# Literature review and analysis on big data stream classification techniques

#### **Abstract**

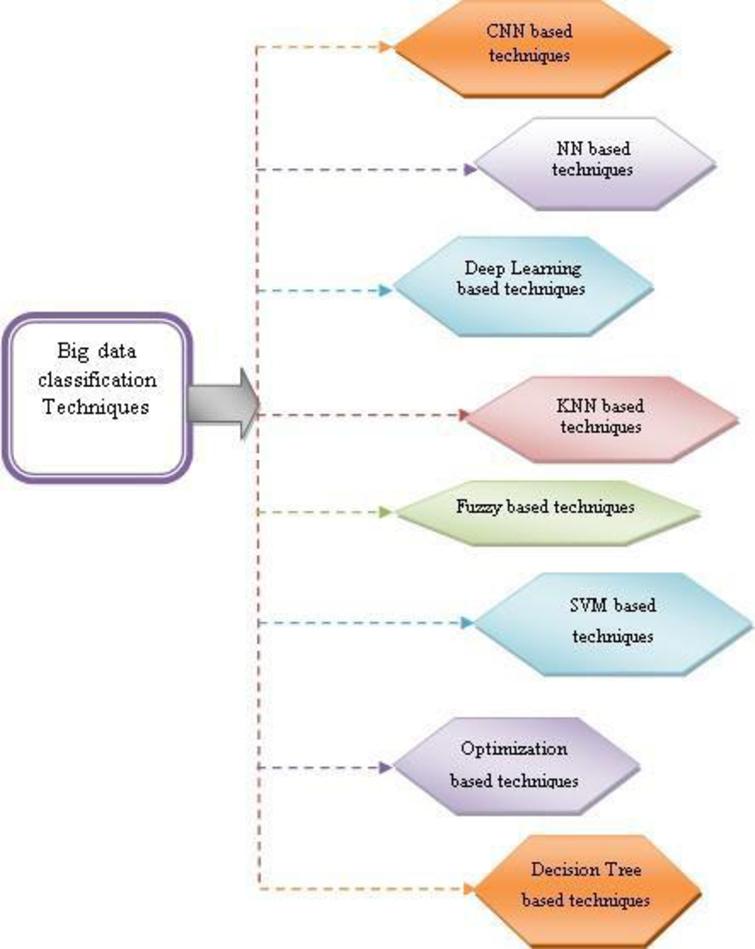
Rapid growth in technology and information lead the human to witness the improved growth in velocity, volume of data, and variety. The data in the business organizations demonstrate the development of big data applications. Because of the improving demand of applications, analysis of sophisticated streaming big data tends to become a significant area in data mining. One of the significant aspects of the research is employing deep learning approaches for effective extraction of complex data representations. Accordingly, this survey provides the detailed review of big data classification methodologies, like deep learning based techniques, Convolutional Neural Network (CNN) based techniques, K-Nearest Neighbor (KNN) based techniques, Neural Network (NN) based techniques, fuzzy based techniques, and Support vector based techniques, and so on. Moreover, a detailed study is made by concerning the parameters, like evaluation metrics, implementation tool, employed framework, datasets utilized, adopted classification methods, and accuracy range obtained by various techniques. Eventually, the research gaps and issues of various big data classification schemes are presented.

## 1.Introduction

Rapid improvement of World Wide Web (WWB), internet data rises and becomes big data. Based on the various properties of big data, big data classification is utilized in network bandwidth, security Filtering network data management, category management, network reputation management, green internet, etc. Big Data is described using velocity, volume, and variety, [[57](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref057)]. If the variety, velocity, and volume of the data are enhanced, the recent technologies are used to manage processing as well as storage of the data. The term Big Data Analytics is nothing, but the process of understanding and analysing the characteristics of vast datasets to extract statistical and geometric patterns [[56](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref056)]. Most of the data generated is originally streaming data. This is because of the data representing actions, measurements, and interactions, which come from the internet. Data is produced from an interval of time. In the streaming framework, high speed data, and algorithms must process very strict constraints of time and space. Streaming algorithms utilize data structures to give speed, and optimal answers [[44](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref044)].

##### **Figure 1.**

Classification of big data techniques.



One of the beneficial tasks in the applications of marketing, biomedicine and social media is the big data classification. The commonly employed model for resolving the challenges of the big data is the single traditional classification model [[58](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref058)]. The classifiers used for big data classification are Naive Bayes (NB), KNN, SVM, and so on. NB classifiers are widely used in information. Data fusion machine analytics is included for big data classification [[11](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref011)], NB is used for classification and target tracking in cloud computing [[12](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref012)] as well as robotics control [[13](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref013)]. Image and text are required for big data analysis, and for cyber analysis it consists of elastic learning and scalable methods [[14](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref014), [10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010)]. By its nature, KNN classifier is a slow classifier and does not have a small fixed-size training model for testing [[9](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref009)]. SVM is employed as a binary classifier, but nowadays, Multiclass SVMs (MSVMs) are used to classify vectors into different sets with the usage of trained oracles, are also being widely studied [[5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005)]. In SVM, the training activity is executed by introducing an optimal hyper plane [[7](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref007)].

