ISSN: 0067-2904

A review of Medical Diagnostics Via Data Mining Techniques

Sarah Sameer Rasheed[1*], Suhad Faisal Behadili[1], Iyden Kamil Mohammed[2],Mustafa S. Abd[1]

[1]University of Baghdad, College of Science, Computer Science Department
[2]University of Baghdad, Al-Khawarzmy Engineering College, Biomedical Department

Received: 7/3/2020 Accepted: 1/8/2020

Abstract

Data mining is one of the most popular analysis methods in medical research. It involves finding patterns and correlations in previously unknown datasets. Data mining encompasses various areas of biomedical research, including data collection, clinical decision support, illness or safety monitoring, public health, and inquiry research. Health analytics frequently uses computational methods for data mining, such as clustering, classification, and regression. Studies of large numbers of diverse heterogeneous documents, including biological and electronic information, provided extensive material to medical and health studies.

Keywords: Classification, Data mining, Decision tree, Hierarchical clustering, K-means.

مراجعة التشخيص الطبي من خلال تقنيات استخراج البيانات

سارة سمير رشيد*[1] ، سهاد فيصل بهادلي[1] , ، ايدن كامل محمد[2] , مصطفى سلمان عبد [1]

[1]جامعة بغداد، كلية العلوم، قسم علوم الحاسبات
[2]جامعة بغداد، كلية الهندسة خوارزمي، قسم الطب الحيوي

الخلاصة:

يعد استخراج البيانات أحد أكثر طرق التحليل شيوعًا في البحث الطبي. وهي تنطوي على إيجاد أنماط وارتباطات في مجموعات بيانات لم تكن معروفة من قبل. يشمل استخراج البيانات مجالات مختلفة من البحوث الطبية الحيوية، بما في ذلك جمع البيانات، ودعم القرار السريري، ومراقبة المرض أو السلامة ، والصحة العامة وأبحاث الاستفسار. كثيرا ما تستخدم التحليلات الصحية الطرق الحسابية لاستخراج البيانات، مثل التجميع والتصنيف والانحدار. قدمت الدراسات لأعداد كبيرة من الوثائق المتنوعة غير المتجانسة، بما في ذلك المعلومات البيولوجية والإلكترونية، الدراسات الطبية والصحية.

1. Introduction

The information technology has led to huge amounts of data stocking in various formats, including recordings, documents, images, sound recordings, videos, scientific data, and many new formats in various fields of human life. The data gathered from various applications require an appropriate

*Email: sarahsameer@scbaghdad.edu.iq

mechanism to obtain knowledge / information for better decision making from the larger repositories. The goal is to find useful information from large data sets in the process of knowledge exploration in databases (KDD) [1, 2], also known as data mining (DM) [3]. DM has flourished, with varied combines and progressing in analytics, databases, machine learning, pattern recognition, artificial intelligence etc. [4]. DM has been expanded into other areas of human life. This review article is structured to describe the DM method in section 2, DM tasks in section 3, DM algorithms in section 4, DM resources in section 5, literature review in section 6, DM implementations in section 7, and a final conclusion section.

2. Data Mining Process

DM is defined as one of multiple steps of knowledge discovery, involving the application of data analysis and the discovery of algorithms that accurately list patterns on data in any allowable computational proficiency [5]. This process is collaborative and repetitive, hence involves numerous steps, user decisions, and attempts to complete a specific discovery task, each performed using the discovery method. For some phrases KDD, DM synonymously used, for example Figure- 1, consider it to be a key step in KDD, which results in favorable patterns or data models [1,2,6]. The manipulated information, i.e. if it is positive or negative, could commonly affect the investigative effects of KDD [7]. A description of the phases of the DM process is provided in the following sections.

2.1 Data Cleaning

Data are gathered from various origins, involving unwanted, misleading and lacking data, so that they ought to be cleaned and screened to render them useful [8]. Data will be cleaned in this step of the DM process [9]. In reality, data are noisy, conflicting, inconsistent, and deficient [2,10]. The process incorporates various techniques. For illustration, the missing values are filled in, combined, and manipulated. The yield of the data cleaning process is sufficiently cleaned data [11,12].

2.2 Data Integration

Data from different data sources are incorporated into one set in this phase of the DM technique Data are stored in various formats, including spreadsheets, text files, databases, records, data cubes, documents etc. The processing task of data is difficult and complex. This is because data do not necessarily outfit various sources [2,6,11].

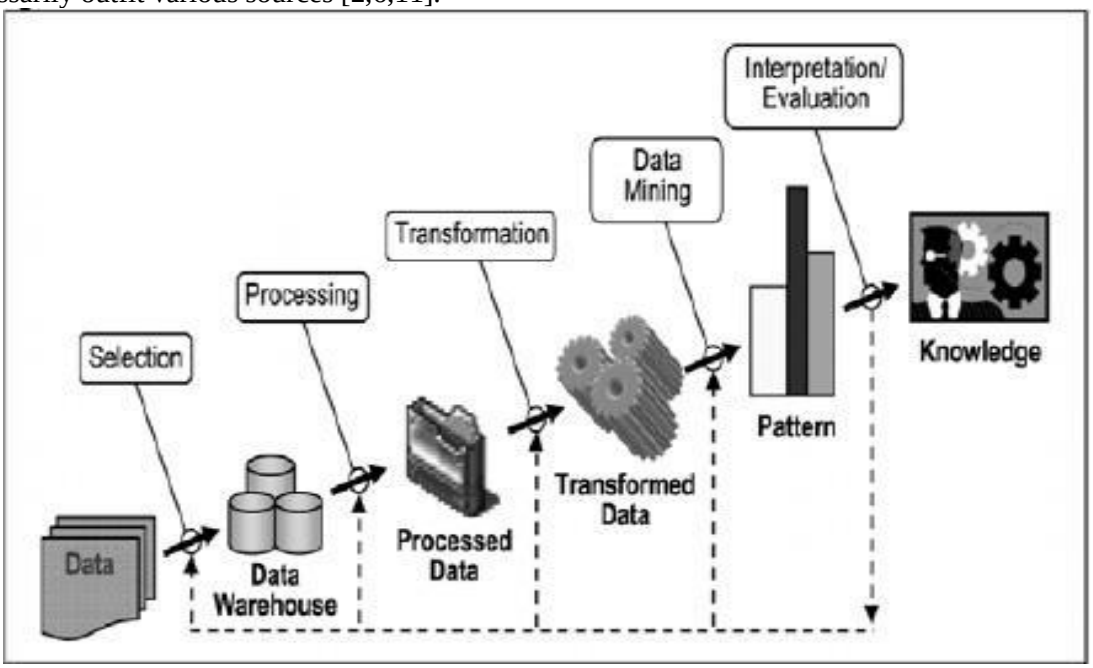

Figure 1- KDD process steps [7].

2.3 Data Selection

The process of acquiring data from the database is an important detail for the analysis approach. This process demands a wide range of objective historical data [1,2]. However, the data archive in the repository usually contains far more data than it is truly required for data integration. Hence, valuable data should be extracted, stored, and processed from the existing data [6,11].

2.4 Data Transformation

The data consolidation process is known as data transformation. In this phase, the collected data are converted into forms suitable for the mining process [1,2]. This process usually involves normalization, aggregation, generalization, and other steps [6, 9,11].

2.5 Data Mining

This is the key step in implementing smart techniques to derive potentially useful patterns [1,2,10]. Additionally, there are multiple tasks in this step. For instance, prediction, classification, clustering, etc [11].

2.6 Pattern Evaluation

The pattern evaluation process states the patterns that are really intriguing. However, the Knowledge represented relied on various kinds of relevant measures [2]. If a pattern is potentially helpful, then it is regarded as interesting and, therefore, easy for people to recognize. Nevertheless, there are several hypotheses that one needs to affirm with some degree of certainty for each new available data [11].

2.7 Knowledge Representation

This is the final phase of visual representation of the observed knowledge to users [1,10]. This is the important phase where the user can recognize and perceive DM outcomes using the visualization techniques [2,11].

3. Data Mining Tasks

There are two major groups in DM tasks, which are the predictive and the descriptive. These two groups are mainly the goals of the DM [13]. There are six major functions of the DM, described namely as clustering, classification, dependency modeling, regression, anomaly detection, and summarizing [5]. The classification, regression, and anomaly detection are classified as predictive categories, while clustering and dependency modelling are categorized as descriptive categories. The predictive model utilizes some dataset variables in order to predict unknown values for certain variables, while the descriptive model categories patterns or relationships and incorporates human understandable patterns and data trends. The definitions of the DM tasks can be found in the following sections [1]:

3.1 Data Classification

The supervised learning model utilizes the classification approach in order to learn how to categorize data classes [13]. This is critical for decision-making management [10]. It identifies common properties among a set of objects in a database and classifies them according to the classification model into separate classes [4, 14]. The major objectives of this step are the analysis of the training data and the creation of a precise description, i.e. a model for each class, by using data features [15]. Mathematical methods, such as decision-trees, neural network, and statistics, are used in this approach [1,16]. The classification method is commonly categorized into two parts, represented by the training and the testing phases. The classification algorithm is charged with creating a classification model supported by the training set. Subsequently in the test phase, model performance is assessed [12,17]. Examples may include a bank loan authority that has to evaluate the data to find out who is "free" and who is "risky" to the borrower, an AllElectronics marketing manager who wants data analysis to determine if a consumer with a specific profile should purchase a new device, or a medical researcher who requires to analyze breast cancer data to anticipate the suitable treatment among three options. However, in each of these cases, the classification method is used for data analysis, where a model or a classifier is constituted to anticipate class (categorical) labels. It requires the classification of the data, as in the example in Figure- 2 of loan application; “yes” or “no” for marketing data; “treatment A” or “treatment B”, or “treatment C” for medical data. In each case, a data analysis process involves classification, which implies the estimation of a formula, or classification. Both groups can be expressed by discrete values that do not require ordering of values. The values 1, 2, and 3, for example, can be used to stand for the A, B, and C treatments, where the following treatment regimens have not implied ordering [18]. Hence, the classification is more effective than the clustering in the medical field, because it is based on more information which is closer to reality, as compared to the approximation used in the clustering method. This represents an important part in such type of data.

