State-of-the-Art Tools and Techniques for Big Data Mining: A Survey

Jaideep Kaur, *M.Tech Student*,GNDU Amritsar, Amit Chhabra, *Assistant Professor*, GNDU Amritsar,Dhanpreet Singh Dhingra, Sr.System Manager, GNDU Amritsar

***Abstract*— Data Mining refers to the discovery of data models in large data sets. It is a process of extracting useful information from the datasets. With the developments in Web 3.0 and Information and Communication Technology (ICT), the cyber world is flooded with petabytes of unstructured and distributed data. Specifically, data mining is a form of machine learning as most data extraction techniques use machine learning algorithms for obtaining useful data segments. Moreover, with the increase in internet utilization, data mining is indeed a necessity for realizing specific applications. Big data, a term coined for the accumulation of large data like petabytes, Exabytes, and so on, requires advance techniques for efficient data abstraction. From the perspective of software, the traditional mining algorithms are applicable only for small scale data. As a consequence, this paper focuses on a reviewing several state-of-the-art data mining techniques and algorithms that are used to process massive data. In addition to these, various advantages and limitations of different approaches are discussed to provide an overview of pros and cons of each technique for efficient utilization in practical environments.**

***Index Terms*— Big Data Mining, Classification, Clustering, Tools and Techniques**

I. INTRODUCTION

DATA mining is defined as the process of extracting useful information from the large, unstructured, and distributed datasets [1]. Basically, it is the mechanism of discovering hidden patterns, sequences, and information from the existing internet-data [2]. Big data, a term used for accumulation of large data sets, depicts enormously diverse, complex, unstructured, and distributed data segments that are generated by Sensors, Actuators, RFIDs, Videos, and Internet-connected ICT sources which are used by users around the world [3]. It is difficult for conventional data management tools and software for effective processing and deployment of big data applications [4]. For instance, internet-generated graphical data on a massive scale easily overwhelms the memory and computational resources available in the computational server systems. Big data includes several inescapable components in the data space to which the big data technology has to respond effectively [5]. In fact, 5 V model has been developed to indicate heterogeneity in data elements as shown in Figure 1. As a consequence, number of mechanisms were developed in the form of data mining tools, for the extraction of useful data sets. In data mining, one of the most premier tasks includes cleansing of data so as to make it feasible for efficient processing.

J. Kaur is M.Tech Student in Department of Computer Engineering and Technology, GNDU Amritsar. Email: Jaipannu6@gmail.com
A.Chhabra is working as Assistant Professor in Department of Computer Engineering and Technology, GNDU Amritsar. Email: amit.cse@gndu.ac.in

This cleanliness of useful data segments is achieved using tools and techniques developed for extraction of specific set of data elements. Data mining deals with static and dynamic data. Static data is easy to handle as most of the information is already known. However, dynamic data varies constantly and unlike static data is not stored beforehand. The main difference between static data and dynamic data is the variability in time that dynamic data processes require as compared static data for effective processing [6].

As discussed earlier, there are numerous techniques and mechanism for extracting useful data. In fact, since data are heterogeneous in nature as it comprises of multi-format, sequential, audio, video, spatial-temporal, and time series data elements, mining useful data is a part of a bigger framework which is defined as Knowledge Discovery in Databases (KDD)[7]. It encompasses a complex procedure from initializing from data accumulation to information modelling.

One of the major data mining technique includes data-classification among several segments which assigns a pre-defined class to each record of a database. Apart from these, clustering represents another important data mining technique, which gathers multiple set of data instead of single record that are similar to each other according to pre-defined metrics. Association, Prediction, and Pattern Generation are other important techniques, which are developed to ensure effective data analyzation [8].

Data mining is the crucial phase in the knowledge discovery as shown in Figure 2. Normally for data pre-processing, it goes through various processes like cleaning, processing, accumulation, transformation, and visualization. Specific algorithms and techniques have been utilized depending upon the decision-making procedure of the applications.

Inspired from these, this paper reviews some of the important contribution in the field of data mining tasks, and techniques that have been developed around the world for extraction of useful data segments. In addition to this, numerous state-of-the-art data mining tools have been listed that are used for effective deployment of data mining procedures.

