ISSN: 0067-2904

A Review of Data Mining and Knowledge Discovery Approaches for Bioinformatics

Fatin Kadhim Nasser[1*], Suhad Faisal Behadili[1]**
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Received: 21/7/2020 Accepted: 26/9/2022 Published: 30/7/2022

Abstract

This review explores the Knowledge Discovery Database (KDD) approach, which supports the bioinformatics domain to progress efficiently, and illustrate their relationship with data mining. Thus, it is important to extract advantages of Data Mining (DM) strategy management such as effectively stressing its role in cost control, which is the principle of competitive intelligence, and the role of it in information management. As well as, its ability to discover hidden knowledge. However, there are many challenges such as inaccurate, hand-written data, and analyzing a large amount of variant information for extracting useful knowledge by using DM strategies. These strategies are successfully applied in several applications as data warehouses, predictive analytics, business intelligence, bioinformatics, and decision support systems. There are many DM techniques that are applied for disease diagnostics and treatment, for example cancer diseases that are investigated using multi-layer perception, Naïve Bayes, Decision Tree, Simple Logistic, K-Nearest Neighbor. As will be explored in this paper. Consequently, for future perspectives there is research in progress for real Iraqi data of Breast Cancer using Data Mining techniques, specifically the Tree decision and K-nearest algorithms.

Keyword: Bioinformatics, Data Mining, Knowledge Discovery Database, Gene Ontology, Similarity Function.

استعراض لطريقة تعدين البيانات واكتشاف الحقائق في مجال المعلومات البايولوجية

فاتن كاظم ناصر, سهاد فيصل شيحان

قسم علم الحاسوب، كلية العلوم، جامعة بغداد ، بغداد، العراق

الخلاصة

يستكشف هذا الاستعراض طريقة اكتشاف الحقائق التي تدعم مجال المعلومات البايولوجية بكفاءة وايضا توضح علاقتها بتعدين البيانات. من المهم أن نوضح مزايا طرق استخراج البيانات مثل التأكيد بفعالية على دورها في التحكم في التكاليف ، ودورها في إدارة المعلومات. وكذلك ، قدرتها على اكتشاف المعرفة الخفية. ومع ذلك ، هناك العديد من التحديات مثل عدم دقة البيانات وقد تكون مكتوبة باليد ، وتحليل كمية كبيرة من المعلومات المتغيرة لاستخلاص المعرفة المفيدة باستخدام استراتيجيات تعدين البيانات. وتطبق هذه الاستراتيجيات بنجاح في العديد من التطبيقات كمستودعات البيانات ، والتحليلات التنبؤية ، و الأعمال

*Email: fatin.nasser1201@sc.uobaghdad.edu.iq

التجارية ، والمعلومات البايولوجية. هناك العديد من تقنيات تعدين البيانات التي تطبق لتشخيص الأمراض والعلاج ، على سبيل المثال الأمراض السرطانية التي يتم التحقيق فيها باستخدام استراتيجيات **multi-layer perception, Naïve Bayes, Decision Tree, Simple Logistic, K-Nearest Neighbor**. وبناء على ذلك ، هناك تطلعات مستقبلية لتنفيذ بحوث على بيانات حقيقة لمركز اورام الثدي في العراق باستخدام تقنيات استخراج البيانات مثل **Decision Tree and K-Nearest Neighbor**.

1. Introduction

Biological databases keep rising rapidly, it expands in complexity and volume of their database, in addition to spreading new database effect in this development. Thus, a large body of data is available for high-level knowledge discovery, that's including development of new ideas, conceptual interrelationships, and important hidden patterns within databases. Hence, biological data contain errors and missing precision, so their filtering must be done to obtain a good Knowledge Discovery Database (KDD) result [1]. In the biological KDD process several requirements must be taken into consideration. The most important end-product of computation is Knowledge Discovery, which adds a new concept or enhances knowledge (low data level). This has more benefits than optimizing production processes or inventories. This problem is not straightforward, and it is one of the most difficult computational tasks to perform [2]. It is also possible to accept KDD as an immediate, production analysis and patterns selection from big datasets. However, KDD is a structured mechanism by which massive and complex data sets define real, novel, useful, and understandable patterns [3]. It is important to recognize how to collect knowledge before attempting to extract it. Therefore, information must be checked by Data Mining (DM), which concerns the application of low-level DM processes through personal input, which is described as algorithms designed to analyze data, or to select patterns from data in particular classes [4]. KDD has several steps which involve data preprocessing, pattern discovery, assessment of knowledge and improvement, and the foundation of KDD is DM [5]. On the other hand, DM deals with several principles such as data warehouse and entering data, the usability of large datasets, presentation of outcomes, the interaction between humans and machines, and also used for analysis data [5], especially the analysis of biological processes. For example, the genes discovery in DNA sequences [6], legal structure of genomes [7], and knowledge discovery on both transmembrane domain and signal peptide sequences [8].