3.2 Regression

Regression is a supervised learning tool for continuous performance variables. Also, it is a statistical analysis tool, which quantifies the relationship between a dependent variable and one or more independent variables to explore the data patterns [5,14]. However, there are two main categories: normal predictive regression and logistic classification regression [12]. The predictive variable can be a continuous variable in normal regression. Assume that the marketing manager can anticipate how much a customer spends on AllElectronics during a sale process. The data analysis task of this data is an illustration of numerical anticipation. However, in the comparison to the class label, the concept design predicts a continuous valued function or ordered value. It is a predictor model. Whereas, regression analysis is a statistical method that is employed most frequently for numerical prediction and hence, the two terms are used synonymously, although there are other numerical prediction methods. The two main types of prediction issues are the classification and numeric prediction issues [18].

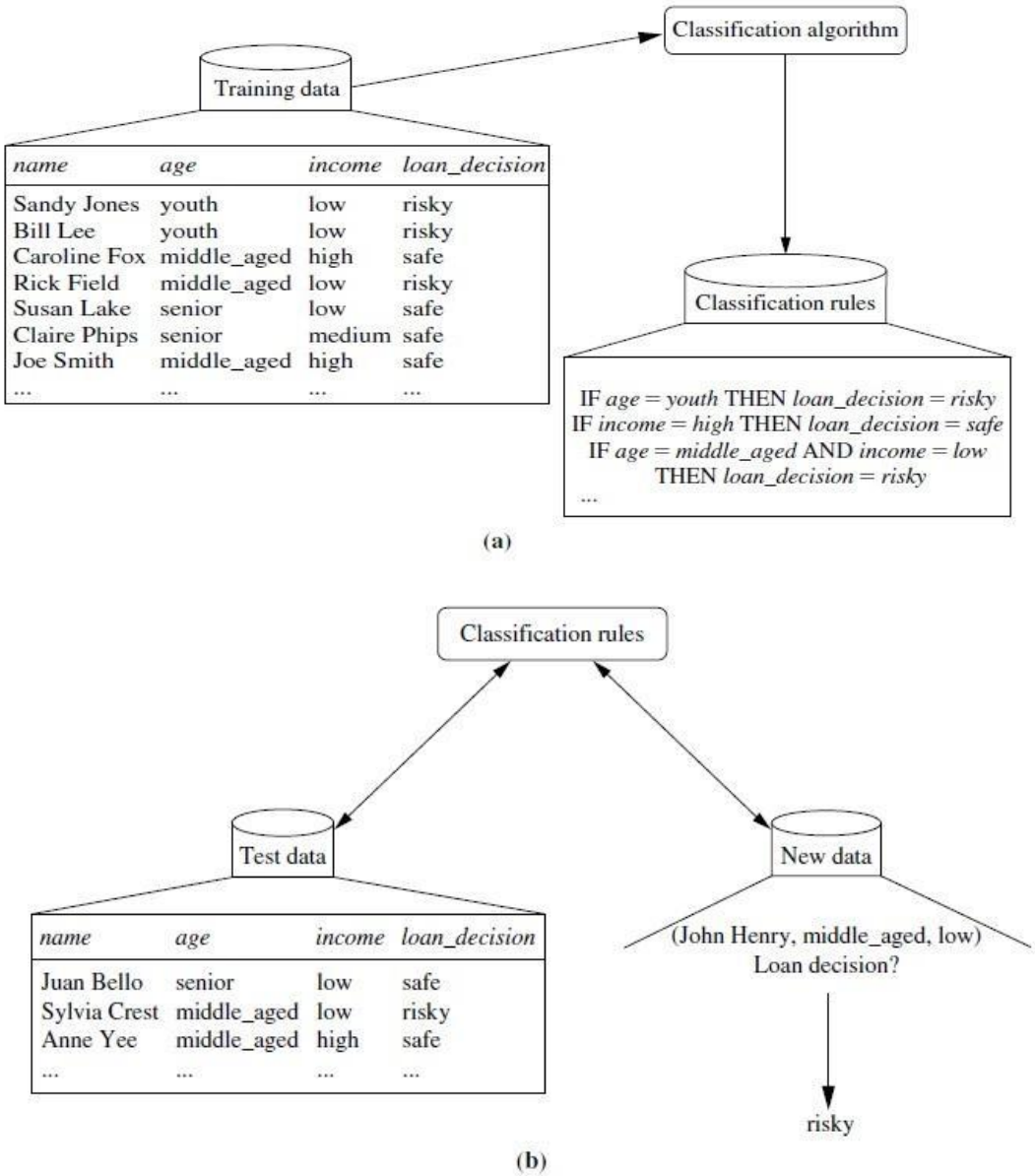

Figure 2- The classification phases of data: Learning:
classification algorithm is used for analyzing training data. The class label is a loan decision, and the learned classifier is viewed as a form of classification rules. (b) Classification: test data are used to determine the classification rules' accuracy [18].

In addition, real valued prediction variables are mapped with elements of learning in regression [1]. The most common technique in this category is linear regression. An estimation analysis of data for doctors (patients' records) or decision support systems [16] are examples of its implementations. A liable indicator 'y' and a standard predictor 'x' are included in the straight-line regression analysis. This technique is the simplest regression method, where 'y' is a linear function of 'x', as shown in equation 1.

$$y = b + wx \quad \ldots\ldots\ldots (1)$$

However, y variance is assumed to be constant. Also, b and w are regression coefficients, which determine respectively the y-intercept and line slope. In addition, the regression coefficients 'w' and 'b' can also be viewed as weights in order to write the equation in the same way as in equation 2 [19]:

$$y = w0 + w1x \quad \ldots\ldots\ldots (2)$$

3.3 Clustering

Clustering is a DM approach that groups abstract or physical objects in classes with similar objects or grouping datasets into several clusters (groups) on the basis of similarities, so that there is considerable relation between them inside specific cluster, whereas the clusters are quite different among each other [10,14,18,20]. Clustering discovers effectively the previous unknown groups in the data Also, it may be employed in the detection of outliers. The outlier values may be far away from any cluster, hence they could be more appealing than the general cases [18,21]. The clustering as an DM approach may be used as an autonomous tool for gaining insights into data distribution in order to perceive the features of every cluster and to concentrate on specific cluster sets for additional analysis. Otherwise, it could be regarded as a pre-processing phase for other algorithms, such as characterization, subset attribute selection, and classification, in order to be implemented in the detected clusters and in the selected attributes or features. Unsupervised learning is known as clustering, because the information about a class label is absent [10]. Clustering is therefore a form of observational learning rather than example learning. Various clustering processes can generate different clusters on the same data set [18]. In considerable applications, such as business intelligence, web search, image pattern recognition, biology, and security, clustering was pervasive [1,3]. Hence, it is complicated to give the clustering approaches a clear categorization, since these categories could be overlapped. Hence, the approach may have features from different categories [18] to provide a specific categorization of clustering methods. Notwithstanding, a fairly organized picture of clustering approaches can be given. The key fundamental clustering approaches can generally be classified, as shown in Table- 1, to the following approaches [17,22]:

Table 1-Clustering methods of data in a general perspective [18]

Method	General Characteristics
Partitioning methods	– Finds mutually exclusive clusters of spherical shape – Distance based – May use mean or medoid (etc.) to represent cluster center – Effective for small to medium size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

1- Hierarchical techniques: these techniques integrate data objects into subgroups, which are incorporated into larger and higher-level groups, and the process continues in this manner to formulate a hierarchical tree. Hierarchical clustering strategies have two main classifications: bottom-up (agglomerative) and top-down (divisive). The agglomerative clustering begins with one-point clusters, and then combines two or more of the clusters recursively. In comparison, divisive clustering is a top-down technique, beginning with a single cluster containing all data points, and then splitting it into suitable subclusters repetitively [10,17,18,22]. A brief description of the various connection-criteria to other types of hierarchical clustering is presented in equations (3-7), where single linkage, complete linkage, average linkage, centroid method, and ward's method are determined respectively [23]:

$$min_i \|a_i - b_i\|. \quad \ldots\ldots\ldots (3)$$

$$max_i \|a_i - b_i\| \quad \ldots\ldots\ldots (4)$$

$$mean_i \|a_i - b_i\| \quad \ldots\ldots\ldots (5)$$

$$\|c_a - c_b\| \quad \ldots\ldots\ldots (6)$$

$$\frac{\sqrt{2m_a - m_b}}{\sqrt{m_{a+m_b}}} \|c_a - c_b\| \quad \ldots\ldots\ldots (7)$$

where a_i and b_i are all objects in cluster a and cluster b, respectively. Whereas, c_a and c_b are the centers of these clusters. Furthermore, m_a and m_b are the number of objects in clusters a and b, respectively.

In fact, the usage of this algorithm is widespread with data that do not contain class label. But, there is a drawback which makes it inaccurate to calculate the true number of clusters due to the manual selection of each cluster number. Therefore, prior knowledge from the physician is an important part in such algorithms to approximate the number of clusters.

2- Partitioning methods: the number of clusters (k) ought to be calculated, where k should be smaller or equal to the number of subjects. However, the partitioning approach constructs k data partitions, and each partition constitutes a cluster. This implies that the data is divided into k groups, so each group should contain at least one object. That is to say, partitioning methods perform partitioning at one level on data sets. Distance-based methods form the majority of partitioning methods. Predetermined k the partitions number to form. Hence, a partitioning approach formulates the initial partitioning by the number of partitions to construct. Thereafter, it utilizes an iterative relocation approach, which tries to enhance the partitioning process by moving objects from a group to another [17, 23]. A good partitioning is usually based on the fact that objects in one cluster are close or related to one another, though objects in different clusters are far apart or very distinct. For subspace clustering, conventional partitioning approaches can be generalized instead of looking for entire data space. If many attributes exist and the data are sparse, this is beneficial. Global optimization is often computationally prohibitive in partitioning-based clustering, which might entail an exhaustive enumeration of all potential partitions. Preferably, many applications endorse common heuristic methods, such as greedy approaches, for instance $k-means$ and $k-medoids$ algorithms, which gradually enhance the clustering efficiency and achieve local optimum. Such heuristic clustering methods are well functioning to identify spherical shaped clusters in databases of small to medium size. Partitioning methods should be expanded to identify clusters of complex shapes and for very broad sets of data [18].