The primary intention of this study is to provide a detailed survey of the various big data classification approaches. This review deliberates the existing methods employed in the distinct research works. The survey is done by analysing the employed framework, implementation tool, evaluation metrics, and so on, provided in the research works. Moreover, the adopted framework and the accuracy range are taken for performance evaluation of the suggested big data classification methods. The existing approaches have been categorized into distinct classification and then, the further survey is performed for the exploitation of research gaps and issues. Thus, it acts as the motivation for the future extension of effective big data management.

This paper is organized as: Section 1 gives the brief introduction about the paper, Section 2 provides the literature review of the existing classification approaches and Section 3 discusses the research gaps and issues. The analysis and discussion of the survey is deliberated in Section 4, and Section 5 concludes the paper.

## 2.Literature survey on different big data classification methodologies

This section extensively discusses the survey of different big data classification techniques. The categorization of distinct big data classification systems is pictorized in Fig. [1](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "kes-24-kes200042-g001). The techniques of big data classification are categorized into eight, namely CNN based techniques, NN based techniques, Deep learning based techniques, KNN based techniques, Fuzzy based techniques, SVM based techniques, Optimization based techniques, and Decision Tree based techniques. The review of big data classification techniques provide a clear visualization of the suggested methods along with their significances and drawbacks.

### 2.1Classification of big data classification techniques

The distinct research works adopting various big data classification techniques are elaborated in this section.

#### 2.1.1CNN based classification techniques

The research papers employing CNN based techniques in the big data classification systems are dis-cussed below.

Plathottam et al. [[16](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref016)] presented CNNs for huge datasets. Here, CNN is employed to control complex power grids. The CNN is trained by power data for performing multi-class multi-label classification from Midcontinent ISO (MISO). Tensor Flow is utilized for constructing CNN and to train the network.

Qin et al. [[29](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref029)] developed CNN to control features from raw data and mitigate the application of external resources as well as tool kits. Entity Tag Feature (ETF) is utilized for indicating the location of target nominal, which is easy, but more effective than Position Feature (PF). Dropout strategy is used for alleviating the over-fitting issue.

Cui et al. [[32](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref032)] developed Multi-scale Convolutional Neural Network (MCNN) in a single framework for extraction of features and classification. Based on learnable convolutional layers and multi-branch layer, MCNN automatically extracts features at various frequencies and scales for feature representation. MCNN is computationally efficient, as it naturally leverages Graphics Processing Unit (GPU) computing. However, the method did not use time series classification by other side information from various sources, like image, text, and speech. Deepak Puthal et al. [[43](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref043)] developed Dynamic Key-Length-Based Security Framework (DLSeF) from synchronized prime numbers. Here, the key is updated dynamically at short intervals for ensuring end-to-end security. The training is performed using CNN model for predicting lights at the night time from daytime imagery, at the same time learning features is used to predict poverty.

Dong [[46](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref046)] presented CNN based multiclass classification approach from Electrical Medical Records (EMRs). This framework comprises of two phases. In the first phase, pre-processing is performed with word embedding to represent the samples. In the second phase, training data is segmented to various subsets and training is done based on CNN on every subset.

#### 2.1.2NN based classification technique

The distinct research works employing the NN based techniques for the big data classification systems are presented below, Shan et al. [[21](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref021)] addressed stochastic optimization approach and applied Shuffled Frog-Leaping Algorithm (SFLA) in NN based classifier based parameter optimization. Initially, the NN classifier is developed and comparison is performed with SVM. NN is essential for huge database and it is difficult for extracting high level data. Then, the data set is established, covering speech and cancer data. Both databases have a large number of sample data with complex low level variance. After that, the NN parameter is optimized based on modified version of SFLA.

Sharma and Mangat [[31](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref031)] developed Relevance Vector Machine (RVM) based on data mining approach for big data classification using epidemic outbreak of Ebola virus and compared with other epidemic diseases and generalizing error and intra class separation is performed based on RVM classifier. Dai et al. [[50](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref050)] developed parallel randomized NN for classification. In this framework, optimization of the RNN parallelization is introduced for reducing the communication overhead in the Shuffle phase, like cache-based optimization, partitioning, and thus, enhances the performance and scalability.