The presented study is organized in different sections. Section II provides an overview of different data mining techniques that have been developed by various researchers around the world. Section III provides numerous state-of-the-art data mining tools

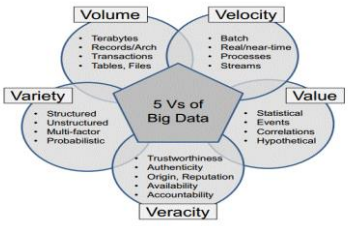

Figure 1: 5V Model of Big Data [9]

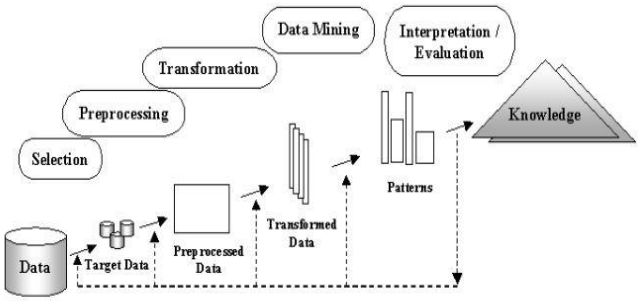

Figure 2: Data Mining Steps [10]

that have been utilized in different data mining applications. Section IV provides a comprehensive overview various works data have been put forth by researchers in the field of data mining and its applications. Finally, Section V concludes the paper with some important discussions.

II. Data Mining Techniques

As discussed in previous section, Data mining techniques have been utilized in different applications like healthcare, weather forecasting, and agriculture. Based on different applications, researchers have developed numerous techniques for extracting useful data sets from heterogenous data repository. Some of the important techniques have been discussed ahead in detail (Figure 3).

A. Association

The association represents an important data mining function for discovering probability of data co-occurrence in a big data environment [11]. The relationships among different co-occurring datasets are represented as association rules. These rules are often used to analyze data elements for effective decision-making. For instance, in e-commerce industry, numerous association rules are used for personalization of Web pages. Infact, an association model is capable of finding that a user who visits pages 1 and 2 is 72%, is likely to visit page 3 in the same or different session. According to this rule, a dynamic link is created among several users who are likely to show interest in page 3. Association is considered one of the best-known data analyzing technique for effective data mining. In addition to this, a pattern is also identified based on the association between data segments. Consequently, association technique is also considered as an effective relation technique.

Even though, several researchers have defined association mining differently, we consider the most cited of them as follows. Let $I=(I^1,I^2,,,,,I^n)$ be a set of n binary features called *items*. Let $D= (d^1,d^2,,,,,d^m)$ be a set of m transactions called database. Each database in D has a unique data segment identifier ID and contains a subset of the data elements in I. A rule is defined as an implication of the form:

For every d^j belonging to D
$A \rightarrow B$, where A, B < I.
Extending this definition,
$A \rightarrow i^j$, where i^j belongs to I.

The above definition provides an overview of the association-based data analyzation. In addition to this, numerous mathematical functions are also defined for increasing the effectiveness of association. These functions include, confidence, support, lift, and conviction.

B. Classification

Classification assigns new data items to already specified categories or classes[12] with major objective to effectively predict the target class of new data element. For example, it can be used to determine loan applicants based on low, medium, or high credit risks. As an instance, a classification model that predicts risks in loan department can be developed using past data of loan applicants for given time. In addition to this, the historical credit rating is capable of tracking employment history, ownership, number of residential years, and type of investments. Even though, there are numerous classification methods, the most fundamental type of classification problem is considered as binary classification. Here, output feature has two values: for instance, high-credit (1) and low-credit (0). In addition to this, other classification involves multi-class targets have more than two values: for instance, low (0), medium (1\2), high (1), or undefined rating [13].