3- Density based methods: Almost all partitioning methods of the cluster structures are based on the distances between objects. These techniques may only discover spherical shaped clusters, while the clusters of arbitrary shapes are hard to be discovered. Moreover, the notion of density has led to the development of other clustering methods. Their common concept is to further grow a certain cluster, under the condition that the "neighborhood" overrides some threshold density (number of data points or objects). For instance, the neighborhood of a determined radius must have a minimum number of points for each data point within a certain cluster. This way, outliers or noise can be filtered off and clusters of arbitrary shapes can be discovered [18, 23]. Hence, a sequence of objects or hierarchy of clusters can be distinguished by density-based methods. In addition, density-based methods may be expanded from full space to subspace clustering [10].

4- Grid based methods: Grid based methods quantize the object space into a limited number of grid-shaped cells. Therefore, all of the clustering processes are carried out in the quantized space (grid

structure). The rapid processing time represents its principal gain. That is, independent of the data objects number, while dependent on the cell number in the quantized space for each dimension [23]. Grids are frequently an effective approach to many problems in spatial data mining, that are comprising clustering. Grid-based approaches can therefore be combined with other clustering techniques, like hierarchical and density-based methods. In addition, there can be clustering requirements for certain applications, which involve the incorporation of various clustering techniques [18].

3.4 Association Rule Mining (Dependency Modelling)

Association Rule Mining is one of DM techniques, classified under unsupervised DM techniques, which strives to discover links or associations between records or items that belong to massive dataset and labels essential dependencies among variables. This technique detects a hidden pattern within the dataset [5,14]. Association rule mining implies the $X \rightarrow Y$ formula, where X and Y are discrete items, or item sets constructing $if - then$ statements concerning attribute values. This rule is popular in market basket analysis, . It is used to analyze clients who purchase such products present insight into consumer combinations which purchase together commonly [1,10].

However, support and confidence are principal measures in association rule mining. The support refers to the relationship degree that appears in the data. While, confidence is about the probability of consequences if a precedent is present. If min_sup percent of transactions support $X \cup Y$, then the rule $X \rightarrow Y$ has minimum support value min_sup. Whereas, the rule $X \rightarrow Y$ holds with minimum confidence value min_conf if min_conf percent of transactions that support X also support Y [24]. For instance, from the transactions stored in supermarkets, an association rule like "$Butter\ and\ Bread \rightarrow Milk$" might be determined by means of the association mining [12].

3.5 Anomaly Detection

The data mining technique of this type involves the identification of data items in the dataset that do not correspond to anticipated behavior or pattern. It can be used in diverse fields, like intrusion, detection, fault detection, or fraud, etc. [1,5,14]. However, a medical dataset could be used with the tumor images to detect the malignant tumor.

3.6 Summarization

It is a result, but not a part, of DM techniques that also deals with the determination of a compact representation for a data subset, synonymously known as a generalization or description, for a subset of data [1,5].

4. Data Mining Algorithms

During the early days of DM, the quantifiable algorithms was used for analyzing data for the purpose of better explanation to the circumstances that to be confronted. The question-specific approaches for DM have also taken account of the value of data obtained by the approach of quantifiable algorithm [25]. The algorithm selection is based on the characteristics of the cleaned and preprocessed data set (width, height, and quality of variable values). Then, the selected DM function and the end user preference are determined. This represents the primary purpose of the study and the algorithm complexity [26]. However, DM algorithms are labeled as supervised or unsupervised. The supervised learning means prediction of a known results of the target through a training set that contains classified data in order to evaluate the inference or classify testing data. There are no reliable outcomes for unsupervised learning, so researchers seek to identify patterns or grouping naturally occurring in unlisted data [27]. Prediction, modeling, and inference are the analytical objectives for the medical data. Nevertheless, in these contexts, the classification, clustering, and regression are conventional methods [13]. Such examples of DM algorithms are explored in the following sections:

4.1 $K - means$: A Centroid Based Technique

Suppose a data set D that includes n objects of Euclidean space. Hence, the partitioning methods diverse the objects in D into k clusters $C1, \ldots, Ck$, that is, $Ci \subset D$ and $Ci \cap Cj = \emptyset$ for $(1 \leq i, j \leq k)$. In order to evaluate the partitioning quality, an objective function has been utilized, in such a way that the cluster objects are similar to each other, though they are dissimilar to objects of other clusters. The analytical feature is used to measure partition consistency. This feature is designed to achieve high intracluster similarity and low intercluster similarity. In order to represent this cluster, a centroid based partitioning technique employs the C_i centroid of a cluster. Theoretically, the cluster center point is its centroid. Correspondingly, the centroid can be described in many ways, such as by means of medoid of the objects or points appointed to the cluster [22]. The distinction between an

object $p \in C_i$ and the representative of the cluster c_i is determined by *dist* (p, c_i), where *dist* (x,y) is the Euclidean distance between two points x and y. The cluster quality C_i can be determined by cluster variation, where it is the squared error total between all objects in C_i and the centroid C_i , as described by equation 8 [18,28]:

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} dist(p, c_i)^2 \quad \ldots\ldots\ldots (8)$$

where E is the squared error total for all objects in the dataset, p is the space point of a particular object, and c_i is the centroid of cluster c_i. Also, p and c_i are both multidimensional. That is to say, the distance from the object to the center of its cluster is squared, and the distances are summed for each point in each cluster. The resulting $k - clusters$ are rendered as separate and as compact as possible by this objective method. The $k - means$ algorithm identifies a cluster centroid as the average value of points within the cluster. It is arranged as in the following order. First, the random selection of k from the objects in D, each representing a mean or center of the cluster. The cluster is allocated for each of the other objects that is most similar, and on the basis of the Euclidean distance between the object and the mean of the cluster. The $k - means$ algorithm enhances the variance within-cluster. It calculates the new mean for each cluster with the objects assigned to it in the preceding iteration. Hence, all objects are reassigned to the new cluster centers with the modified means. The iterations remain so long till the assignment is steady, which is to say that the clusters created in the current cycle are the same as in the preceding cycle, as represented in Figure- 3. Also, the procedure of k-means is outlined in algorithm 1 [18,23,28].

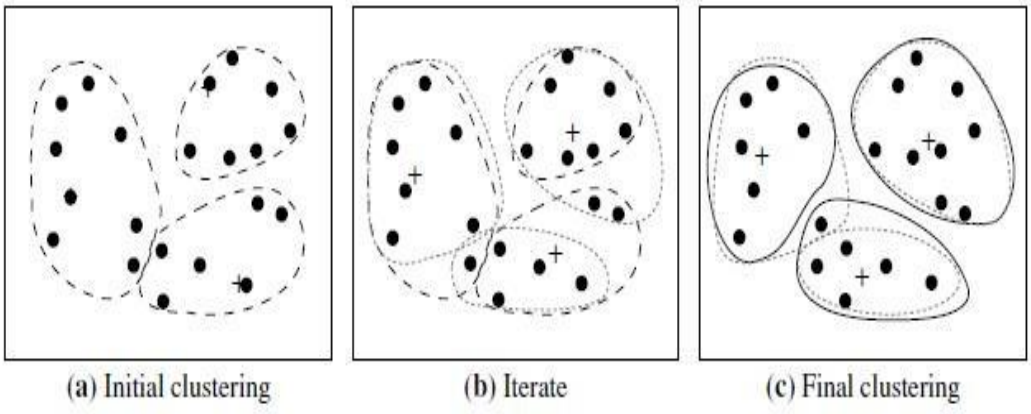

Figure 3- k-means method (mean of each cluster is marked by a +) [18]

Algorithm 1: The $k - means$ algorithm for partitioning, where the mean value of the objects in the cluster represents each cluster center.

Input:

- k: Number of clusters,
- D: Data set involving n objects.

Output: A set of k clusters.

Method:

(1) Choose k objects arbitrarily from D as initial cluster centers;
(2) R**epeat**
(3) Reassign each object to its cluster, to which the object is most similar, according to mean value of the
objects in the cluster;
(4) Update the cluster means, by calculating the mean value of the objects for each cluster;
(5) **Until** no change;

This approach is successful generally in most areas , but particularly in the medical field, because it is easy to be understood and applied on data, in addition to the addition of the silhouette method as an improvement to detect the number of clusters automatically.

4.2 Apriori Algorithm

The algorithm name is based on a knowledge of frequent itemset properties that the algorithm used previously. An iterative approach, called a level-wise search, is used by Apriori, where $k-itemsets$ are employed in order to elaborate the $(k+1)-itemsets$. Firstly, by scanning the database to collect the count per item, and accumulating the items that satisfy minimal support, the collection of frequent $1-itemsets$ is found. The set is labelled with $L1$. The next step is to use $L1$ in order to determine $L2$, a regular collection of $2-itemsets$ that are utilized to find $L3$, and so forth, until no further frequent $k-itemsets$ are identified. Each L_k must be determined with a complete database scan. Hence, a significant property, termed Apriori property, is exploited to diminish the search space [18], with a view to improve the efficiency of level reasonable generation of frequent itemsets. Further, Apriori implies that all frequent itemset nonempty subsets ought to be frequent. On the subsequent observation, the property Apriori supposes that if an $itemset\ I$ does not meet the minimum standard threshold (min_sup), then I is not frequent; in other words, $P(I) < min\_sup$. On the other hand, if an $item\ A$ is added to the $itemset\ I$, then the resulting itemset, i.e. $I \cup A$, cannot occur more frequently than I. Therefore, $I \cup A$ is not frequent either, that is, $P(I \cup A) < min\_sup$ [24]. So, the clinical data could be helpful in a particular disease prediction, such as for the frequent appearance of symptoms or biological tests associated with a disease.