#### 2.1.3Deep learning based classification approaches

Alsheikh et al. [[2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref002)] presented scalable spark based approach for deep learning in big data. Deep learning uses a tool for adding value from raw mobile big data. Particularly, distributed deep learning is implemented based on Map Reduce on several spark workers. Every worker is capable of learning a partial deep model on a partition of the overall mobile, and a master deep framework is established for all partial models.

Li et al. [[6](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref006)] developed deep convolutional computation model using tensor representation for big data learning features from the vector space to tensor space. Tensor based approach is utilized for representing every heterogeneous object for revealing the hidden correlations. Tensor CNN uses current CNN to share the parameters. Additionally, a tensor full-connected layer, and a tensor pooling layer are devised for learning high-order features. Back-propagation approach is established for training the parameters of deep convolutional computation model.

Double-projection Deep Computation Model (DPDCM) is developed by Zhang et al. [[8](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref008)] for feature learning. DPDCM projects nonlinear subspaces for learning interacted features to replace each layer based on double-projection layer. Then, mapping is performed using weight sensor for modelling the interactions. After that, back propagation approach is introduced for training the parameters of DPDCM. At last, Privacy Preserving Double Deep Computation Model (PPDPDCM) is established for preventing private information.

Deng et al. [[13](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref013)] developed Fuzzy Deep Neural Network (FDNN) for extracting data from neural and fuzzy representations. The knowledge gained from these two views is merged to form final representation. Here, fuzzy representation mitigates the uncertainties and neural representation eliminates the noises in the raw data. Then, the developed approach used two best representations for final classification. Hence, FDNN is essential for applying data noise, and ambiguity.

Ravi et al. [[14](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref014)] developed deep learning approach for data classification. Initially, pipeline gathers the original data from the inertial sensors. Then, the input data is extracted into segments, in which the features from shallow and deep learning method are estimated in parallel. At last, these two feature sets are combined together and classification is performed based on soft-max and fully connected layers.

Long et al. [[18](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref018)] designed Deep Adaptation Network (DAN) for enhancing the feature transferability. The hidden representations are embedded for producing kernel Hilbert space, in which the embedding’s of various domain distributions are grouped. The domain discrepancy based on optimal multi-kernel selection approach for mean embedding matching is then minimized.

Xiong et al. [[24](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref024)] examined deep learning approach for geosciences data. This approach is utilized for learning meaningful patterns from huge amount of input data to map minerals in the south western Fujian metal organic zone of China. This framework is relevant for discovering difficult features required to detect and classify from the original data.

Koyamada et al. [[25](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref025)] developed DNN-based subject-transfer decoder. Large-scale functional Magnetic Resonance Imaging (fMRI) database is applied for attaining maximum decoding accuracy than other baseline methods. For addressing the knowledge obtained based on decoder, Principal Sensitivity Analysis (PSA) is applied to visualize the discriminative features for all subjects in the dataset.

Zou et al. [[26](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref026)] modelled Tencent deep learning approach for training, simplifying, and support large model in Tencent. Mariana includes three frameworks. They are multi-GPU model parallelism and data parallelism framework, a CPU cluster framework for large scale DNNs.

Kuang and He [[30](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref030)] presented DBN for feature extraction and classification. This framework is showed to be effective in discriminating Attention Deficit Hyperactivity Disorder (ADHD) from subtypes and control. The accuracy is enhanced with the results published in ADHD-200 Competition.

Shafiq and Torunski [[49](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref049)] presented formal model to structure and organize logs. Then, Bayesian technique is used for predicting and detecting any possible faults. Also, Map Reduce based distributed, single-pass, incremental approach is introduced to build, train and implement the developed framework.

#### 2.1.4KNN based classification techniques

The research papers utilizing the KNN based approaches for the big data classification systems are elaborated as follows.

Maillo et al. [[3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref003)] developed KNN classifier for big data classification. Here, the map phase determines k-NN in various splits of the data. Then, the reduce stage is used for computing the definite neighbours from the list attained in the map phase. This approach employs KNN for scaling just by adding more computing nodes. This execution involves the better classification rate based on the original KNN model.

Ramírez-Gallego et al. [[5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005)] employed KNN classifier using apache spark for processing massive and high speed data streams. This method is designed for high speed, large scale, and streaming issues. This approach organizes the examples based on distributed metric tree, which contains top level tree for query routing and a set of distributed sub trees for performing the parallel searches. Then, the instance selection approach is introduced for updating and eliminating outdated instances from the case-base.

Hassanat [[9](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref009)] developed KNN classifier for inserting training instances to binary search tree to speed up the searching time for text samples. Here, two approaches are utilized for sorting the training examples. The primary one is the Norm-based Binary Tree (NBT) for computing maximum or minimum scaled norm. Instances with 1-norms are arranged in the right child, while with 0-norms are arranged in the left child of the node. This process is repeated till the similar norm in a leaf node is obtained. Secondary one is the improved NBT. Here, each example is inserted to search tree depending on the maximum, and the minimum Euclidean norms. At last, sequential KNN classifier is used for classifying feature vectors.