In the model training procedure, the classification algorithm determines appropriate relationships between the values of predictors and target classes. These relationships are usually summarized in a specific-application oriented model, which is effectively applied to a dataset element for which class allocations are unknown. These models are verified by comparison of the predicted values to known values in a dataset. The historical accumulated data for

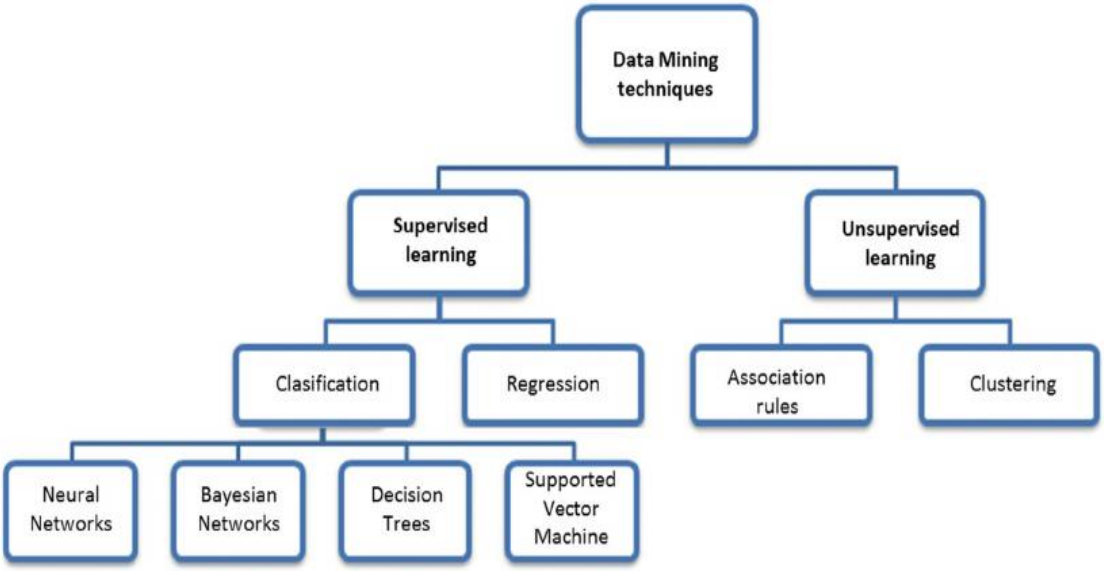

Figure 3: Taxonomy of Data Mining Techniques

Classification is divided into two data set categories: one for building the model; the other for verification of the model.
In addition to this, classification technique is capable of predicting group membership for numerous data instances in real-time. Classification is supervised learning as classes are already labelled. Under training, the model is trained to predict class of datasets. Specifically, there are two types of characteristic features available namely, dependent attribute and independent attribute. In supervised technique of classification, there is one-to-one or one-to-many mapping of input data elements to finite set of discrete classes. Input data set D^i € A, where i is the dimension of input data and discrete class label A € 1.....T, where T is the total number of class types. is modelled in the term of equation D=D(u, w), w is the adjustable parameters.
There are numerous classification techniques provided by the researchers. Few of these have been discussed ahead in detail.

(i) Decision Tree
Decision tree is formulated from the class of labelled tuples [14]. Decision tree is a tree like data structure which comprises of internal nodes, branches, and terminal node. Internal node represents the test of several attributes, branches indicate test outcome, and leaf node or terminal represents the label of class. Two steps involved are learning and testing. The main objective is to determine output class for attribute.

However, there may be several errors in prediction of the appropriate classes using decision tree approach. Consequently, effective pruning algorithms are used in the building of effective decision trees.

(ii) Rule – based classification
It uses IF-THEN rules [15] which are anayzed depending on type of data segments involved. The next step is to analyze how these rules are formulated from training data using different algorithms. Expression for rule is:

IF
{condition}
THEN
{conclusion}

(iii) Backpropagation based Neural Networks
Backpropagation is also known as Neural Network-based learning algorithm [16]. Neural Network (NN) learning is called connectionist learning because it builds association or connections among several data segments. Moreover, it is feasible for deploying in applications where training is required for long time. Neural Network based algorithm proceeds in such a way that it repeatedly performs data processing in a continuous manner and formulates a learning model by comparing the output results with target value provided earlier.

(iv) Bayesian Classifier
Bayes classification is another form of a classification technique which involves the categorization of data segments in different classes [17]. This technique is based on Bayes' theorem with strong (naive) independence assumptions between the data features. In addition to a conventional naïve Bayes classifier, several variations have been put forth by researchers for this effective classification technique. Moreover, the deployment feasibility of naive bayes classifier has led its implementation in several applications.

C. Clustering

Clustering is considered as unsupervised classification technique as it is capable of performing explanatory analysis on unlabeled data [18]. The main objective of this technique is to disperse unlabeled data elements into discrete and finite set of hidden data formats. It is concluded from the fact that there is no procedure of obtaining accurate categorization of unidentified data samples which are generated by same probability distribution function. Specifically, this technique is further classified into two categories as follows.