4.3 Decision Tree (DT)

Decision tree is a supervised classification algorithm. The criterion employed in DT is generally the gain ratio or information gain. However, the gain information is an entropy change of information, when information state is changed [9,10]. C may be the class values $\{c_1, c_2, \dots, c_n\}$ and A is the attribute of values $\{a_1, a_2, \dots, a_k\}$, while $H(C)$ is the entropy of C attribute. Also, $H(C|A)$ is the conditional entropy, which indicates the entropy of C if the state of attribute A is well-known. Then, the information gain, $I(C, A)$, is computed as in equation 9 [29,30]:

$$I(C, A) = H(C) - H(C|A). \quad \dots\dots(9)$$

The entropy of attribute, $H(C)$, is determined as in equation 10:

$$H(C) = -\sum_{i=1}^{n} p(C = c_i) \log_2(p(C = c_i)) \quad \dots\dots(10)$$

Whereas, $P(C = c_n)$ is the relative frequency of class value c_n. Also, the conditional entropy, $H(C|A)$, is determined as in equation 11:

$$H(C|A) = -\sum_{j=1}^{k} p(A = a_j)\, H(C|A = a_j) \quad \dots\dots(11)$$

The information gain prefers an attribute of higher number of values. For the purpose of avoiding this, the gain ratio could be employed, which is determined as shown in equation 12:

$$\frac{I(C, A)}{H(A)} \quad \dots\dots(12)$$

Whereas, the attribute entropy, $H(A)$, is determined as in equation 13:

$$H(A) = -\sum_{j=1}^{k} P(A = a_j) \log_2 \left(P(A = a_j)\right) \quad \dots\dots(13)$$

The tree algorithm is specifically designed for Iterative Dichotomiser3 (ID3), C4.5 algorithm, and Classification and Regression Tree (CART), and is usable in several ways [2]. However, ID3 is one of the most important algorithms of DT. In this manner, data gains ahead of time. In addition, in order to decide appropriate properties for each resulting DT hub, the property of most impressive data might be selected as test property, which depends on the current node [15,30]. Additionally, Random forest algorithm (RF), which classifies enormous amount of data with high accuracy, is one of the best classification algorithms [17]. Figure- 4 shows the conceptual structure of DT.

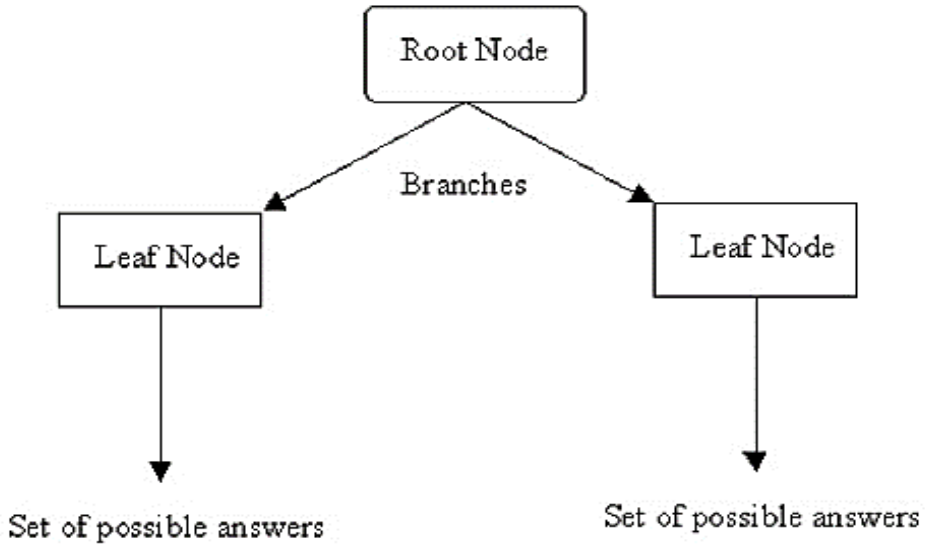

Figure 4-Structure of decision tree [9].

As a result, it is preferred to use the information gain principle, which is a successful algorithm for medical data, especially the balanced class data. Also, the algorithm is not affected by the bias of feature values.

4.4 Naïve Bayesian (NB)

The Bayes' theorem was named after Thomas Bayes, an English non-conformist clergyman, who worked early on probability and decision theory in the 18th century. Suppose X be a tuple of data. X is considered as " evidence " in Bayesian terms. Same as always, measurements made in relation to a set of n attributes are explored. Suppose that H is a hypothesis of the tuple of data X that relates to a given class C. Moreover, $P(H|X)$ will assess the probability of the hypothesis H getting the evidence or the investigated data tuple X found for classification problems. That is to say, it is about the probability of tuple X belonging to class C, because the description of the attribute of X is known. Whereas, $P(H|X)$ is an $H-conditioned$ posteriori probability on X. Also, $P(H)$ is the prior probability of H, where $P(X|H)$ is the posterior probability of X conditioned on H, and $P(X)$ is the prior probability of X [9,18]. The Bayes' theorem is valuable since it offers a way to measure the posterior probability, $P(H|X) from\, P(H), P(X|H),$ and $P(X)$, as represented in equation 14 [9,15,18].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad \ldots\ldots\ldots (14)$$

However, the Bayes' theorem has been employed in the NB classifier, which is used in supervised learning. It acts as in the following [15]:

1. Suppose D be a set of tuples and their corresponding class labels. Normally, the $n-dimensional$ attribute vector is defined for each tuple. Also, $X=(x_1, x_2, \ldots, x_n)$, which shows n measurements made on the tuple of n attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. Suppose m classes, $C_1, C_2, \ldots, C_m$, Provided the tuple X, the classifier predicts that X belongs to the class of the highest posterior probability conditioned on X. That is to say, the NB classifier anticipates that tuple X belongs to the class C_i if and only if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i$. Hence, $P(C_i|X)$ has to be maximized. The class C_i for which $P(C_i|X)$ is maximized is named the maximum posteriori hypothesis. According to Bayes' theorem, the $P(C_i|X)$ in equation 14 is determined in equation 15 [9, 18]:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad \ldots\ldots\ldots (15)$$

In the medical domain, the features are usually related to each other, such as patients' symptoms and disease. Since Naïve Bayes Classifier treats the features as independent of each other, it is rarely used in such a field.

4.5 k- Nearest Neighbors (KNN)

The algorithm $k-$Nearest Neighbors is intended to determine the nearest point for the object observed [2]. An object is categorized by a majority vote from its neighbors, in addition to the specified object. Further, k is a positive integer, typically small, and is assigned to the most common group among its k closest neighbors. However, if $k = 1$, then the object is simply appointed to the next neighbor class (single nearest neighbor) [9]. Nevertheless, the research in this subject and the task of classification are incomprehensible. It is an approach of supervised classification that is commonly used. Whereas, the arrangement procedure is easy to implement and the training is fast [10]. Hence, KNN uses the feature of distance function, as in the following equations (16-18) [15,21,27]. Thus, KNN is employed in several applications of diverse areas like health datasets, cluster analysis, image field, online marketing, and pattern recognition. Figure- 5 represents the conceptual description of KDD technique [17,21].

$$\sqrt{\sum_{i=1}^{k}(xi - yi)^2} \quad \text{Euclidean} \quad \ldots\ldots\ldots (16)$$

$$\sum_{i=1}^{k} |xi - yi| \quad \text{Manhattan} \quad \ldots\ldots\ldots (17)$$

$$[\sum_{i=1}^{k}(|xi - yi|)^{q}]^{1/q} \quad \text{Minkowski} \quad \ldots\ldots\ldots (18)$$

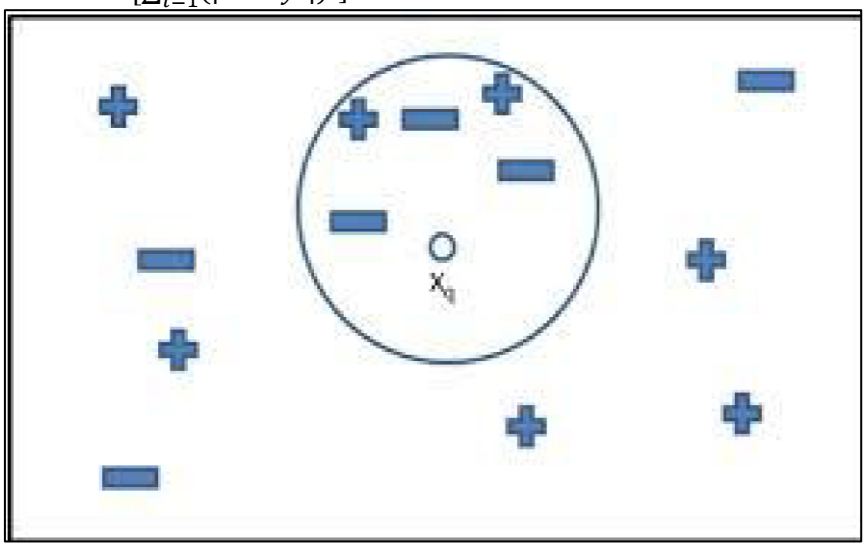

Figure 5-Conceptual description of KNN technique [9]

As long as the medical field needs large amount of data, KNN will demand more computations. Consequently, this requires more time and cost.

4.6 Artificial Neural Networks (ANN)

Neural networks is an artificial intelligence-based method. In ANN, neurons or nodes are the fundamental elements. Hence, the neurons become interconnected and operate in parallel within the network in order to generate the output functions [30]. Through current investigations, even in cases where neurons or nodes in the network fail or fall due to their failure to operate in parallel, they are capable of generating new observations. A growing neuron is assigned an activation number, while a weight is assigned for each edge within the neural network [26]. It is a common knowledge that ANN is mainly used to perform classification and pattern recognition tasks. Hence, ANN is based on biological neural networks in the human brain and represented by its neuron, a cell that processes information in the human brain, as a connectionist model [15,31]. Indeed, there are two different branches in the neuronal cell body that contains the nucleus, the axon and the dendrites. The dendrites receive incoming signals or impulses of other neurons. While, the axon transmits signals or impulses to other neurons. Hence, all neurons are linked and transmitted via the short pulse trains. Anyway, the nodes are artificial neurons and the directional edges are considered as a connection between the

output and input neurons. Hence, the internal weights of neural networks would be adapted during the training phase in accordance to the employed transactions in the learning process. Also, the neural network receives the predicted output for every training transaction, which allows to modify the weight. Figure- 6 presents the structure of ANN [9].

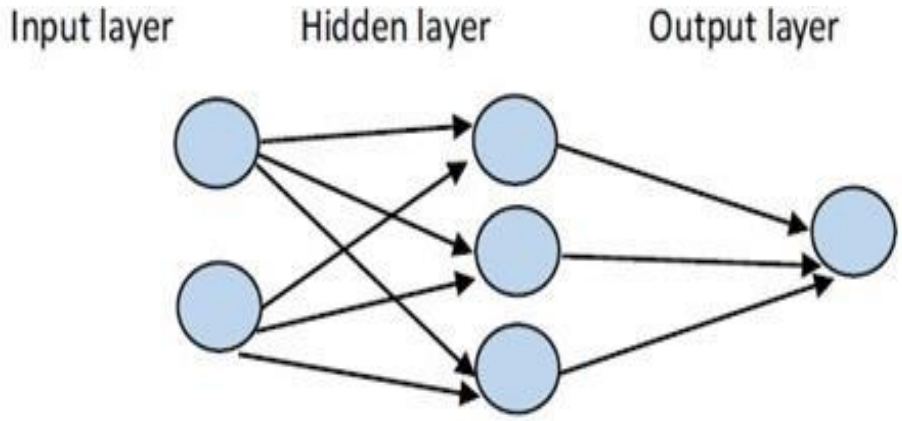

Figure 7- Artificial neural network structure [15].