Popescu and Keller [[54](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref054)] presented a fusion strategy for Random Projections Fuzzy KNN (RPFKNN) for big data classification. This method is introduced using the values of class membership, is achieved in each projection using classification accuracy and FKNN.

#### 2.1.5Fuzzy based classification schemes

The research works adopting the fuzzy based techniques for the big data classification techniques are discussed below.

Fernandez et al. [[17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017)] designed evolutionary Fuzzy Rule Based Classification Systems (FR-BCSs) for classification. It consists of Knowledge Base (KB) construct by mans of the Chi-FRBCS-Bi Data approach, along with genetic tuning of the database based on 2-tuples representation. After that, the KB from every map are merged together to produce an ensemble classifier.

Iniguez et al. [[19](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref019)] presented Chi Big Data Support Filtering for classifications. Initially, the classification accuracy and geometric mean are considered. Then, the interpretability of the system is analyzed using rule base size. At last, the scalability is performed based on elapsed training times.

El Bakry et al. [[20](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref020)] developed fuzzy KNN classifier based on Map Reduce for classification. This approach comprises of mapper and reducer. The mapper is utilized for dividing the datasets to chunks and generates intermediate records. These records are produced using map function in the “key, dat” set form. Mapper executes the computing process and passes the results to the reducer. Then, the outputs obtained from the reducer to achieve the final output.

Akil Kumar et al. [[22](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref022)] presented FRBCS for solving the problems in big data, termed Chi-FRBCS-BigData. This approach uses interpretable model for managing vast data and provides accuracy with optimal performance time with Hadoop framework.

Segatori et al. [[28](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref028)] developed distributed Fuzzy Decision Tree (FDT) based on Map Reduce framework to generate multi path and binary way FDTs from big data. This approach uses a distributed fuzzy discretized that produces fuzzy partition for every repeated attribute using fuzzy information. The fuzzy partitions utilized FDT learning for choosing attributes at the decision nodes.

Azar and Hassanien [[36](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref036)] developed Linguistic Hedges Neuro-Fuzzy Classifier with Selected Features (LHNFCSF) for dimensionality reduction, selection of features and classification. In this framework, connection weights, propagation, and activation functions vary from NN.

#### 2.1.6SVM based classification approaches

The research works practicing the SVM based techniques for the big data classification are discussed below.

Rebentrost et al. [[1](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref001)] used optimized binary classifier for classification. A least-square formulation of the SVM permits the usage of phase computation and the quantum matrix inversion approach. The speed of the quantum approach is high, when the kernel matrix is dominated using less number of principal components. The quantum technique does not require representation of all the features of every training example, but it generates the appropriate data structure. After generating the kernel matrix, the individual features of the training data are fully hidden from the user.

Bishwas et al. [[7](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref007)] employed all-pair quantum multiclass-SVM for classification of big data. Here, k(k-1)/2 classifier is established for classifying the given hidden quantum query state. Classical multiclass-SVM can be executed with polynomial run tim; this framework uses quantum version for speed up the system. This framework uses various classification approaches, and speed up gain is attained with respect to their classical counterparts.

Bishwas et al. [[12](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref012)] developed multiclass SVM and Quantum One-Against-All Approach for handling the classification of big data. At first, k quantum binary classifiers are introduced. Then, these classifiers are utilized for classifying unknown quantum state and the prediction is done based on the maximum confidence score.

Zou [[38](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref038)] developed Max-Relevance-Max-Distance (MRMD) feature ranking technique that manages stability and accuracy of feature prediction as well as ranking task. This framework avoids the difficult calculation for ensuring the stability of feature selection.

#### 2.1.7Optimization based classification techniques

The research works practicing the optimization based techniques for the big data classification are discussed below.

Meera and RosilineJeetha [[23](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref023)] developed Acceleration Artificial Bee Colony-Artificial Neural Network (AABC-ANN) to solve high dimensionality and feature selection issues. This framework comprises of pre-processing, selection of features and classification. In the pre-processing step, the KNN approach is utilized for removing irrelevant, and redundancy attributes, for managing the unknown values. After that, the feature selection is done based on AABC for generating the optimal values. The selected features are fed to training and testing. Based on ANN classifier, the features are classified more precisely.

Hegde and Mundada [[47](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref047)] developed Cognitive Particle Swarm Optimization (PSO) in integration with Deep Learning for Big Data classification. This approach is utilized for extracting the features from the given dataset and classification is carried out based on Deep learning. The developed method can adjust its learning rate and the features are extracted efficiently from the obtained data.