(a) Hard clustering: In this clustering technique, one object can belong to single category only. In other words, it ensures one-to-one mapping between data elements and given number of classes.

(b) Soft clustering: In this clustering technique, one data object can belong to more than one category, simultaneously. In other words, it supports one-to-many mapping of data elements and specified number of classes. Formally, clustering is illustrated as mentioned ahead.

Let there be set of input data segments X ={x^1,x^2,,,,,,,x^M}, where xi= (x^{i1},x^{i2},,,x^{id}), such that x^{jd} is known dimension, feature, and attribute.

(a) Hard clustering results in the following output:

(i) A= {A^1...., A^K} where (K≤M) and $A^i \neq \phi$
for all i=1,...., M

(ii) $A^i \cap A^j = \emptyset$
for all i= 1,2, …. .K *and* $i \neq j$

(b) Soft clustering has several aspects for representing the generation of output in the form of tree like structure. Clustering of X, $P = P^1$, … … P^r *where* $r \leq$ M *and* $A^i \in P^l$ *and* $A^j \in P^m$ and l > m imply $A^i \in A^j$ *for all* $i,j \neq i,l$ m = 1,2, … . r.

C.1 Clustering Procedure

The clustering procedure involves several sequential steps. In other words, it is a step-by-step process which generates variable output. Following are the steps for performing effective clustering on a distributed data element.

Step 1: Selection and Extraction of Features

Feature selection is identifying distinguished character set of data segments. In addition, to this, extraction transforms data segments to obtain novel and unique features from original data.

Step 2 Designing Clustering Algorithm

Step 2 involves optimization of clustering solutions. In this step, a clustering algorithm is effectively incorporated to form different sets of clusters for accumulated data.

Step 3 Validation

Validation includes determining the validity of the formed clustered in previous steps. Specifically, these all are verified by three indices for testing and validation. These include (a) External indices (b) Internal indices (c) Relative indices. These indices are defined for different clustering data structures like partitioning, and hierarchical clustering.

Step 4 Output Generation

Finally, results are interpreted and generalized. In other words, visualization mechanism is utilized for generation of the clustering procedure.

C.2 Methods of clustering

There are several methods developed by researchers for clustering of heterogenous data elements for solving given problem [19]. Specifically, two main types of clustering techniques have been devised namely, partitioning and hierarchical. In hierarchical based clustering, data sets of n elements are divided into hierarchy of groups resulting in a tree like structure. There are two sub-categories of hierarchical clustering (a) Bottom-up Agglomerative Approach (b) Top-down Decisive Approach. On the other hand, in partitioning methods, the output is partitioned into N datasets elements. This technique results in generation of n-partition with k-objects. Different partitioning-based approaches include (a) Grid based method (b) Subspace based method (c) Density based method (d) Relocation based method. In addition to this, numerous other approaches are also developed for provisioning effective data clustering.

D. Prediction

Prediction is another important technique for mining big data and decision-making analysis. It discovers the relationship among distributed data segments, including dependent and independent variable [20]. This technique of predictive analysis encompasses a variety of statistical techniques including data prediction, predictive modelling, machine-learning, and visualization. Predictive analysis depends on determining output on a priory basis for effective decision mining. Several techniques and tools have been developed to implement this technique like Artificial Neural Networks, Support Vector Machine, Self-organized Mapping, and Bayesian Analysis. As far applicability is concerned, there are several domains where predictive modelling has been extensively utilized for provisioning effective results. These include weather forecasting, healthcare, and agriculture. In healthcare predictive analysis is used to determine the patient's health condition and medical improvement. Weather forecasting used, predictive analysis for determining weather conditions like rain, temperature, and pressure. Similarly, in agriculture, it is used to determine the growth rate of crops, and plants and corresponding growth-risk assessment.

E. Regression

Regression is another important technique for extracting useful data, based on supervised learning [21]. It is used to determine a continuous, numerical data segment. In other words, it is based on training procedure similar to neural networks. Moreover, it estimates values by comparing predefined value with the predicted values. These values are then compiled in some specific model. The generated error, also known as residual, is the difference between predicted and expected value. The major objective of this technique is to eliminate errors for obtaining accurate results. Regression techniques are of two types namely Linear regression and Non-linear regression.