Accordingly, ANN is the most common data modeling algorithm, which has been utilized in clinical medicine, such as classification of diabetic disease cases.

4.7 Genetic Algorithms (GAs)

Genetic algorithms are graded as evolutionary, stochastic, and those that provide an optimal solution [2]. Some specified data structures are important to genetic algorithms. They work on a categorically represented population of characteristics. However, the comparison with genetics is that the population (genes) are made of characteristics (alleles). Hence, the method of using genetic algorithms is to use operators (reproduction, crossover, selection) with the mutation property for improving the generation of probably superior combinations [17]. Therefore, it is constructed of randomly selecting parents that reproduce over crossover, where the reproduction is the operator that selects which individual entity may survive; that is, to decide its survival, it requires some objective features or selection characteristics. Meanwhile, the crossover is associated with entity changes in future generations. Also, survivors are selected through fitness features for the next generation. In addition, the mutation is the process through which, for future operations, randomly selected attributes of randomly selected entities are modified. Eventually, the process continues until a preset number of iterations that exceeds either a certain fitness level is obtained. Moreover, GA will map data to the shape of features that have discrete values. This will lead to loss of information when treating continuous values [12].

4.8 The *k*-Fold Cross Validation

This is a statistical method employed for predictive model performance in unknown dataset [12]. The full dataset is randomly split into k mutually exclusive subsets of approximately the same size in $k-fold$ cross validation, also known as the rotation estimation. The classification model is trained and tested for k times. It is trained on all folds except one each time, then tested on one remaining fold, as illustrated in Figure- 8 [18,31]. Simply, the overall accuracy of the model is determined based on cross validation, that is by averaging the k individual accuracy measures, which is defined in equation 22 [12].

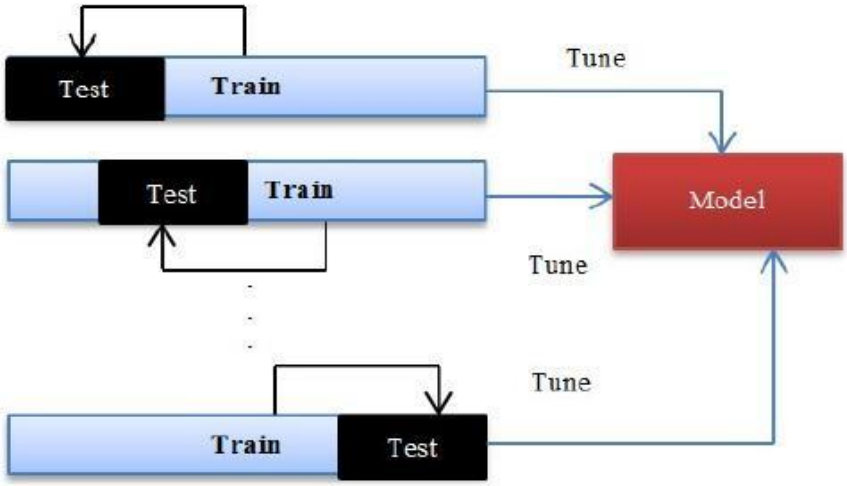

Figure 8- *k*-fold cross validation architecture [31].

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (22)$$

where CVA is cross validation accuracy, the number of folds used is k, and the accuracy measure is A, for example, the hit-rate, sensitivity, specificities, etc. of each fold. Because the accuracy of cross-validation will depend on the random assignment in K distinct folds of the individual cases, it is common practice to range the folds themselves. The folds are generated to an approximately equal proportion of predictor labels in a range of k-fold cross-validations, i.e. classes like in the original dataset. The experimental researches have shown that range cross-ventilation produces comparable results with lower bias and lower variance than regular cross-validation. Frequently, $K - fold$ cross-validation is termed $10 - fold$ cross-validation, since the k obtaining the value of 10 is the most ordinary practice. Actually, observational studies have demonstrated that 10 tends to be an optimum number of folds, which optimizes time to complete the test as well as the bias and variance related to the validation process [12,18,31]. In the case of $k - fold$, cross validation is represented in Figure- 9.

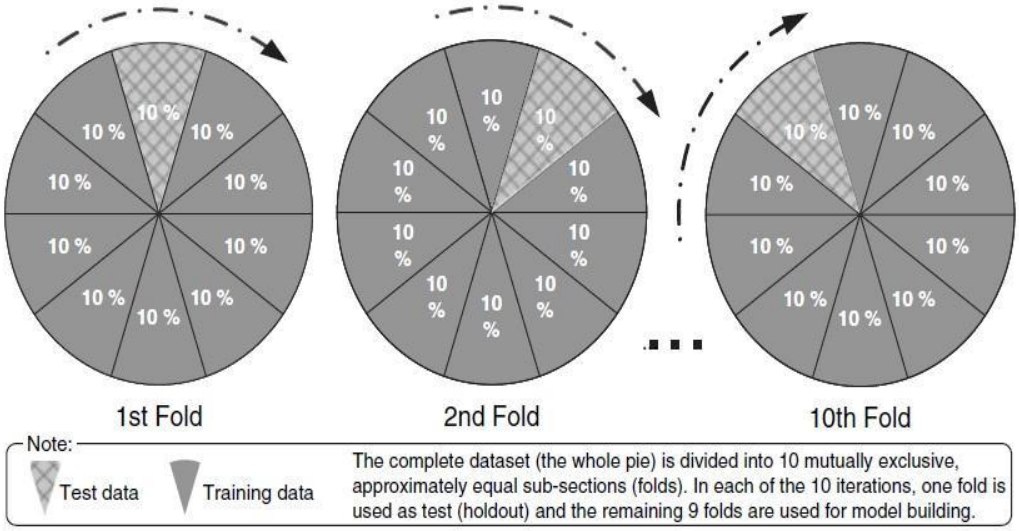

Figure 9- 10-fold cross-validation representation [12].

The cross-validation technique step by step details are described as follows [12]:
Step 1: The total dataset is divided into k disjoint subsets randomly, i.e. folds, each including almost the same number of records. The class labels apply the sampling to guarantee that the proportional

representation of the classes to the initial date set is approximately the same to that in the original dataset.

Step 2: For every fold, all records, except those in the current fold, are built with the classifier. The classifier is then checked on the current fold to provide an estimation of its error rate for cross-validation. Hence, it registers the result.

Step 3: Actually, the ten cross validation evaluates, after repeating step 2 for all 10 folds, are averaged to supply the estimation of aggregated classification accuracy for each model. Note that $10-fold$ cross-validation involves no further data than the conventional single split 2/3, 1/3 test experiment. In addition, $k-fold$ experimental methods are recommended for comparative studies with relatively smaller datasets. In fact, the principal benefit of $10-fold$ cross-validation of a number of folds is that they minimize the bias associated with a random sampling, and holdout data samples by repeating the experiment 10 times with a discrete portion of the data used as a holdout sample. In comparison to only once repetition, the downside of this technique is the need to conduct the training and test k times [12].

4.9 Performance Metrics for Predictive Modeling

For classification issues, a coincidence matrix (contingency table or classification matrix) is the main source of performance measurements. Hence, Figure- 10 explores a coincidence matrix for two-class classification issues. The numbers from upper left to lower right in the diagonal stand for the correct decisions made, and those outside the diagonal reflect the errors. The true positive rate of the classifier is determined by dividing the correctly classified positives (the true positive count) by the total positive count, also called hit rate, recall, and sensitivity. The true negative rate, also known as the specificity rate, of a classifier is calculated by dividing the correctly classified negatives counts (true negative count). The total accuracy of the classifier is calculated by dividing the total positives and negatives by the total number of samples, known as recognition rate. Also, the accuracy is known as the estimated detection rate. Furthermore, performance measures, such as the error rate, which is known as the misclassification rate, and the accuracy that may be considered as a measure of exactness, $f-measure$, e.g. area under the Receiver Operating Characteristic (ROC) curves, are also utilized in measuring other aggregated performance measures. However, the quality assessment is estimated in the following equations (23 - 28) [15,18].

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Figure 10- Classification matrix of two-class issue [12].

$$true\ positive\ rate = \frac{TP}{TP+FN} \quad \text{..... (23)}$$

$$true\ negative\ rate = \frac{TN}{TN+FP} \quad \text{.... .... (24)}$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (25)}$$

$$error\ rate = \frac{FP+FN}{TP+TN+FP+FN} \quad \text{..... (26)}$$

$$precision = \frac{TP}{TP+FP} \quad \text{.....(27)}$$

$$f - measure = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \quad \text{..... (28)}$$

4.10 Support Vector Machines (*SVMs*)

The Support Vector Machines are considered as supervised learning methods, which produce functions for input-output mapping through a set of training data [17]. The mapping function may be either an input data categorization function or a regression function, which is utilized for desired output estimation [32]. However, for classification, non-linear kernel functions are commonly employed to convert input data, which are inherently representing highly complex nonlinear relationships into high dimensional feature space, where the input data become more segregated, that is, the mean linearly is separated in comparison to initial the input space [12,15]. The classification of the supporting vector investigates for an optimal separating surface, known as hyperplane, that is equal to each of classes, as presented in Figure-11 [17,27]. Therefore, *SVM* concentrates on data division into two classes, namely P and N, which relate with the cases $y_i = +1$ and $y_i = -1$, respectively. Moreover, when training data are linearly separable, then a pair (w, b) exists, as shown in equations 29 and 30 [10,19].

$$w^T x_i + b \geq 1 \text{ for all } x_i \in P \quad \text{..... (29)}$$

$$w^T x_i + b \leq -1 \text{ for all } x_i \in N \quad \text{..... (30)}$$

Hence, w is a weight vector and b is a bias. Whereas, the prediction rule is obtained from equation 31 [19].

$$f = sign\ (\langle w.x \rangle + b) \quad \text{..... (31)}$$

Furthermore, SVMs have demonstrated highly competitive performance in many practical applications, including bioinformatics, medical diagnoses, image processing, face recognition, and text mining. Alongside their strong mathematical foundation for statistical learning theory, they have made SVMs one of the most common and state of the art tools for knowledge discovery and data mining [12].