2. Data Mining Techniques

There are many types of datasets in DM [9] that can be used as outlined next as data collection of particle physics, physiological set of data, data collection of brain-computer interface, Gene/Protein position data set prediction, molecular bioactivity estimation for drug development: connecting to thrombin dataset, internet data base commercial, gene expression vectors of cerevisiae, The colon cancer data, the leukemia data set, the humane splice web data items, data intrusion for network. A great deal of knowledge is collected in medical databases. Therefore, the DM algorithms play a great role in selecting useful knowledge and making decisions for the disease diagnosis [10]. Whereas, the medical industry is becoming ambiguous because of the complexity growing of diseases. For diseases detection, many algorithms are used, such as Naive Bayes, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor, as well as Artificial Neural Networks. Therefore, DM algorithms have proved to diagnose certain diseases successfully [11]. So, DM in research for cancer is one of the most analyzed areas in biomedical science [12], and fourteen classification algorithms can be used for this type of dataset [13].

Bayes Net is used for probabilistic relations between a series of random variables, Bayes Nets or Bayesian networks that are graphical explanations. An explained Directed Acyclic Graph (DAG) mapping a propagation probability is a Bayesian network [14]. There is another type

of algorithm Naive Bayesian on the Bayes conditional likelihood rule used to execute classification processes, the Naive Bayesian classifier is created as development, assuming attributes are uncorrelated, the name Naive means solid. All data set attributes are assumed to be separate from each other and robust [15]. As well as, Simple Logistics is a classification algorithm that performs predictions of linear logistic regression. Basic learners matching the logistic models are LogitBoost with simple regression methods. The perfect number of LogitBoost iterations to perform is cross-validated, resulting in an automated set of attributes [16]. Another algorithm is Multilayer Perceptron is a nonlinear classification algorithm dependent on the Perceptron, Multilayer Perceptron (MLP) is a neural network that may have one or more levels between input level and output level [17]. Another one is Sequential Minimal Optimization (SMO) using RBF kernels or polynomial that can perform support vector classifiers for training. It substitutes all null values and translates nominal values into binary [18]. Moreover, IBk (Instance Based learner) is a k-nearest-neighbor (KNN) classification algorithm that is equal in distance measure. K-NN is lazy learning where the function is locally estimated and all computation is postponed until classification, an object is assigned to its class depending on the vote of its neighbors [19]. In addition, KStar (K*) algorithm, three example-based learners of increasing complexity are identified by Aha, Kibler and Albert. IB1 (Instance-Based) is a representation of a particular distance feature of a nearest neighbor algorithm. IB3 is an extension to improve noisy data. Instances with a very bad classification history are ignored but instances with a good classification result are used for classification. IB4 and IB5 can manage unrelated and new attributes, were defined by Aha [20]. Another one is the Non-Nested Generalization (NNge) algorithm which is "lazy" in the way that when studying from the data set, they perform little research, but spend more time classifying new instances. The easiest strategy, the closest neighbor, when learning done no action performed. NNge explores generalized examples as a means for enhancing the execution of instance-based learners in the classification instead of trying to outperform others classifiers for machine learning [21]. PART algorithm based on a separate and conquer strategy for constructing a rule used to eliminate processed elements and continuous in this manner for treating other cases, where C4.5 and RPPER optimize globally for accurate rules, the key benefit of PAT is this enhanced simplicity [22]. Furthermore, ZeroR is the simplest technique of classification that relies on the goal and avoids all prediction; ZeroR algorithm estimates the class group and helps in evaluating other classification techniques [22]. Another algorithm is Alternating Decision Tree (ADTree) used in machine learning, involving decision and prediction nodes. An example is characterized by an ADTree all decision nodes are valid and any traversed prediction nodes will be summed up, which cause ADTree differs from a simple tree classification model only single way will be followed via a tree [23]. Another algorithm is J48 considers as Weka implementation of C4.5. It useful in reducing pruning error, but it consider as greedy algorithm [24].

As well as another one is Random Forest, it is a classification technique composed of several decision trees and providing outputs in the form of class, classification tree in Random Forest may produce with poor pruning [25]. Finally, simple cart algorithm which is in repeated and incremental style for generating decision tree and estimate classification state for new data of the known variable value of the input. This algorithm is produced depending on Classification and Regression Trees (CART) [26]. For predicting Breast Cancer several algorithms have been applied like Multilayer Perceptron, Naive Bayes, Simple Logistic, KNN, and J8 that is used to distinguish whether Breast Cancer is malignant or benign [27]. Many heuristic strategies, such as the Particle Swarm Optimization Algorithm (PSO) [28] and the Genetic Algorithm (GA) [29] are used for breast cancer prediction.

Open source data available like Surveillance Epidemiology and End Results (SEER) instances. SEER dataset includes numerical and nominal. Age, tumor size, number of positive

nodes, number of nodes, and number of primaries are numerical attributes, while the nominal characteristics include race, marital status, primary site node, histological type, behavior code, grade, tumor extension, involvement of the lymph node, site-specific surgery code, radiation, and cancer level. Wisconsin Breast Cancer (WBC) datasets involves Wisconsin Breast Cancer Diagnosis (WBCD) dataset, and Wisconsin Breast Cancer Prognosis (WBCP) dataset and other publicly available or real-world datasets, example in previous studies data obtained from hospitals, but in Arabic countries are difficult to perform this work. There are several open source database in bioinformatics like National Centre for Biotechnology Information (NCBI), European Molecular Biology Laboratory (EMBL), and DNA DataBank of Japan (DDBJ) [30].