(a) Linear regression Technique

Linear regression is used in scenarios where the relationship between target data element and predictor is represented in straight line given by the equation Z = Ax + B + C. For multivariate linear regression, the regression line is represented as Z = A+ Bx+ Cz+ ⋯ + Xx+ e.

(b) Non- Linear Regression Technique

In this type of technique, data elements cannot be represented in the straight line. As a result, it requires multi variable regression mechanism for generation of predictive results.

III. DATA MINING TOOLS

Data mining tools refer to different software components that are available for deployment in practical scenarios [22]. There are several open source tools available for implementation of data mining. Some of these tools are used for clustering, classification, regression, association, and others. This section describes features of some of the important state-of-the-art data mining tools which can be used to implement effective data mining.

A. Data Mining Tools and Features

(i) *Orange*

It is an open source technology, which uses visual programming or Python language-based scripting. Regression method is used in Orange tool for generation of time-sensitive results. Orange is analysis and visualization tool [23].

(ii) *WEKA*

WEKA is one of most famous data mining tools available commercially. It means Waikato Environment for Knowledge Analysis. Based on the Java programming language, it comprises of tools for data pre-processing, data classification, data clustering, data association and, decision visualization. ARFF (attribute relation file format), CSV (comma separated values) are common formats used by input data. In addition to this, it can be read from a URL and is compatible with SQL database and JDBC. In addition to this, results are visualized in graphical manner for effective analyzation. Several parameters like sensitivity, specificity, accuracy, and f-measure can be obtained using WEKA toolkit [23].

(iii) *SCaVis Scientific Computation and Visualization Environment*

SCaVis stands for Scientific Computation and Visualization Environment. It provides an effective environment for computation of scientific data, analyzation, and visualization. It is especially designed for scientists, engineers, and research students for implementation purposes. The program of SCaVis incorporates several open source software toolkits into a unique interface based on dynamic scripting. Moreover, it provides freedom to select a specific programming language, operating system, and freedom to share and distribute code. In addition to this, there is a provision of multi-document support, and multi-Eclipse bookmarks. Moreover, there is extensive LaTeX support for building effective documentation [24].

(iv) *Apache Mahout*

The Apache Mahout's main objective is to develop machine learning-based library which scalable to enormous data set. For performing data classification following algorithms have been included in toolkit (i) Logistic Regression (ii) Naive Bayes/ Complementary Naive Bayes (iii) Random Forest (iv) Hidden Markov Models and (v) Multilayer Perceptron. In addition, to this several clustering algorithms are also incorporated like Canopy Clustering, Streaming k-Means, Fuzzy k-Means, k-Means Clustering, Spectral Clustering. Data visualization is performed using graphical visualization and statistical analysis [24].

(v) *R Software Environment*

It is free software environment for graphical and statistical purpose. It is an incorporated collection of software services such as calculation, graphical display, and data manipulation. Graphical as well as statistical techniques like linear and non-linear modelling, classification, clustering, and classical statistical tests are provided by software environment [25].

(vi) *ML Flex*

To predict the values of variable that are dependent, ML Flex uses artificial intelligence techniques to derive models of independent variables [24].

(vii) *ESOM (Emergent Self Organizing Maps) tool*

Visualization, Training, Analyzing, Data pre-processing, and Grouping is done with the help of ESOM. Data space is a space that consists of collection of data points from high dimensional space. Online training and the batch training are the two most important training algorithms that are used in ESOM tools. This model is searched for each data input using the online and batch training algorithms, and the optimal matched model is selected. The best matched model is updated instantly in online training algorithm. In other scenarios, all the best matches are grouped together and then updation is performed [26].

(viii) *NLTK (Natural Language Tool Kit)*

To understand the human language, the NLTK works as a platform for building the Python programs. It incorporates highly reliable interfaces to over 50 data formats. It consists of the collection of data processing techniques for parsing, stemming, classification, tagging, and tokenization. In addition to this, it also consists of resources like WordNet. Windows, Mac OS X, and Linux use the NLTK for understanding human language efficiently. It is the project that is open source and easily accessible. Naive Bayes Classifier, Decision Tree Classifier, Maxent Classifier, Conditional Exponential Classifier, and Weka Classifier are the different types of classification techniques used in NLTK [27].