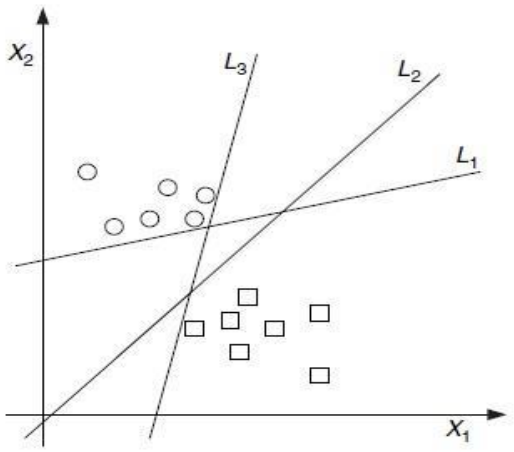

Figure 11-Separating the data using linear classifiers (hyperplanes) [12].

5. Data Mining Tools

Almost all DM tools are dataflow architectures based on software applications. DM tools facilitate choosing the right algorithm occasionally. These tools are graphical integrated environments, such as

KNIME, RapidMiner, Orange, Weka, and Tanagra, which enable the visual component's placement, connection, and dragging. There are different tools, for example. R and Scikit-learn, that are essentially extensions of the language underlying in the format of specialized packages and/or add-ons for the Graphical User Interface (GUI) [33,34]. However, they will be defined in the subsequent sections and represented in Table- 2.

5.1 RapidMiner

RapidMiner is an interactive user environment for machine learning and data mining processes [34]. Earlier RapidMiner environments include Rapid-I, YALE, which is a mature, Java-based, general DM tool from RapidMiner, Germany [35]. In RapidMiner processes, R and Python could be integrated [33]. RapidMiner totally concentrates on processes that may include sub-processes. The processes include visual components operators. DM algorithms, data sources, and data representation are applied by these operators. On the other hand, the data flow is created by dragging and dropping the operators and connecting the inputs and outputs of respective operators. In addition, RapidMiner provides the application wizards option, which automatically creates the process according to project objectives, such as sentiment analysis and direct marketing [14]. Despite the fact that RapidMiner has a basic set of operators, which are quite effective, its extensions make it actually more valuable. However, the operator sets for text mining, web mining, and time series analysis are common extensions. There is currently limited support for the use of inductive programming algorithms, deep learning approaches, and some of the more advanced machine learning algorithms, like extremely randomized trees. Nevertheless, the Hadoop cluster (Radoop) supports big-data analysis, which includes medical data analysis [32,33].

5.2 Weka

Weka can be considered as an open source DM-platform, which was initially evolved at the University of Waikato, New Zealand, with a toolkit to machine learning and DM [12,33,35]. It includes a large set of state-of-the-art machine learning, along with DM algorithms that are Java-written. Weka includes several tools for regression, classification, clustering, association rules, visualization, and data pre-processing software [14,34]. Additionally, Weka offers four DM features: Command Line Interface (CLI), Explorer, Experimenter, and Knowledge flow. Hence, Weka has four DM options. The Explorer is the preferred option to define the data source and prepare data, machine learning algorithms, and visualization [36]. The Experimenter is primarily used to compare the performance of different algorithms on the same dataset. Whereas, Knowledge flow is the same as the RapidMiner operator model, which enables the data flow definition using properly connected visual components. Furthermore, Weka supports various model evaluation procedures and metrics, but lacks multiple data surveys and visualization methods [32]. This needs the storage of data files in uncommon Attribute Relation File Format (ARFF), despite it reads some Comma Separated Values (CSV) files with some issues [33].

5.3 Orange

Orange is a Python based tool for DM that is built at the Bioinformatics Laboratory of the University of Ljubljana, Faculty of Computing and Information Science [34]. It can be used either via Python scripting as a Python plug-in, or via visual programming [32]. Orange computer-intensive parts are written in C++, and its upper layers are built in Python [35]. Its visual software development interface, Orange canvas, provides a structured view of supported features in 8 categories: data processes, visualization, classification, regression, evaluation, unsupervised learning, association, and prototype implementations. However, functionalities are visually defined by various widgets, such as reading file, discretionary, SVM classifier training, etc. Additionally, software development takes place through the use of widgets on canvas, and connecting their inputs and outputs. The interface is highly polished, desirable, and optically pleasing to the user [14]. An obvious downside to Orange is that the number of widgets available seems limited, particularly because of the lack of integration with Weka, compared to other tools like RapidMiner or KNIME. Nevertheless, traditional DM techniques are very well known [32].

5.4 Konstanz Information Miner (KNIME)

The DM platform of KNIME was developed and operated by the Swiss company KNIME.com AG within the Eclipse development environment of IBM [14, 34,35]. The first version was released in 2006 as an open-source [32,35] and began in 2004 at the Constanz University in Germany. However, KNIME is written in Java and can expand its library to include Weka supplied builtin supervised and

unsupervised DM algorithms. Visual programming of KNIME is structured like a data flow. Also, dragging nodes from the repository node to the central part of the benchmark is used by the software programs. Every node is documented in detail, so that the documentation will automatically be shown within the interface when the node is selected. Hence, each node performs some function, including reading data, filtering, modeling, visualization, or similar functions. Nodes have ports of input and output. Most ports send and receive data, whereas others manage models of data, such as classification trees. The integration with Weka and R is one of KNIME's greatest strengths. Furthermore, Weka integration allows to use almost all applications available as KNIME nodes in Weka. Whereas, R integration allows to operate R code as a step through the workflow, and to open R views and learning models in R. There are also several other important free extensions, such as JFreeChart extension that enables advanced charting, OpenStreetMap extension that enables working with geographical data, etc. [32].

5.5 Tanagra

Tanagra was created by Ricco Rakotomalala at the Lumière University, Lyon 2, France, as a free suite of machine learning tools for research and academic purposes [34]. Tanagra is designed around a graphical user interface, hence data processing and analysis components are arranged in a tree-like structure, so the parents pass the data on to their children. Moreover, machine Tanagra data analysis components present their outputs in a satisfactorily formatted HTML [32]. Also, Tanagra supports many typical DM tasks, including visualization, statistic description, instance selection, feature selection, feature construction, factor analysis, regression, clustering, classification, and association rule learning. In addition, Tanagra makes a good balance with machine learning techniques (NN, SVM, DT, RF) and between statistical approaches, such as parametric and nonparametric statistical tests and multivariate analyses (factor analysis, correspondence analysis, cluster analysis, regression) [34].

5.6 R R is considered as a strong choice for DM tasks. Also, it is a statistic computing open source language. Its source code is developed in C, C++, FORTRAN and then interfaced to R. The Bell Labs developed the scripting language initially. However, the main language is expanded by an abundance of packages, which are used for all forms of computational tasks [32,33]. The favored interface to R is its scripting command line. The scripting interfaces have discrete benefits that include data analysis procedure, which is clearly specified and could be saved for subsequent reuse. Nevertheless, the other side of the coin is its scripting, which requires certain competences in programming. Hence, without these components, the users can use R via extensions with GUIs. Rattle is an R interface extension that has been deployed as R library and provides a GUI to several R data analysis and modeling functions [32]. Under the DM perspective users, R provides very rapid accomplishments of several learning algorithms, in comparison to RapidMiner and Weka, from which a wide variety of algorithms are taken, in addition to the complete prospects of methods of statistical data visualization. This incorporates particular types of data for big data management, data streams, supports parallelism, web mining, spatial mining, graph mining, and vastly other sophisticated tasks, comprising little support for deep learning methodologies. Also, C++, Python, and other programming languages can interface well with R [33]. Also, RStudio is an IDE Integrated Environment (IDE) for R [35].

5.7 Scikit-learn

Scikit-learn is a gratis Python package that improves NumPy and SciPy module functionality with various DM algorithms. The matplotlib software package is also used for drawing charts. The INRIA and Google Code Summer project endorse the package [37]. Also, it offers a perfectly written online documentation for all applied algorithms, which is one of its major strengths. This documentation is mandatory for every contributor and is valued over many poorly documented implementations of algorithms [32]. Most main DM algorithms are provided by this package. It is also pretty quick, although it is written in an interpreted language [37]. Notwithstanding its benefits, but Scikit-learn demands one to be a skilled CLI programmer in Python [32].

Table 2- Data mining platforms description [14,32].

S.No	Tool	Year and Author	Company/ Organization	Availability	Core Area	Focused on	Programming language
1	Rapid	2004,	RapidMiner,	Open	Data	Data	Java

	Miner	Ingo Mierswa and Ralf Klinkenber-g	Germany	Source	mining	science, machine learning, predictive analytics	
2	Weka	1993, University of Waikato	New Zealand	Open source	Data mining	Machine learning, data analysis, data visualization	Java
3	Orange	1997	University of Ljubljana	Open source	Data mining	Machine learning, data analysis, data visualization	C++, Python, Qt framew.
4	KNIME	January 2004, KNIME. co-m AG	University of Konstanz	Open source	Data mining	Enterprise reporting, business intelligence, deep learning, text mining, data analysis	Java
5	Tanagra	2003, France	Lumière University Lyon 2	Open source	Data analysis, machine learning, databases area	Some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and non-parametric statistics, association rule, feature selection and construction algorithms	Java
6	R	1993,Ross -Ihaka and Robert	University of Auckland, New Zealand, and is	Open source	Statistical computing	Linear and nonlinear modeling, classical	C, Fortran, R

		Gentleman	currently developed by the R Development Core Team			statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions	
7	Scikit-learn	2007, David Cournapeau and Matthieu Brucher	INRIA and Google Summer of Code project	Open source	machine learning	Classification, regression, clustering and dimensionality reduction	Python+ NumPy+ SciPy+ Matplotlib