Manoj et al. [[52](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref052)] developed Ant Colony Optimization (ACO) and Artificial Neural Network (ANN) based hybrid algorithm for feature selection to text categorization. This hybrid algorithm is found optimal and effective. Here, ACO has the ability of congregating promptly since it has search ability in the state space problem and could find minimal feature subset efficiently.

Lin et al. [[55](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref055)] developed improved Cat Swarm Optimization (CSO) for big data classification. Here, Improved CSO (ICSO) is applied for feature selection, and finally, the classification is performed using SVM. The method did not consider other methods to improve the traditional mode of CSO.

#### 2.1.8Decision tree based classification approaches

The research works adopting the decision tree classifiers for the big data classification are discussed below.

Triguero et al. [[4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref004)] developed big data approach based on apache spark to tackle lack of density issues. Initially, the whole training dataset is divided to chunks, and the extraction is performed using positive samples. After that, the positive sets are broadcasted, so that all the nodes have distinct in-memory copy of the positive samples. For every negative chunk data, a subset of data is obtained based on the positive sample set. After that, Evolutionary under Sampling (EUS) is applied for mitigating the size of both classes and improve the classification performance. At last, various models are merged for predicting the classes of the test set.

Bhagat and Patil [[11](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref011)] developed Synthetic Minority Oversampling Technique (SMOTE) for classification. Initially, Binarization techniques are introduced for decomposing raw dataset to binary class subsets. Then, SMOTE approach is given for every subset of imbalanced binary class for data balancing. At last, Random Forest (RF) classifier is utilized for classification.

Bifet et al. [[39](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref039)] introduced evaluation techniques for streaming big data. This framework analyses unbalanced data streams, in which modification attained on various time scales, and the data is from testing and training from different models. Leszek et al. [[40](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref040)] presented decision trees for data streaming. A new measure for dividing the tree nodes is introduced using impurity measure termed error misclassification. This approach is utilized for deciding the better attribute determined from the whole data stream.

#### 2.1.9Other classification methods

The other classification methodologies adopted for big data classification are elaborated below.

Liu et al. [[10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010)] developed easy and entire system for classification on huge datasets based on NB classifier. Here, additional modules are required for automating the experiment. Yan et al. [[15](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref015)] Developed scalable classifier ensemble technique for combining the results from various classifiers to classify big data. Initially, various“judger” are generated from features and classifiers using training and validation outputs. Then, the judgers are ranked and joined as a boosted classifier. At last, spark-based classification framework is introduced for big data classification.

Peixoto et al. [[27](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref027)] developed Hierarchical Multi-Label Classification (HMC), termed Semantic Hierarchical Multi-Label Classification, for big data classification. Hierarchization, Vectorization, Indexation, Resolution, and Realization, are the steps involved in Semantic HMC. Here, the primary three phases performed label hierarchy from data sources. The remaining two phases are utilized for classifying new items based on the hierarchy labels.

Fong et al. [[33](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref033)] presented lightweight feature selection, termed Swarm Search with Accelerated Particle Swarm Optimization for classification. Here, comparative performance analysis is performed over two groups of classification algorithms, namely batch learning and incremental or data-stream learning, in the light of achieving top accuracy at the shortest possible pre-processing times.

Triguero et al. [[34](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref034)] presented Map Reduce solution for Prototype Reduction (MRPR) for classification. This approach is used to tackle issue that did not comprise of several thousands of examples, because of memory and runtime restrictions. The Map Reduce has transparent simple, and effective environment for prototype reduction computation. Three various reduce types, like Join, Fusion, and Filtering are also analysed for providing more accurate pre-processed sets.

Wu et al. [[35](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref035)] developed HACE for Big Data revolution features, and introduced Big Data processing model from the perspective of data mining. This approach uses mining, demand-driven aggregation of information sources, security as well as privacy considerations, and modelling of user interest.

Chen [[37](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref037)] presented Hierarchical Extreme Learning Machine (H-ELM) using Flink is one of the Graphics Processing Units (GPUs), and in-memory cluster computing platforms. Various optimization techniques are employed for improving performance, like reasonable partitioning strategy, cache-based approach, and memory mapping for mapping specific Java virtual machine objects for buffering. This framework is used for accelerating several variants of ELM and machine learning approaches.

Puthal et al. [[41](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref041)] developed Dynamic Prime Number based Security Verification (DPBSV) for big data streaming. This approach uses common shared key for updating dynamically to generate synchronized pairs of prime numbers. Li et al. [[42](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref042)] developed Privacy-Preserving Outsourced Classification in Cloud Computing (POCC) approach for processing and storing. Proxy fully holomorphic encryption mechanism Using Gentr’s is utilized for preserving the privacy data. Based on POCC, the Crypto Service Provider (CSP) and S evaluator are used to train a classification model from encrypted data with various public keys. This model is then stored in encrypted format in the evaluator that is used for providing service for the clients. However, in between CSP and S, there are some connections. The method failed to mitigate the computation cost and communication.