Area Under the Curve (AUC), Classification accuracy (CA), F test (F1), Precision, and Recall are metrics used evaluation algorithm performance [31]. Feature selection must be carried out to maximize the accuracy of the model, but constrained and unconstrained optimization problems may appear with selection, so solving this problem by using a Genetic Algorithm (GA) by developing population to gain optimize solution[32]. Normally, self-examination, mammography, ultrasound, cytology, and core-cut biopsy are tested for the existence of breast cancer in women. After the diagnosis of breast cancer, this disease can be treated through surgery (mastectomy and lumpectomy), radiation therapy, chemotherapy, and hormone therapy [32]. So, the accuracy for each algorithm can be calculated in Eq. (1) [33].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots\dots\dots (1)$$

Where, TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative. TP is a test outcome in which the model properly defines the positive class, while TN is a test outcome where the model classified the negative class correctly. FP is a test outcome where the positive class was incorrectly identified by the model, while FN is a test outcome where the negative class was incorrectly identified by the model.

DM has several tools [34][35] which are useful for inexpert user, and illustrated in Table 1.

Table 1- Data Mining Tools

DM tools	Description
Xplenty	offers a service that performs data integration, collection in order to facilitate data analytics. With the assistance of Xplenty, companies can use advantages offered by a big data without investing staff members, hardware, and software. It is a full tool for constraining data pipelines
Rapid Miner	It is one of the highest classification analysis methods produced by the Rapid Miner company of the same name. It was developed by the programming language JAVA. This tool produces a platform which is used for text analysis, machine learning, deep learning, and predictive analysis. Many applications like commercial, education, machine learning, research, analysis are useful from this tool. Rapid miner organized as client\server structure. Rapid Miner designed with template increased execution speed with minimal error rate.
Orange	It is an application for data mining and machine learning. It is a component-based program that best supports the visualization of data. Python programming language used for writing it. Widgets are the components of Orange. These widgets focus on data analysis, preprocessing, evaluation algorithms, and predictive modeling. By easily comparing and analyzing the data, users can make intelligent decisions in a short time by using Orange [34].
Weka	It is a software for machine learning, data mining, and estimating classes, it is the most suited tool. It includes algorithms and tools for analysis that enable machine learning. With a GUI Weka designed which makes it easy to access its features. Weka, developed by a programming language called JAVA, provides jobs of data mining involving data mining, processing, analysis, regression, etc. deal with data stored as flat

	file and accessing SQL Databases and after processing data by the query can obtain results.
KNIME	It is a suitable tool for analysis and reporting data, it consists of different components of data mining and machine learning integrated together. For pharmaceutical science KNIME is used and also useful for analyzing business data, user data, and market data. Then performs this operation perfectly. KNIME advantages are fast deployment and efficiency of scaling.
Sisense	It is an effective and appropriate tool and useful for reporting task inside a company. It has an excellent capacity for large or small organizations to manage and process information, Great advantage of Sisense enables the integration of information from multiple resources and stored in a warehouse and then provides common access. Sisense helps non-expert people, also includes many tools like drag and drop and the output will be in shape like bar graphs, line charts, pie charts, etc.
SSDT (SQL Server Data Tools)	It is a common declaration pattern that in the Visual Studio IDE extends all stages of database creation. Developers use SSDT transforms to create, manage, debug and refactor databases, and design the framework of SQL. A user can operate either with a database directly or a linked database, thus allowing on-site or off-site services. For database creation such as IntelliSense, code navigation tools, and programming support through C #, visual basic, etc. users may use visual studio tools. SSDT allows the Table Designer with the ability to construct new tables and edit tables in direct databases and linked databases.
Apache Mahout	It plays the primary objective of developing algorithms for machine learning. It deals primarily with clustering, classification, and removing the error from data. It is developed by the Java programming language including java library, which makes it easy to execute mathematical functions such as linear algebra and statistics. Algorithms involved in Apache Mahout are always developed, Mahout continues to expand. A stage above Hadoop by mapping/reducing templates has been implemented by Mahout Algorithms
Oracle data mining (ODM)	It offers data mining techniques that allow analysts to analyze insights, create good estimates, attract the best customers, identify sales, and avoid errors in the classification of data, prediction, regression, and advanced analytics. The algorithms built within ODM exploit the Oracle database strong points. SQL data mining tools can extract data out of tables, views, and schemas of databases. An extended version of the Oracle SQL Developer is the Oracle data GUI (Graphical User interface). It provides users with a direct drag-and-drop service for data within the database, thereby providing a good understanding.
Rattle	It is useful in the statistical domain especially with programming language R in DM; it is GUI depending on DM algorithm, Rattle keep a copy for code executed in GUI [34].
DataMelt (DMelt)	It is a platform for processing and analysis. It is primarily intended for engineers, scientists, and students. By using Java DMelt written, it can execute with any operating system combined with JVM (Java Virtual Machine), also includes a scientific library and mathematical library [34].
IBM SPSS Modeler	at first generated by SPSS and later by IBM for using DM and text research to facilitate estimation classes. Also, make inexperienced people dealing with DM without needing to learn a programming language, such that it will eliminate unwanted information and then predict classes.
Statistical Analysis System (SAS)	It is software created for analytics and data management by the SAS organization. SAS can mine and change data, integrate information from multiple resources and analyze it, for non-expert people SAS provides GUI. SAS enables users to analyze large datasets. SAS architecture is an expansion of distributed memory computing. It is appropriate for data mining, text mining, and visualization.
Teradata database (Teradata)	It is a data warehouse and also contains a DM algorithm that is used for treatment and analysis companies' data like user selection, sales, position of the product, etc. Teradata also distinguishes among 'hot' and 'cold' details, which means in a slow storage section it positions less commonly used data. Teradata operates on the 'Sharing Nothing' design meaning each server point has its own storage unit and processing function.
Board	It is a data mining analytics, and organizational performance management software application. For businesses looking to enhance decision-making, the Board is the most suitable tool for this aim. The Board collects data from various sources, and to get a wanted report the data must be simplified. The Board offers multi-dimensional analysis,