(ix) *ELKI (Environment for Developing KDD Applications Supported by Index- Structures)*

The ELKI is an algorithm that is used for research and development. It focuses on unsupervised learning that is used for analyzing data cluster and for outlier detection. It is an open source software written in Java. In order to achieve performance and scalability at the higher index, it uses data structures index like R*-tree. However, the algorithms like data mining, data types, distances, distance function or file parsers are not the part of this approach [24].

(x) *Unstructured Information Management Architecture*

To abstract the appropriate information, wide range of amorphous data is analyzed. The function is decomposed into a number of modules. To maintain the data flow between parts and to manage them, this framework is utilized. Frameworks, parts and infrastructure are the basic availability characteristics of this tool [24].

(xi) *GraphLab*

GraphLab consist of the toolkit in which several data mining algorithms are implemented simultaneously. In GraphLab programming, an algorithm can be visualized using graphical data structure [24].

(xii) *Mlpy machine learning Python*

Mlpy incorporates algorithms for regression, clustering, and classification. In addition to this, it supports dimensionally reduction and wavelet transform. Moreover, different statistical algorithms like feature ranking, resampling algorithm, error evaluation [24].

(xiii) *KEEL (Knowledge Extraction Evolutionary Learning)*
KEEL is open source data mining software based on Java language, which have license of GPLv3(General Public License version 3). It provides access to evolutionary learning and basic soft computing techniques for different data analysis problems [28].

(xiv**)** *Scikit-learn*
Scikit-learn is also an open source software available for data analyzation. Based on Python, it is an extension of NumPy and SciPy packages. It also incorporates matplot.lib package for plotting graphical charts. This tool supports many data mining algorithms except classification rules, and association rules [24].

IV. DATA MINING: APPLICATION REVIEW

As discussed in previous sections, data mining has a wide range of industrial applicability, extending from healthcare industry to social network industry. Apart from these, numerous data mining techniques have been discussed, that are capable of extracting useful data segments based on data formats, patterns, and state of data generation like time, and space. Consequently, numerous researchers have been working on deploying these techniques different applications. *Rochd et al.* [5] have reviewed several algorithms for extraction of distributed data suing Hadoop and Spark models for big data. Specifically, authors have focused on identifying frequent item-sets in the collection of large data sets, in terms of frequency, and existential features. *Leung et al*. [29] presented a constrained-based data mining technique which focuses on extraction of frequent patterns that are specific to users. In addition to this, the presented technique runs in fog computing environment. In such environment, computation is performed at network edges of the computing architecture. *Hoi et al.* [30] presented a constrained-based data mining algorithm which enables large number of users to collectively vote for their specific patterns. The proposed technique attains advantages of crowdsourcing, simultaneous crowd voting, and data filtering for data analysis and mining of constrained repeated patterns in big data applications and services. *Manigandan et al*. [18] have used data analytics tools to analyze business related decision making. In other words, authors have developed intelligent business model that is capable of provisioning decision-making results in time sensitive manner. Additionally, authors have used data mining tools to generated business reports like trends, growth rate, and profit. Moreover, classification and clustering techniques have been effectively incorporated in the model to present numerous enhancements in the overall model. For implementational purpose, authors have used WEKA tool for generation of results on different datasets. *Ahmad et al.* [31] have provided a comprehensive literature review across five massive healthcare databases from 170 articles, from which nearly 7 articles were identified in the final procedure. Authors concluded that many prediction models utilize different classification techniques, like decision trees, artificial neural network, dynamic support vector machines, and Bayes classification algorithms in which cardio-related diseases are studied. Furthermore, the presented research had a useful applicability in obtaining similar algorithms for developing risk prediction models for different diseases. *Chaturvedi and Saritha* [32] proposed a procedure to determine and abstract frequently occurring patterns from a social-network based data utilizing the concepts of thresholds for support and confidence. Social Network analysis for pattern mining has been performed on Facebook, and Twitter. For implementation purpose, parallel computation is achieved with Apache Spark programs. The acquired patterns are useful in making decisions with r4espect to social media. *Ahuja et al*. [33] has analyzed a comparative performance of several clustering and classification techniques which are applied on the online educational dataset. Educational Data Mining (EDM) incorporates data mining algorithms to explore different educational statistics for identification of patterns and predictions for indicating learner's performance. Numerous design challenges like objective, accuracy, functionality, and overheads are determined for large dataset. Data mining algorithms indulged are classified as centroid-based clustering, graphical clustering, and supervised classification algorithms. Implementational results depict high accuracy of the presented techniques. *Si et al.* [34] used data mining approach to present a novel technique for time series data, namely Three-Dimensional Piecewise Cloud Representation (TDPCR). The presented technique comprises of new partitioning mechanism which secure information between consecutive points by overlapping two neighboring segments. Utilizing cloud model theory, the proposed technique acquires minimization of data dimensionality and captures data distribution. Moreover, a new distance measure, that has adaptive weight factors is defined to highlight relationship among two three-dimensional clouds. For comparative analysis, a successful performance evaluation for the proposed technique is obtained in classification and query of content tasks. Malik et al. [35] have reviewed various data mining technique deployment in healthcare applications. The presented researcher enables feasible employment of the data mining mechanism for various health are applications like predictive healthcare, real-time healthcare delivery, and mobile health surveillance.
Based on the works, it can be depicted that data mining and related techniques are efficiently deployed in different domains like pattern mining, frequency mining, and statistical mining.