Vu et al. [[45](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref045)] developed first distributed streaming algorithm for learning decision rules. This framework utilizes horizontal and vertical parallelism for distributing Adaptive Model Rules (AMRules) on a cluster. The decision rules introduced by AM are comprehensible models, in which the antecedent rule is a conjunction on the attribute values, and the consequent is nothing, but the combination of the attributes.

Twardowski and Ryzko [[48](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref048)] presented multi-agent architecture for Big Data processing. This model uses lambda system and it shows autonomous agents are utilized for enhancing architecture, providing robust processing of data in real-time.

Zhai et al. [[51](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref051)] presented fuzzy based ELM classifier for classification. Initially, for every positive case, its nearest neighbor enemy is identified using Map Reduce, and p positive are produced in a random manner in its enemy nearest neighbor based on uniform distribution. After that, balanced Datasets are established and l classifier is trained using constructed data subsets with non-iterative learning. At last, the trained classifier is combined using fuzzy integral for classifying unseen cases.

Bifet et al. [[53](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref053)] developed Scalable Advanced Massive Online Analysis (SAMOA) tool for big data stream mining. This framework is used to gather distributed streaming approaches for machine learning and data mining tasks, like classification, clustering, regression, and programming abstractions. Also, this method measures a pluggable design to run on various processing engines in a distributed manner, namely Samza as well as Storm S4.

## 3.Research gaps identified

This section deals with the various research gaps and challenges existing in the literature works Elaborated, as follows.

The limitations of deep learning based approaches are as follows: In [[2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref002)], deep learning classifier is utilized for speeds up the deep learning model consists of several hidden layers and millions of parameters on a computing cluster. The method in [[13](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref013)] did not consider linear regression model for price value prediction. In [[30](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref030)], DBN does not consider personal characteristic data and Functional Magnetic Resonance Imaging (fMRI)-based information for analysis.

The challenges faced by CNN based classification techniques were as follows: The method in [[29](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref029)] did not consider target nominal for classification, and also, failed to add some linguistic prior knowledge into system. The major challenges in the NN based data classification approaches are: NN [[21](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref021)] failed to consider more data set and compare Shuffled Frog-Leaping Approach (SFLA) with stochastic search algorithms. The challenges faced by the KNN based classification of data are, KNN classifier [[3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref003)] did not use recent techniques, like spark for faster computation.

The major challenges in the fuzzy based classification approaches are deliberated as follows: The method did not consider other programming model, like Spark for improving the efficiency and scalability [[17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017)]. In [[20](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref020)], fuzzy KNN classifier does not consider fuzzy mechanism in the reducer. The technique in [[22](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref022)] requires further improvement in the performance using techniques, like FCBF and Bagging algorithm. The method in [[36](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref036)] did not estimate neural fuzzy in various medical diagnosis issues, such as micro array gene selection, and other data mining problems.

The challenging issues of the optimization based data classification techniques are, ACO [[52](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref052)] did not introduce other classifiers, and failed to use other kinds of data, such as audio, video, images, and so on.

The challenges faced by the Decision tree based data classification techniques are as follows: The method in [[4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref004)] did not consider efficient big data strategies that can deploy pre-processing procedures, like under sampling as well as oversampling approaches in the big data context. Synthetic Minority Oversampling Technique (SMOTE) [[11](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref011)] did not address the problem caused by oversampling rate as it varies with respect to the dataset. The method failed to consider regression, multi-label, and multi-target learning [[39](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref039)].

The challenges faced by other approaches are: The method in [[10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010)] failed to consider information fusion over text and imagery, cyber analysis using cloud computing, and distributed robotics applications. Semantic HMC failed to add integrate between the Reduce and the Map phases to minimize computation time and size of the data to the reducers [[27](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref027)]. In [[45](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref045)], the method did not consider new approaches to deal with both high dimensional and large-scale datasets. In [[34](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref034)], the method did not use structured and multi-target learning.

## 4.Analysis and discussion

The analysis and discussion of the distinct research works for the big data classification in accordance to the employed framework, utilized datasets, accuracy range, evaluation metrics, implementation tool, and adopted classification methods are elaborated in this section.