	process management, and performance planning monitoring facilities.
Machine Learning python (Mlpy)	It is an open source tool that offers broad methods of machine learning for problems and tries to find proper solutions, designed with several platforms and depending on python.
H2O	It is another great open source platform for performing analysis of large amounts of data, and analyze the stored data in cloud computing.

3. Knowledge Discovery Database(KDD)

KDD is an evolving area that incorporates database, statistical and artificial intelligence techniques to extract knowledge (high level data) from huge information (low level data). KDD involves several statements to find important facts, DM is an important phase in KDD , Knowledge gets from information through conventional methods depending on understanding it and manual analysis [2]. Slow, costly and highly subjective manual testing of a data set [1]. Manual data analysis is becoming unsustainable with big data drastically, so KDD needs are much more complicated, particularly with bioinformatics information, as relevant data is distributed across heterogeneous and geographically dispersed databases. KDD approach is supplementary for laboratory studies and discovery action in biology can be accelerated by decreasing testing number and increasing the efficiency to analyze biological data. KDD has several processes for extracting useful information, indeed using DM techniques (algorithms) for discovering classes from it, and testing DM items for classifying subsets of enumerated patterns considered as knowledgeable [2], the overall KDD method involves testing of the extracted patterns to decide which patterns will be assigned for new knowledge. However, the elementary steps could be recognized as the following [36][37]:

A. An understanding of the application domain and the necessary prior knowledge is obtained and the goal of the KDD process from the view of the consumer is identified.

B. In this step target variables will be discovered in order to enable finding feature vectors.

C. Most important step is eliminating inappropriate data (noise data) and keeps suitable data this process known as "Data cleaning and preprocessing", sometime important information lost "missing data" so it must identified suitable strategy for handling missing data.

D. Based on the aim of task features must be identified to represent information and this operation performed by eliminating data.

E. Fit KDD process objectives with DM methods like classification, clustering, regression and summarization which will be explained later.

F. Algorithm for DM will be selected for assigning data to a class; this operation requires identifying models and parameters will be acceptable and fitting KDD process with DM algorithm.

G. It is DM, looking for patterns of concern or a collection of such patterns in a specific framework investigation, involving clustering, regression and classification tree or rule.

H. This process may also include the representation of patterns and models extracted or representation of the data provided the models extracted. probably go for further iteration to each of stages A through G

I. This step focuses on discovering knowledge by using the knowledge directly, integrating information, and then transferring it into new form, and also discovering knowledge by checking if there is unwanted knowledge, and trying to solve this problem.

J. Newly discovered knowledge assessment of the intent of KDD is also used to develop new assumptions; new questions can also be posed by using enhanced knowledge. Evaluating KDD process performed in this stage which allows using it again with enhancement and growth situation. So previous steps illustrated in Figure 1.

A motivating factor behind KDD is the field of databases; actually the challenge of efficient data manipulation has a critical importance to KDD when the main memory is enough for data. Algorithms expand to deal with larger data sets and accessing it under control. Data

warehousing is a connected field emerging from databases, its function gathering data (collecting data) and eliminating noisy data (clearing data) and transformational data to facilitate analysis of data online and discover knowledge. Data warehouse enables KDD to perform: data cleaning and access to data [2][1]. Whereas, data cleaning is performed when companies are challenged to learn about a single rational big data they have. Hence, they forced the problem of how to assigned unique names to data, how to treat missing data, and eliminate noise and error data. As well as, data access defines the way which enables entering data which is hard to access (offline data). After execution these operations KDD will execute. Data warehouses and especially multi-database systems are the most interesting from the KDD perspective, On-Line Analytical Processing (OLAP) is a standard technique for data warehouse research [38].

