pjesivac-Grbovic, J. (2009). MapReduce: The programming model and practice.
72. Reddy, C. K., & Vinzamuri, B. (2018). A survey of partitional and hierarchical clustering algorithms. In *Data clustering* (pp. 87-110). Chapman and Hall/CRC.

73. Lamsal, R. (2021). Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, 51, 2790-2804. https://doi.org/10.1007/s10489-020-02029-z
74. Lopez, C. E., Gallemore, C., "An Augmented Multilingual Twitter dataset for studying the COVID-19 infodemic" Soc. Netw. Anal. Min. 11, 102 (2021). DOI: s13278-021-00825-0
75. Lloyd, S. P. Least squares quantization in pcm. Information Theory, IEEE Transactions on 28, 2 (1982), 129–137.