### 4.1Based on the methodologies employed

The analysis performed on the basis of employed big data classification schemes is elaborated in this subsection. The distinct classification methodologies adopted for the effective classification of big data are depicted in Fig. [2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "kes-24-kes200042-g002). From the Fig. [2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "kes-24-kes200042-g002), it is conveyed that among 55 research papers, about 11 research papers, i.e. 27%, have used deep learning classifier, 11% of the research works have utilized Fuzzy classifier. The CNN classifier is adopted in 12% of the researches, whereas, 7% of the researches, i.e. 3 research papers, employ NN classifier and the remaining works employed other techniques, such as KNN, SVM, and Optimization, and decision tree classifier.

### 4.2Based on implementation tools

This subsection deliberates the analysis carried out considering the simulation tools employed in the research papers. Table [1](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "T1) represents the distinct implementation tools employed for performing the effective big data classification. The commonly practiced implementation tools are Java, Cloudera, MATLAB, Hadoop, and Weka. From Table [1](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "T1), it is clearly shown that the MATLAB tool is the frequently adopted implementation tool for effective big data classification.

### 4.3Based on evaluation metrics

This subsection demonstrates the analysis based on the different evaluation metrics used in the re-search papers.

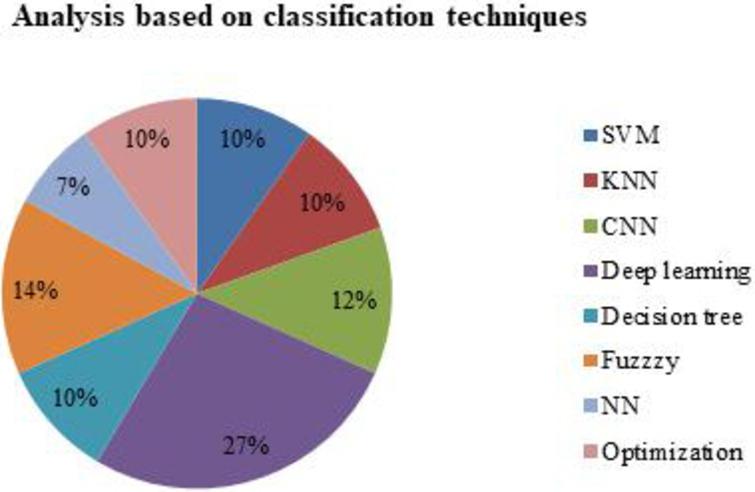
##### **Table 1**

Analysis based on the implementation tools used for big data classificatio

| **Implementation tools** | **Research paper** |
| --- | --- |
| **Java** | [[2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref002)] |
| **Cloudera** | [[3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref003), [4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref004), [5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005)] |
| **Matlab** | [[23](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref023), [36](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref036), [41](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref041), [44](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref044), [46](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref046), [51](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref051)] |
| **Weka** | [[38](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref038)] |
| **Others** | [[6](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref006), [8](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref008), [14](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref014), [16](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref016), [26](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref026), [29](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref029), [21](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref021), [29](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref029), [53](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref053)] |

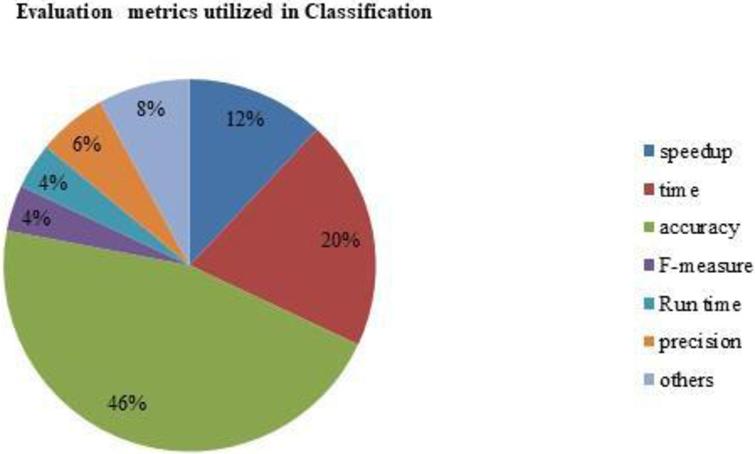
##### **Figure 2.**

Analysis based on the classification schemes.



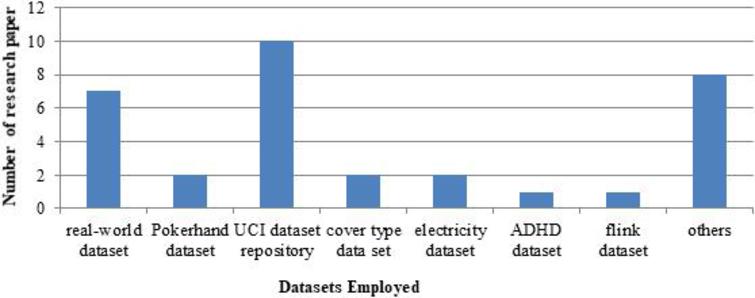
##### **Figure 3.**

Analysis based on evaluation metrics utilized in classification.