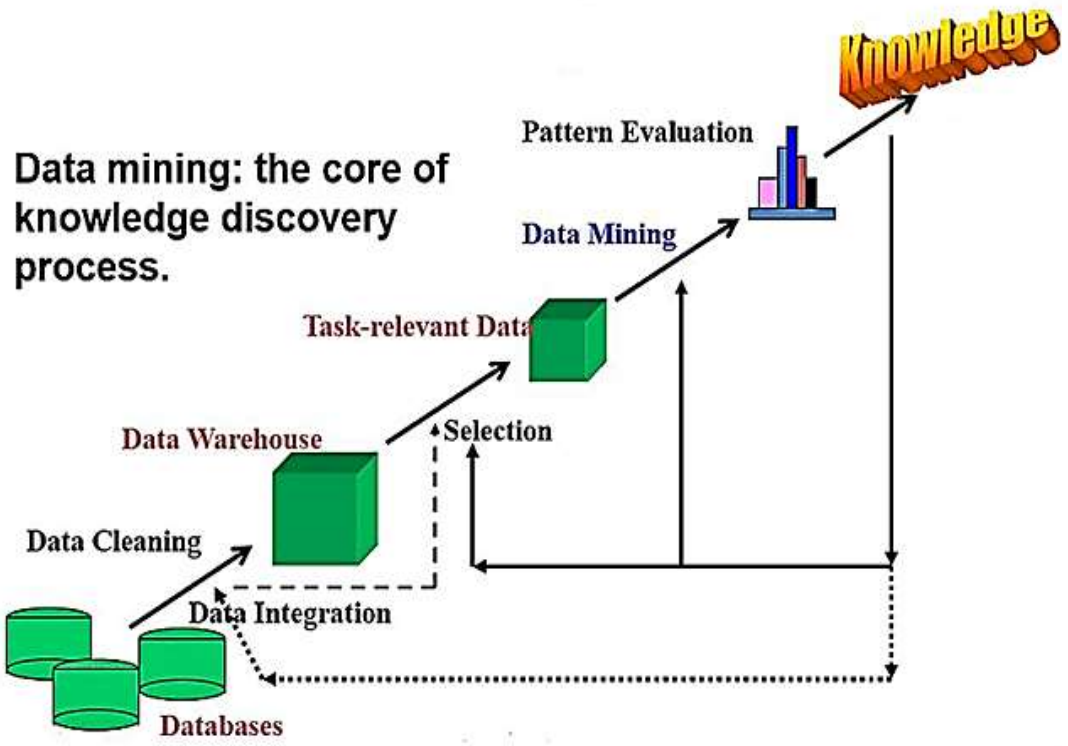

Figure 1-Knowledge Discovery Process Diagram [39]

4. Data Mining phase in KDD

The data-mining aspect of the KDD process also requires repeated iterative application of unique methods of data-mining [40][1]. This section describes an overview of the key data mining objectives, a summary of the techniques used to address these objectives, and a short summary of the algorithms for data mining that implement these techniques. It must differentiate between two types of target verification and discovery. With verification, this method is restricted to checking the hypothesis of the user. With discovery, new patterns are autonomously detected by the patterns. Further subdivide the discovery goal into prediction; it can also divide the purpose of exploration into prediction, where the system finds patterns for predicting certain entities future actions, and description, where the system finds patterns for a user's presentation in a human-understandable form [2]. Data mining includes fitting models from input data, or identifying patterns. In model fitting, two main mathematical constructs are used, logical and statistical. The statistical strategy in the model allows for non-deterministic outcomes, where as a logical model is strictly deterministic. DM algorithms depend on machine learning, pattern recognition, and statistics (classification, clustering, and regression) and so on. However, the suitable algorithms are determined by using fitting

criteria [1][2][40]. In general, the two key high-level objectives of data mining tend to be prediction (supervised) and description (unsupervised), prediction requires the use specific variables in the database to estimate target values of other variables, and description based on identifying user explanation patterns that explain information, in spite of difference among description and prediction is weak [2]. Differentiation is helpful in understanding the ultimate purpose of discovery. Advantages and disadvantages of each type are illustrated in Table 2.

Table 2-pros and cons for supervised and unsupervised strategies

DM type	Pros	Cons
Supervised	• Classes indicate the terrestrial characteristics. • Training data can be reused except for changes in features.	• Consume time and cost will selection training data. • Classes shall not correspond to spectral classes.
Unsupervised	• Less complexity. • Useful in real time application. • Does not require previous acknowledge in image field.	• Does not care to spatial correlation among data. • Interpretation for spectral classes will consume time. • Does not necessary that spectral classes will describe attributes.