V. CONCLUSION

Big data has been a keen area of research over the past few years. Researchers around the world have developed numerous techniques and mechanism to provide effective data mining tools for extracting useful data in widely distributed heterogenous data sets. Inspired from these aspects, this paper provides a comprehensive overview of several state-of-the-art data mining techniques that have been effectively utilized in several applications. In addition to this, several data mining tools have also been discussed to review different features and characteristics of these tools for specific data analyzation. In Moreover, this paper reviews several important literatures that have incorporated data mining techniques for obtaining decision-making services with high efficacy. Henceforth, it can be concluded that numerous applications have been developed to incorporate data mining techniques. Moreover,

time-sensitive applications are immensely benefited from these techniques.

VI. REFERENCES

[1] F. A. M. Zaki and N. F. Zulkurnain, "Frequent Itemset Mining in High Dimensional Data: A Review," Springer, Singapore, 2019, pp. 325–334.

[2] F. Amato, G. Cozzolino, F. Moscato, V. Moscato, A. Picariello, and G. Sperli, "Data Mining in Social Network," Springer, Cham, 2019, pp. 53–63.

[3] J. Kaur and K. Garg, "Efficient Management of Web Data by Applying Web Mining Pre-processing Methodologies," Springer, Singapore, 2019, pp. 115–122.

[4] G. Muggenhuber, "Geospatial Data Mining and Analytics for Real-Estate Applications," Springer, Cham, 2019, pp. 225–240.

[5] Y. Rochd, I. Hafidi, and B. Ouartassi, "A Review of Scalable Algorithms for Frequent Itemset Mining for Big Data Using Hadoop and Spark," Springer, Cham, 2019, pp. 90–99.

[6] H. Yao, M. Xiong, D. Zeng, and J. Gong, "Mining multiple spatial–temporal paths from social media data," *Futur. Gener. Comput. Syst.*, vol. 87, pp. 782–791, Oct. 2018.

[7] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-Temporal Data Mining," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–41, Aug. 2018.

[8] J. Kozak, "Evolutionary Computing Techniques in Data Mining," Springer, Cham, 2019, pp. 29–44.

[9] S. Sapna, "Fusion of big data and neural networks for predicting thyroid," *undefined*, 2016.

[10] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2/3, pp. 107–145, 2001.

[11] R. Dautov and S. Mosin, "A technique to aggregate classes of analog fault diagnostic data based on association rule mining," in *2018 19th International Symposium on Quality Electronic Design (ISQED)*, 2018, pp. 238–243.

[12] S. Anwar Lashari, R. Ibrahim, N. Senan, and N. S. A. M. Taujuddin, "Application of Data Mining Techniques for Medical Data Classification: A Review," *MATEC Web Conf.*, vol. 150, p. 06003, Feb. 2018.

[13] R. Sahani, Shatabdinalini, C. Rout, J. Chandrakanta Badajena, A. K. Jena, and H. Das, "Classification of Intrusion Detection Using Data Mining Techniques," Springer, Singapore, 2018, pp. 753–764.