6. State of Art

The scientists have many subjects for the DM method, hence [38] suggested to optimize Singular Value Decomposition (SVD) entropy, which is a recent, unsupervised feature filtering. The features were ranked corresponding to the Entropy values (CE). Thus, it was implemented in four means. First, Simple Ranking (SR); second, Forward Selection by accumulation of features so that the set produces the highest entropy (FS1); third, Forward Selection by accumulating features by selecting the best CE out of the remaining ones (FS2); fourth, Backward Elimination (BE) of lowest CE. Accordingly, three biological indices were used and their utility was tested, which include the Viruses Dataset of MLL dataset, the MLL dataset, and the leukemia dataset. Consequently, the quality of the data clustering in the selected feature spaces is evaluated in each case. Also, the Extended Naïve Bayes classifier (ENB), described earlier [39], suggested to deal with mixed data as a conventional NB that only handles categorical data. Five datasets were used to compare the efficiency of the proposed system with further algorithms, like Multiplayer Perceptions (MLP), CART, and DT. The used five datasets are the real mixed datasets collected (Australian credit approval, German credit data, hepatitis, horse colic, and UCI breast cancer repository). The Average Reciprocal Rank (ARR) was utilized as estimation metric in this analysis. In contrast to other algorithms, ENB has the highest ARR score. Whereas, another work [36] applied the C4.5 algorithm, J48 in Weka, SVM, and Sequential Minimum Optimization (SMO) in Weka for prediction and examination of the best possible classification of anemia using a complete blood count (CBC) dataset collected from CBC test reports. The dataset was preprocessed from missing values and duplicates. Also, the attribute selection method was used to select relevant attributes and remove redundant and/or irrelevant attributes. An evaluation was performed by the utilization of in Weka, test options, Mode1 cross validation 2-fold, Mode2 cross validation 4- fold, and Mode3 cross validation 10-fold. It showed accuracy along with correct f-measure results using various test options. C4.5 algorithms classify anemia more accurately compared with SVM. However, in a previous study [40], the modified $k-means$ algorithm was presented for the usage of temporal views for the purpose of producing yearly and monthly frequent patterns of diseases, through the construction of a prototype application and by using the medical data collected from the Reputed Private Hospital. The reports encompass the 2012 and 2013 Electronic Health Records (EHRs). Likewise, in another study [28], an improved $k-means$ was submitted, which used a greedy approach to generate the preliminary centroids and took k or lesser passes. It was expanded to data from different providers of healthcare services. Thus, they compared the traditional *k*-means and selected the enhanced one via the f-measure complexity to measure the accuracy of test results in order to equate not only the results accuracy, but also the efficiency of both algorithms. Furthermore, a previous investigation [34] exploited the Indian Liver Patient Data Set (ILPD) by using three

classification algorithms (DT, KNN, and NB) to classify people with and without Liver Disorder. Five DM methods were used (Weka, RapidMiner, Tanagra, Orange, and KNIME). The classification algorithm output was evaluated with the confusion matrix. Accordingly, it appeared that from all three classification algorithms, the KNIME method predicts greater accuracy. Additionally, DT and KNN were more reliable than NB. Whereas, in another work [41], the logistic regression, the neural network (NN), DT, and KNN predictive models were developed. The authors used test data from Multiparameter Intelligent Monitoring Intensive Care II (MIMICII) physiology patients for predicting death within the next 24 hours. They reported good results with NN and logistical regression with radial kernel models, where the configuring parameters played a key role in model success. For each approach in the test set, they used performance metrics. Logistic regression approaches are more likely to predict which patients died (true positive) in hospital, while NN is doing better predicting those who leave the hospital a live (true negative). Alongside, the Self-Organizing Map (SOM), Principal Component Analysis (PCA), and NN have therefore established an intelligence system with the clustering, noise reduction, and classification approaches. These were implemented on the Pima Indian Diabetes (PID) dataset from University of California and Irvine (UCI) to enhance the predictive accuracy of diabetes. The evaluation of the predictive model was performed on 10-fold cross validation. In addition, the UCI machine learning data repository was compiled on breast cancer, diabetes, heart disease, tuberculosis (TB), and liver disease [42]. The other source of datasets includes data from Amana Hospital in Dares-salaam, Tanzania for human immunodeficiency virus (HIV). Accordingly, the authors created an eight separate prediction classifiers method for disease prediction. The classifiers are: NB, J48, Instance Based Learning (IBK), SMO, MLP, DT Reduced Error Pruning (REP) Tree, Projective Adaptive Resonance Theory (PART), and RF. Such research algorithms would increase the classification accuracy of the merged hybrid classifiers. Two experiments were also established for evaluating the electing performance. The first was a learning experiment with Wisconsin Breast Cancer Dataset, which analyzed the selected performance of eight learning algorithms. Better performance was obtained with combinations of SMO+RF+IBK, and SMO+RF+MLP. While, HIV dataset was used in the second experiment. The SMO + J48 + MLP combination delivered better results. As a test method, the researchers used $10-fold$ cross validation, along with the confusion matrix as performance metric of training and testing data. In another study [43], the authors suggested three DM approaches, namely Apriori algorithm, association rule mining, and NB, so that the output of each algorithm is employed as an input data for the execution of the next algorithm. The dataset was collected from a general hygiene survey form, developed and distributed by students for a sample of 200 students in two secondary schools in Baghdad. The statistics represented general characteristics of environmental health. The data was encoded with the Weka DM tool and then analyzed. The recommendations were for general hygiene researchers to effectively and logically follow this approach, as it was uncovering unknown relationships and correlations among investigated attributes. Moreover, the authors of another work [17] explored the gene document relationships by utilizing Apache Hadoop MapReduce, which is an open source distributed data processing platform. They aimed to find textual patterns from a big dataset of medical document used for obtaining quality gene disease information. This large set of medical records was obtained from the databases of Medline and PubMed repositories. This model was implemented to classify the likelihood of a document belonging to a certain gene-related clusters based on functional relationships. As a semantic indicator, gene-based vector representation was found to be possibly beneficial. In addition, Apache Hadoop MapReduce was used to enhance the medical care given to the patients and to allow advances in decision making by health decision-makers [44]. The application was implemented on EHRs, a huge complex biomedical data, and a high-quality-omics data. Thus, they analyzed these big data with parallel processing methods using cloud computing with efficient multicore Central Processing Units (CPUs), Graphics Processing Units (GPU), and Field Programmable Gate Array (FPGAs). DM technology was used to define optimal functional recommendations in hospitals on the EHR, networks, and social media data, along with the association rules on the EHRs for the tracking of diseases and health patterns. DM techniques were used in the field. In another study [45], Weka algorithms (J48, basic logistics, and MLP) were used for machine learning on real data from several Iraqi breast cancer cases in early detection hospitals. As a test choice, the authors employed $10-folds$ cross-validation as a test option, and a performance metric of a confusion matrix to evaluate the best among the suggested algorithms. The researchers also

analyzed if, after several algorithm iterations, the error ratio decreases. They found that it is lower than the basic logistic, with J48 algorithms for the MLP algorithm, after 5-10 iterations. Also, the implementation of machine learning algorithms was subject of another work [46]. The sample contained 370 employees in Iraq, where the data were preprocessed to represent the class attribute based on the gender value. Two DM approaches, namely the supervised greedy stepwise subset evaluator (CFS) and the Ranker as a search process, along with the Gain Ratio Attribute Evaluator with Ranker, were utilized to select the attribute for reducing the feature space. Also, the Apriori and association rule algorithms were then used to classify the key factors driving the feature of job apathy inn the study sample.

7. Data Mining Applications

The increase in the use of DM techniques leads to DM applications, including healthcare, manufacturing, electronic commerce, municipal government, education, and transportation, etc.

7.1 Data Mining in Health Care

In healthcare systems, DM has been significantly useful, although its success depends on clean data availability. DM is used to identify patterns of successful treatment for various illnesses, alongside enabling patient habits to be identified by incoming office visits [12]. Doctors can predict the best and efficient methods to improve the quality of patient care. DM offers approaches and methods to transform data into information for successful decision-making, because huge amounts of healthcare data are complicated and thoroughly collected and analyzed [1,7].

7.2 Data Mining in E-Commerce

The fact that data records, such as product data, customer data, and users' log data are many, E-commerce is one of the most eventual fields of DM. In order to obtain an indication of what goods combinations have been bought, the researchers profit from association analysis and clustering, thus inspiring customers to buy items that might have been overlooked or neglected [14]. The competencies and habits of consumers in the web surfing comportments are tracked and analyzed [10]. DM helps businesses to identify the hidden patterns of buying transactions, hence contributing rapidly and cost-effectively to prepare and introduce new marketing campaigns [47].

7.3 Data Mining in Industry

In domains like banking and telecommunications, DM could be of great use. This sector implements the classification and clustering techniques [48]. The estimation of the credit value of borrowers in advance during the process of credit assessment is one of the principal factors for performance of insurance companies and banks [49]. Retailers collect customer information, product information, and related business transactions to boost product demand predicting accuracy, product recommendation, and product ranking for retailers and manufacturers significantly [50]. The SVM [51], or Bass Model [52], aids the researchers in predicting products demand [10, 11].

7.4 Data Mining in City Governance

DM techniques have been used in public service research, service performance improvement, and automated systems-based decision making to decrease risk. The information management systems for city disasters incorporate DM techniques for obtaining a thorough evaluation of effects of natural catastrophes on agricultural production and accurately identify disaster-affected areas [53]. Hence, the researchers can predict, by using data analytics, which factors of city and urban life contribute to the decision of the resident to leave the city [54]. Similarly, police authorities utilize grouping algorithms to identify patterns of crime. In addition to exploring previously unknown systemic patterns on criminal networks, DM can also be used for the identification of criminal identity by analyzing information about people; for example address, name, birth date, and the social security number [55].

7.5 Data Mining in Transport

DM may be utilized in the transport systems to refine the maps made by GPS tracks. Also, researchers may discover the most interesting places, along with the traditional travel sequences for recommendation of location and recommendation of travel [10, 56], according to the multiple users GPS routes.

7.6 Data Mining in Education

DM supports the educators to access student data, forecast their performance levels, and classify students or groups that require additional attention. Hence, in relation to coherent data-driven

principles of student progress, the educators will predict the student's success prior to the beginning, and build intervention strategies to hold them up with the course [1,57].