DM algorithms divided into multiple groups As well as Clustering is an unsupervised machine learning, its work depends on the principle of dividing data into groups or clusters for describing data [19][41]. In a knowledge discovery sense, samples of clustering activities include the discovery in marketing databases of homogeneous subpopulations for customers and the detection of subcategories of spectra from Measurements of the infrared sky [42]. Prediction of thickness likelihood example of clustering, estimating the joint multivariate likelihood density function of all variables or fields in the database from data [43]. Since little information is available about gene clustering useful in testing microarray data.
[44]clustering algorithms dividing into partition clustering involve dividing dataset into partition clustering, hierarchical clustering, and density based clustering. In partitioning clustering involves K-Means algorithm and K-Mediods. K-Means works on principle of dividing dataset into number of groups (K groups), dataset in same cluster with high similarity inverse with dataset in other clusters with low similarity. Hierarchical clustering works on principle of dividing dataset in a hierarchical manner, generally Hierarchical clustering has two types Agglomerative and Divisive. So, the Agglomerative function depends on a principle where each element in the dataset is considered as a cluster, and then combines these clusters depending on the similarity measurement until they become a single cluster. While, in contrast the Divisive considers the total dataset as a single cluster at first, and then divides it into a number of clusters until it reaches termination condition. In Density Based clustering success in treatment fault of partition and hierarchical clustering, also can deal with arbitrary shape data not only spherical like partition and hierarchical clustering. Moreover, OPTICS and DBSCAN are two algorithms of density based clustering that depend on the principle of analyzing density linking; while DENCLUE clustering performs by analyze distribution values. As well as, Regression is considered as a supervised Learning mechanism, which translates a data into a prediction variable with true value and discovering functional relationships between variables [45]. In addition, Classification is a supervised method that transforms (classifies) a data item within one of several learned class labels [46]. The classification of developments in financial markets is an example of classification approaches for discovering knowledge [47] and image databases, and the automatic recognition of objects of interest [2][48]. Classifications used in medical area for diseases diagnostics and minimize cost. There are many classification algorithms effect successfully

for this purpose such as [49]: KNN is a supervised and simple algorithm, which stores all available instances that consume very large memory, and predicts the numerical target based on the dissimilarity measuring (distance functions), also it is usually implemented either by using Euclidean distance compute in Eq. (2), or by using Manhattan distance also called City Block as computed in Eq. (3), or using Minkowski Distance as computed in Eq. (4), which is considered as generalization of Euclidean and Manhattan.

$$D(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \quad \dots\dots\dots\dots\dots (2)$$

$$D_{(x_1,x_2)} = |x_1 - x_2| + |y_1 - y_2| \quad \dots\dots\dots\dots (3)$$

$$D = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \dots \quad \dots\dots\dots\dots\dots\dots (4)$$

When, $p = 1$ considered as Manhattan, else when $p = 2$ considered as Euclidean. Moreover, Decision Tree (DT) is the result of classification model, or regression as diagram style. This operation could be performed by splitting the dataset into the smallest data, while at the same time establishing an association tree gradually. Finally, the output tree would be constructed from root, leaf nodes, and decision nodes. Some criteria used in DT to determine the attribute of the best partition like Gini Index (GI), which is reducing as much as possible the misclassification probability, so it could be computed as in Eq. (5).

$$GI = 1 - \sum_{j}^{c} p_j^2 \dots \quad \dots\dots\dots\dots (5)$$

Where, GI is the gini index, c is the class labels, p_j is the probability of class$j$. Additionally, the entropy computes the system randomness, thus if all data belong to the same class, then$entropy = 0$, else if each class as the same instance, then entropy will be maximized according to Eq. (6) [50].

$$Entropy = \sum_{i}^{c} -p_i * log_2(p_i) \quad \dots\dots\dots\dots (6)$$

Where, p_i is the probability of class i. Furthermore, Information Gain (IG) is dependent on entropy, and is defined as a metric that determines if the feature is useful or not in the classification, so IG could be computed as in Eq. (7). Hence, the DT structure is illustrated in Figure 2 [51].