[14] H. AbouEisha, T. Amin, I. Chikalov, S. Hussain, and M. Moshkov, "Different Kinds of Decision Trees," Springer, Cham, 2019, pp. 35–48.

[15] S. H. Basha, A. Tharwat, K. Ahmed, and A. E. Hassanien, "A Predictive Model for Seminal Quality Using Neutrosophic Rule-Based Classification System," Springer, Cham, 2019, pp. 495–504.

[16] M. Bhatia and S. K. Sood, "An intelligent framework for workouts in gymnasium: M-Health perspective," *Comput. Electr. Eng.*, vol. 65, 2018.

[17] W. Hadi, Q. A. Al-Radaideh, and S. Alhawari, "Integrating associative rule-based classification with Naïve Bayes for text classification," *Appl. Soft Comput.*, vol. 69, pp. 344–356, Aug. 2018.

[18] E. Manigandan, V. Shanthi, and M. Kasthuri, "Parallel Clustering for Data Mining in CRM," Springer, Singapore, 2019, pp. 117–127.

[19] I. Mishra, I. Mishra, and J. Prakash, "Differential Evolution with Local Search Algorithms for Data Clustering: A Comparative Study," Springer, Singapore, 2019, pp. 557–567.

[20] J. Li, X. Li, and L. Yu, "Ship traffic flow prediction based on AIS data mining," in *2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 2018, pp. 825–829.

[21] P. Chertchom, "A comparison study between data mining tools over regression methods: Recommendation for SMEs," in *2018 5th International Conference on Business and Industrial Research (ICBIR)*, 2018, pp. 46–50.

[22] S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study," Springer, Cham, 2019, pp. 196–211.

[23] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Comput. Sci.*, vol. 85, pp. 662–668, Jan. 2016.

[24] "Top 10 open source data mining tools - Open Source For You." [Online]. Available: https://opensourceforu.com/2017/03/top-10-open-source-data-mining-tools/. [Accessed: 21-Oct-2018].

[25] T. Fujino, "vdmR: Generating Web-Based Visual Data Mining Tools with R," *J. Stat. Softw.*, vol. 82, no. 6, pp. 1–16, Nov. 2017.

[26] A. Ultsch, A. Ultsch, and F. Mörchen, "ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM," *DATA BIONICS Res. GROUP, Univ. Marbg.*, vol. 17, p. 46, 2005.

[27] "Deep Learning using TensorFlow and NLTK — Analyzing corpus's sentiments[Part 1]." [Online]. Available: https://becominghuman.ai/deep-learning-using-tensorflow-and-nltk-analyzing-corpuss-sentiments-part-1-bec9d6c1051. [Accessed: 21-Oct-2018].

[28] J. Alcalá-Fdez *et al.*, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Comput.*, vol. 13, no. 3, pp. 307–318, Feb. 2009.

[29] C. K. Leung, D. Deng, C. S. H. Hoi, and W. Lee, "Constrained Big Data Mining in an Edge Computing Environment," Springer, Singapore, 2019, pp. 61–68.

[30] C. S. H. Hoi, D. Khowaja, and C. K. Leung, "Constrained Frequent Pattern Mining from Big Data Via Crowdsourcing," Springer, Singapore, 2019, pp. 69–79.

[31] W. M. T. W. Ahmad, N. L. A. Ghani, and S. M. Drus, "Data Mining Techniques for Disease Risk Prediction Model: A Systematic Literature Review," Springer, Cham, 2019, pp. 40–46.

[32] S. Chaturvedi and S. K. Saritha, "Parallel Frequent Pattern Mining on Natural Language-Based Social Media Data," Springer, Singapore, 2019, pp. 507–517.

[33] R. Ahuja, A. Jha, R. Maurya, and R. Srivastava, "Analysis of Educational Data Mining," Springer, Singapore, 2019, pp. 897–907.

[34] G. Si *et al.*, "Three-dimensional piecewise cloud representation for time series data mining," *Neurocomputing*, vol. 316, pp. 78–94, Nov. 2018.

[35] M. M. Malik, S. Abdallah, and M. Ala'raj, "Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review," *Ann. Oper. Res.*, vol. 270, no. 1–2, pp. 287–312, Nov. 2018.