##### **Figure 4.**

Analysis based on data set utilized in the classification approaches.



The analysis chart showing the research papers for big data classification system based on evaluation metrics, like time, speedup, accuracy, F-measure, run time, and precision is depicted in Fig. [3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "kes-24-kes200042-g003). From the Fig. [3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "kes-24-kes200042-g003), it is conveyed that among 55 research papers, about 23 research papers, i.e. 46%, have used accuracy metric and 20% of research works have utilized time. Speedup metric is adopted in 12% of the researches, whereas, 8% of the researches, i.e. 4 research papers, employ other metrics. Nearly, 6% of the research works are employed with the precision metric, 4% of the papers used F-measure and the run time is employed in 2 research works. Thus, it is seen that most of the research papers have used the accuracy metric for the big data classification.

### 4.4Based on accuracy range

The analysis carried out in the aspect of the accuracy range is elaborated in this subsection. Table [2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "T2) represents the analysis based on the classification accuracy. From the table, it is elucidated that the accuracy value within the range of 60%–70% is achieved by 5 research papers and 70%–80% is attained by 8 research works. Then, the accuracy between 80%–90% is achieved by 2 research works and 5 research papers attained the accuracy in the range of 90%–99.9%.

### 4.5Based on datasets employed

##### **Table 2**

Analysis based on classification accurac

| **Accuracy range** | **Research paper** |
| --- | --- |
| **60%–70%** | [[19](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref019), [25](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref025), [30](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref030), [40](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref040), [50](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref050)] |
| **70%–80%** | [[8](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref008), [9](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref009), [13](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref013), [17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017), [20](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref020), [22](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref022), [33](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref033), [39](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref039)] |
| **80%–90%** | [[10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010), [18](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref018)] |
| **90%–99.9%** | [[5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005), [6](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref006), [16](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref016), [23](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref023), [47](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref047)] |

The analysis carried out regarding the datasets employed by the distinct research works is elaborated in this subsection. Figure [4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "kes-24-kes200042-g004) depicts the different datasets used for the big data classification. The commonly utilized datasets in the distinct research works are real-world, poker hand, UCI dataset repository, cover type, electricity, ADHD, and flink datasets. From Fig. [4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "kes-24-kes200042-g004), it can be shown that the most frequently utilized datasets are UCI dataset repository, and real-world dataset.

### 4.6Based on the frameworks employed

This subsection deliberates the analysis carried out by considering the adopted frameworks for the big data classification using Table [3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "T3). Through this analysis, it is clearly shown that among the 55 research papers, about 12 research papers have used MR framework and 6 research works are adopting Hadoop framework. Then, 4 research papers have used Apache Hadoop framework and 7 research works have utilized Apache Spark framework. Thus, it is clear that the commonly employed framework to handle big data is the MR framework.

##### **Table 3**

Analysis based on frameworks used for big data classificatio

| **Frameworks** | **Research paper** |
| --- | --- |
| Map Reduce framework | [[3](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref003), [10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010), [11](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref011), [17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017), [19](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref019), [20](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref020), [27](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref027), [37](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref037), [48](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref048), [49](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref049), [50](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref050), [51](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref051)] |
| Apache Hadoop framewor | [[17](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref017), [21](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref021), [34](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref034), [35](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref035)] |
| Apache spark framework | [[2](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref002), [4](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref004), [5](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref005), [11](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref011), [15](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref015), [28](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref028), [39](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref039)] |
| Hadoop framework | [[10](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref010), [22](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref022), [44](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref044), [45](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref045), [48](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref048), [54](https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes200042?resultNumber=7&totalResults=9622&start=0&q=big+data&resultsPageSize=10&rows=10" \l "ref054)] |

## 5.Conclusion

A survey on different data stream mining schemes employed for the effective classification of big data is elucidated in this study. The primitive goal of this paper is to review, categorize and learn the distinct classification utilized for the big data classification through the analysis of various research papers from IEEE, Google Scholar, Elsevier and Science Direct etc. The analysis and discussion were made concerning the evaluation metrics, adopted classification methods, utilized datasets, implementation tool, employed framework, and accuracy range.

Also, this survey suggests the major future scope for the effective big data classification by considering the research gaps and issues. In accordance with the review and analysis, it can be concluded that the accuracy is widely concerned evaluation metric in most of the research works for exploring the effectiveness of big data classification. The vastly adopted framework is MR framewor; and the frequently employed classification methodologies are the deep learning classifier and the fuzzy classifier.