$$IG = entropy(parent) - average\ entropy(children) \dots\dots\dots\dots (7)$$

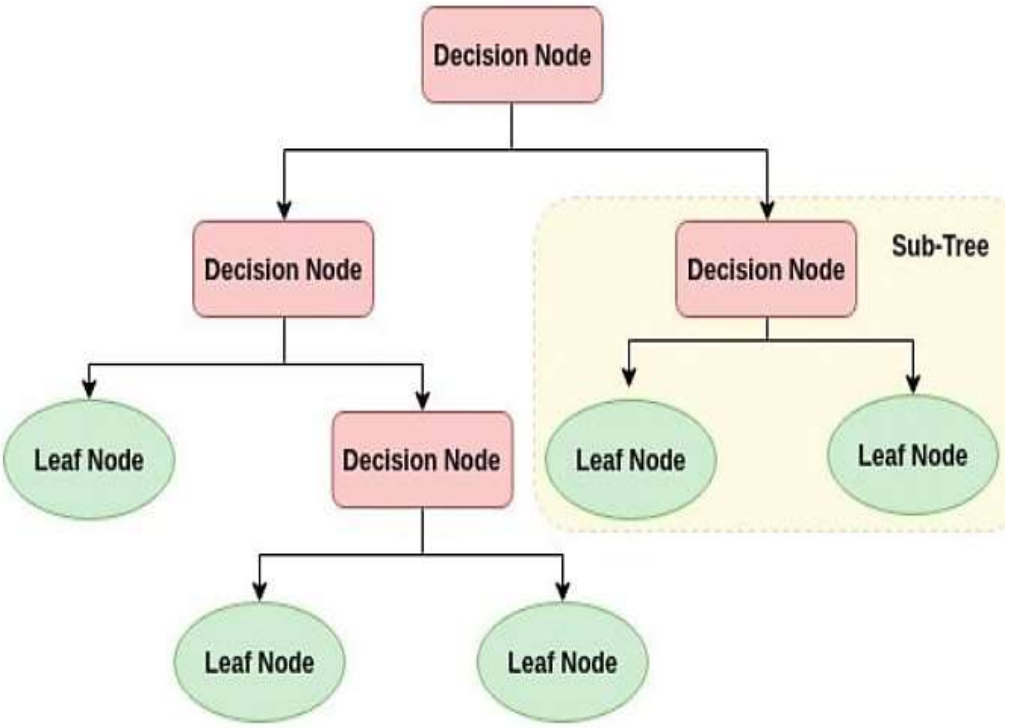

Figure 2-DT structure

Furthermore, Artificial Neural Network (ANN) [52] is another algorithm consisting of interconnected nodes, and weighted path, where the output node sums its input node according to its weighted link, and then compares the output by the threshold t as determined in Eq. (8). ANN has multi types single layer network (Perceptron), which consist of input and output nodes, and multi-layer neural network that contain hidden layers additionally to input and output layers. The ANN structure is shown in Figure 3.

$$y = sign\left(\sum_{i=1}^{d} w_i\, x_i - t\right) \quad \ldots\ldots\ldots\ldots (8)$$

Table 3 shows advantages and disadvantages of these algorithms [11].

Table 3-pros and cons of classification algorithms

Classification technique	Pros	Cons
K- Nearest Neighbor (KNN)	• It is easy to construct. • Use local information.	• Consume large amount of memory • Testing slowly.
Decision Tree (DT)	• Automatically perform feature selection, treat missing data, and does not require normalization on dataset, and it is easy to interpret and visualize.	• Occur in overfitting problem. • Not suitable for categorical data. • It may be biased to some attributes than other attributes and may lead to complex decision.
Artificial Neural Network (ANN)	• Treat missing data and noise data. • Ability to work with big data.	It is complex to change existing network.

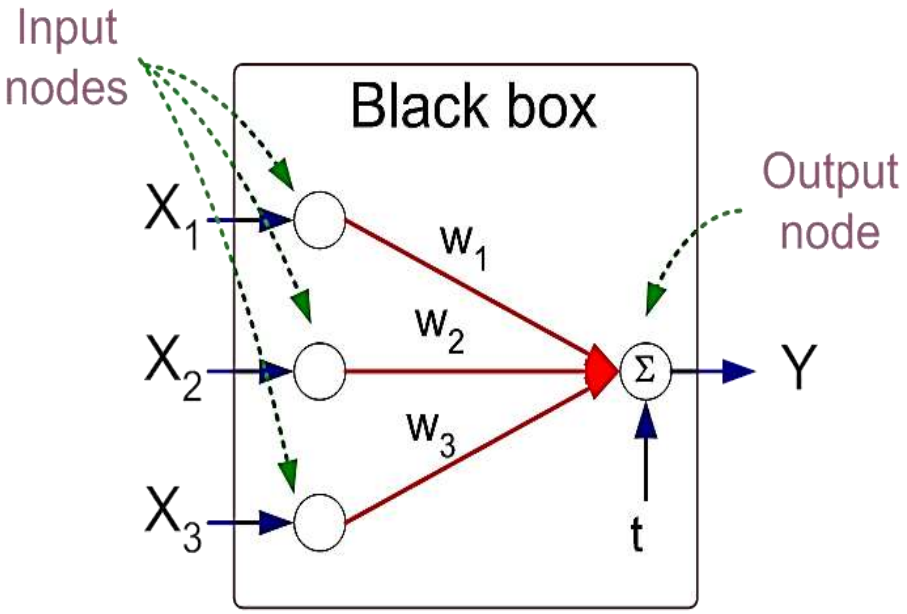

Figure 3-Artificial Neural Network

5. The Bioinformatics

Generally, Bioinformatics is a scientific discipline within computer science, where biology and information technology are integrated and viewed as one framework [53][54]. The goal of this field is to establish some new biological insights to be created, and to establish a global view that discerns unifying principles in biology. Hence, the Biological data could be conventional that are low-throughput biological data, contradictory terminology, and manually assembled, accurate [55], or modern high-throughput, standard, automated, error-prone technological processes [56]. From molecules and biological sequence specially Genome, DNA sequencing and microarray platforms bioinformatics gets useful information and using it in fruitful way [57][58][59]. Nowadays, the high technology of prediction, trends and hypothesis discovery from bioinformatics researches has evolved significantly [60], which suggesting that data mining of bioinformatics has many utilizations, including "gene discovery, protein function domain detection, function motif detection and inference of protein function.". This is clear in identifying meaningful questions, and obtaining appropriate responses. Thus, in order to investigate Bioinformatics data efficiently it is important to use KDD technologies. Therefore, if there is enough data available, and the biological problem is well established, then the studies should be performed using standard statistical methods. As a consequence, the epidemiology [61] has been routinely used, which is defined as an area in this approach. Ordinarily, the objective of gene and protein are produced using statistical analysis [62], the majority of biological studies are particularly in molecular biology that are carried out in fields of insufficient background knowledge, which use the data from different sources and with variable accuracy. On the other hand, the artificial intelligence approaches are more helpful than statistical techniques in such situations. As well as, describing the Data Learning Process (DLP) [1] had been used to implement the bioinformatics analysis for the biological systems. However, the data warehouses and particularly multi-database systems are the most interesting from the KDD point of view, which perform integration, cleaning, and minimize transactional data [1]. Whereas, OLAP is a common strategy for data warehouse analysis [58], and it had been considered better than SQL in the analysis task. The purpose of OLAP tools is to simplify and facilitate interactive data analysis. Accordingly, some fundamental notions of biology need to be implemented next [53]. Meanwhile, the Cells are considered as basic operating units for life. Then, the chemical DNA that contains the