8. Conclusions

In view of the presented studies, it is concluded that data mining is responsible for exploring useful rules from the raw and complex data, which are gathered from various medical sources. Nevertheless, it should be noticed that, for all types of data, no particular method can be performed, since there is a certain suitable approach to extract information, depending on data type. However, hybrid approaches are sometimes more effective than only one approach. Also, an approach that utilizes more than one algorithm could be enhanced and implemented. Also, the accuracy of various algorithms on the same dataset should be compared. Notwithstanding, the calibration of parameters might produce similar results, so it might be more affecting than the selected approach. Also, the preprocessing step is a very important part for medical dataset, due to its high noise, null values, and high complex raw data. This review draws the attention to the important algorithms in the medical diagnostic, which transform such complicated data into information, thus obtaining new perspectives on human health. Consequently, this study elaborated multiple data mining tools. Each tool has its substantial benefits and drawbacks, with no existence of an optimal tool for all purposes. As a result, we note the DT is an effective algorithm in the medical field, because of its ability to handle mixed data, which is a primary feature of medical data, simplicity in results visualization, and the lack of need to normalization and feature selection before application. Whereas, ANN has been used in diabetic disease classification for type1 and type2 by many researchers. Also, NB could be used as a predictor, together with ANN, in this kind of research. However, SVM is less used depending on data type. On the other hand, K-means enhanced by silhouette method is more successful than hierarchical clustering. Consequently, data mining tools support the discovery of the right dataset algorithms.

References

1. Mittal, S. and Zaman, M. **2016**. A Review of Data Mining Literature. *International Journal of Computer Science and Information Security*, **14**(11): 437.
2. Singh, G. S. **2014**. A Review of Data Mining Techniques. *International Journal of Computer Science and Mobile Computing*, **3**(4): 1401.
3. Raheem, H.A. and Al-Mamory, S.O. **2014**. Privacy Preserving in Data Mining. *Journal of kerbala university*, **12**(3): 179-195.
4. Reddy, D.L.C. **2011**. A review on data mining from past to the future. *International Journal of Computer Applications*, **975**: 8887.
5. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. **1996**. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**(11): 27-34.
6. Martin-Sanchez, F. and Verspoor, K. **2013**. Big data in medicine is driving big changes. *Yearbook of medical informatics*, **23**(01): 14.
7. Parvathi, I. and Siddharth, R. **2014**. Survey on data mining techniques for the diagnosis of diseases in medical domain. *International Journal of Computer Science and Information Technologies*, **5**(1): 838.
8. Khajehei, M. and Etemady, F. **2010**. September. Data mining and medical research studies. *Second International Conference on computational intelligence, modelling and simulation*, 119-122. IEEE.
9. Varghese, D.P. and Tintu, P.B. **2015**. A Survey On Health Data Using Data Mining Techniques. *International Research Journal of Engineering and Technology (IRJET)*, **2**(07): 2395-0056.
10. Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V. and Rong, X. **2015**. Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks*, **11**(8): 431047.
11. Ljupce , M., Igor , Z., and Miroslav, A. **2019**. Data Mining Process.
12. Olson, D.L. and Delen, D. **2008**. *Advanced data mining techniques*. Springer Science & Business Media. E-book.
13. Ashour, M. **2016**. Review of Data Mining Concept and its Techniques. 207-216.
14. Shankar, R. and Duraisamy, S. **2018**. Analysis of Data Mining Tasks, Techniques, Tools, Applications And Trends. *Journal of Computer Engineering*, **20**(5):12-19.

15. Deepthi, P. N., Anitha, R. and Swathi, K. **2019**. A review on bioinformatics using data mining techniques. *Journal of Physics: Conference Series*, **1228**(1): IOP Publishing.
16. Lee, C. H. and Yoon, H. **2017**. Medical big data: promise and challenges. *Kidney research and clinical practice*, **36**: 3-11.
17. Bikku, T., Nandam, S. R. and Akepogu, A. R. **2018**. A contemporary feature selection and classification framework for imbalanced biomedical datasets. *Egyptian Informatics Journal*, **19**(3): 191-198.
18. Han, J., Pei, J., and Kamber, M. **2011**. *Data mining: concepts and techniques*. *Elsevier*. E-book.
19. Aljumah, A.A., Ahamad, M.G., and Siddiqui, M.K. **2013**. Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, **25**(2): 127-136.
20. Dutt, A., Ismail, M.A., and Herawan, T. **2017**. A systematic review on educational data mining. *Ieee Access*, **5**: 15991-16005.
21. Hassan, D., Aickelin, U., and Wagner, C. **2014**. Comparison of distance metrics for hierarchical data in medical databases. *International Joint Conference on Neural Networks (IJCNN)*, IEEE: 3636-3643.
22. Saket, S. and Pandya, S. **2016**. An overview of partitioning algorithms in clustering techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, **5**: 2278-1232.
23. Drab, K. and Daszykowski, M. **2014**. Clustering in analytical chemistry. *Journal of AOAC International*, **97**(1): 29-38.
24. Huang, S.M. **2013**. A study of the application of data mining on the spatial landscape allocation of crime hot spots. *Geo-Informatics in Resource Management and Sustainable Ecosystem*, *Springer, Berlin, Heidelberg*. 274-286.
25. Tummala, Y. and Kalluri, H.K. **2018**. A review on Data Mining & Big Data Analytics.
26. Harper, P. R. **2005**. A review and comparison of classification algorithms for medical decision making. *Health Policy*, **71**(3): 315-331.
27. Dinov, I.D. **2016**. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience*, **5**(1): 13742-016.
28. Haraty, R.A., Dimishkieh, M., and Masud, M. 2015. An enhanced k-means clustering algorithm for pattern discovery in healthcare data. *International Journal of distributed sensor networks*, **11**(6): 615740.
29. Polaka, I. and Borisov, A. **2010**. Clustering-based decision tree classifier construction. *Technological and Economic Development of Economy*, **16**(4): 765-781.
30. Gayathri, V., Mona, M.C., Chitra, S.B., and Chitra, S.B. **2014**. A survey of data mining techniques on medical diagnosis and research. *International Journal of Data Engineering*, **6**(6): 301-310.
31. Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., and Shahmoradi, L. **2017**. Accuracy improvement for diabetes disease classification: a case on a public medical dataset. *Fuzzy Information and Engineering*, **9**(3): 345-357.
32. Jovic, A., Brkic, K., and Bogunovic, N. **2014**. An overview of free software tools for general data mining. *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. *IEEE*. 1112-1117.
33. Farantatos, E., Amidan, B., Arghandeh, R., Bienstock, D., Etingov, P., and Murphy, S. **2019**. Data Mining Techniques and Tools for Synchrophasor Data.
34. Naik, A. and Samant, L. **2016**. Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science*, 85: 662-668.
35. Al-Odan, H.A. and Al-Daraiseh, A.A. **2015**, March. Open source data mining tools. *International Conference on Electrical and Information Technologies (ICEIT)*. *IEEE*. 369-374.
36. Sanap, S.A., Nagori, M., and Kshirsagar, V. **2011**. Classification of anemia using data mining techniques. *International Conference on Swarm, Evolutionary, and Memetic Computing*. *Springer, Berlin, Heidelberg*. 113-121.
37. Developers, S.L. **2018**. Scikit-Learn User Guide. *Release 0.19*, **2**: 214-215.
38. Varshavsky, R., Gottlieb, A., Linial, M., and Horn, D. **2006**. Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**(14): e507-e513.

39. Hsu, C.C., Huang, Y.P., and Chang, K.W. **2008**. Extended Naive Bayes classifier for mixed data. *Expert Systems with Applications*, **35**(3): 1080-1083.
40. Khaleel, M.A., Dash, G.N., Choudhury, K.S., and Khan, M.A. **2015**. Medical data mining for discovering periodically frequent diseases from transactional databases. *Computational Intelligence in Data Mining, Springer, New Delhi*. **1**: 87-96.
41. Salcedo-Bernal, A., Villamil-Giraldo, M.P., and Moreno-Barbosa, A.D. **2016**. Clinical data analysis: An opportunity to compare machine learning methods. *Procedia Computer Science*, **100**(100): 731-738.
42. Diwani, S.A. and Yonah, Z.O. **2017**. A novel holistic disease prediction tool using best fit data mining techniques. *International Journal of Computing and Digital Systems*, **6**(02: 63-72.
43. Mustafa, T.K. and Abd, M.S. **2017**. Proposed approach for analysing general hygiene information using various data mining algorithms. *Iraqi Journal of Science*, **58**(1B): 337-344.
44. Ristevski, B. and Chen, M. **2018**. Big data analytics in medicine and healthcare. *Journal of integrative bioinformatics*, **15**(3).
45. Behadili, S.F., M.S., Mohammed, I.K., and Al-SAYYID, M.M. **2019**. Breast cancer decisive parameters for Iraqi women via data mining techniques. *Journal of Contemporary Medical Sciences*, **5**(2).
46. Abd, M.S. and Behadili, S.F. **2019**. Recognizing job apathy patterns of Iraqi higher education employees using data mining techniques. *Journal of Southwest Jiaotong University*, **54**(4).
47. Padhy, N., Mishra, D., and Panigrahi, R. 2012. The survey of data mining applications and feature scope. *arXiv preprint arXi*, **2**(3): 5723.
48. Elgendy, N. and Elragal, A. **2014**, July. Big data analytics: a literature review paper. *Industrial conference on data mining, Springer, Cham.* 214-227.
49. Kambal, E., Osman, I., Taha, M., Mohammed, N., and Mohammed, S. **2013**, August. Credit scoring using data mining techniques with particular reference to Sudanese banks. *International Conference on Computing, Electrical and Electronic Engineering (Icceee), IEEE.* 378-383.
50. Maaß, D., Spruit, M., and de Waal, P. **2014**. Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques. *Decision analytics*, **1**(1): 1-17.
51. Du, X.F., Leung, S.C., Zhang, J.L., and Lai, K.K. **2013**. Demand forecasting of perishable farm products using support vector machine. *International journal of systems Science*, **44**(3): 556-567.
52. Lee, H., Kim, S.G., Park, H.W., and Kang, P. **2014**. Pre-launch new product demand forecasting using the Bass model: A statistical and machine learning-based approach. *Technological Forecasting and Social Change*, **86**: 49-64.
53. Sullivan, B. and Mitra, S. **2014**. Community issues in American metropolitan cities: a data mining case study. *Journal of Cases on Information Technology (JCIT)*, **16**(1): 23-39.
54. Chen, M. **2013**. Towards smart city: M2M communications with software agent intelligence. *Multimedia Tools and Applications*, **67**(1): 167-178.
55. Fauquet, C., Desbois, D., Fargette, D., and Vidal, G. **1988**. Classification of furoviruses based upon the amino acid composition of their coat proteins. *Developments in applied biology*.
56. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., and Korsmeyer, S.J. **2002**. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, **30**(1): 41-47.
57. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., and Bloomfield, C.D. **1999**. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, **286**(5439): 531-537.