necessary instructions to guide their activities. Also, the DNA has the same chemical and physical makeup of all organisms. DNA consists of four types of nucleotides (adenine (A), guanine (G), cytosine (C), and thymine (T)), its sequence is arranged in standard structure for example, ATTCCGGA, as well as, the actual needed instructions of constructing a specific organism with its own unique characteristics. Additionally, a genome is regarded as a DNA full collection of an organism. So, the genomes are widely different in size, Approximately 600,000 DNA base pairs (bp) contained in a bacterium, while the human and mouse genomes have 3 billion bp [63]. Regarding that all human cells contain a full genome, except for mature red blood cells, then a big difference among two living organisms is the genomic variance [63]. Consequently, the DNA for people is placed into 24 chromosomes. Meanwhile, the proteins carry out most activities of life, and even make up the maximum structures of cellular. The proteins are highly complex molecules constructed from tiny subunits called amino acids. Whereas, the proteome is the collection of all proteins in the cell. Indeed, the dynamic proteome alters from minute to minute in response to tens of thousands of internal and external environmental signals, unlike the relatively unchanging genome [64]. In addition to another one, Gene Expression that tests the involvement degree of genes using a chip microarray. Also, there are three types of chains for the Life molecular building blocks that are DNA, RNA, and proteins. Regularly, the Dual-stranded large series consists of four nucleotide types (A, C, G, and T), which is the DNA molecule. It has the form of a double-helix, and saves genetic material. Also, the RNA molecules that are made up of four nucleotides (A, C, G, and U) are very similar to DNAs. Likewise, the proteins are sequences of 20 different main components named amino acids. And, the gene is the basic unit of genomic DNA that holds the needed data of conducting the cell biological functions. Virtually, all important functions in a cell are performed by proteins. Then, it is necessary to encode the corresponding gene into mRNA, and then convert the mRNA to a protein for protein production. Specifically, the examination of gene expression is useful in the medical treatment of cancer and other harmful diseases throughout testing regulator gene abnormalities [13]. Moreover, the Textual Phenotypes had been used to explain in plaintext information about a gene, and then convert this plaintext into a meaningful form.

5.1. Gene Ontology (GO)

Generally, GO defines cellular location, biological process, and molecular function of gene products and to thus enable extraction of biological meaning from these large datasets [65], where the experimenters also use terminology from this ontology in order to mark the genes functions for providing a specific language to define the cellular position, biological system, and molecular role in gene generation. In consequence the Ontology (GO) was initiated to extract the effective biological parameters from huge datasets [66]. The GO terminologies are arranged in a directed acyclic graph, whereas the associations among terminologies could be represented though directed edges. So, the gene product assignment of a GO term is called annotation. Hence, explaining the gene products to perform the computational methods for analyzing high throughput datasets. Accordingly, the GO annotation has become a "standard method". In order to explain the UniProtKB entries (UniProtKB is central hub used for storage accurate, consistent information on protein[67]) with GO terminology, then the Gene Ontology Annotation (GOA) [68] uses a pipeline that includes manually assembled and electronic approaches. Especially, the annotation manual assignment depends on looking for proof across literature that a protein has a specific job, in spite of this process can be costly and long, but the outputs are specific and additionally correct. However, the pipeline electronic form contains data from multiple sources such as electronic annotation that is especially helpful for assigning GO terminology to non-model organism proteins; of course it is unlikely to obtain manual annotations. In connection with the GO assignment, there are many computational annotation pipelines, including [69] merged information from multiple

resources both manually edited, and computationally sent for unique database, [70] for adding features to new mixed series of clustering with homology test. As well as, text mining can be used to estimate gene jobs in order to simplify the manual process of literature annotation of gene items. For example, [71] select automatically the functional annotations from Medline texts for mammalian proteins via establishing normal expression and relationships between GO terminology and proteins [72] [73]. Concerning the text mining usage to cluster term frequency-inverse document frequency ($tf - idf$) arrays to connect phenotypes with genes. This could be done depending on all genes similarities that can run a classification process; this process relies on two data types phenotype obtained using text mining and gene expression. Graphs are then generated such that each node represents the gene, and the edge represents the similarity between gene pairs. Hence, to determine the associated functions with unidentifiable genes, the GO annotation is used as the Training set label [65].

5.2. The similarity functions

Usually, the similarity functions are the basis of the prediction algorithms. This consists of comparing two genes if they are different, then returns 0, else returns 1, as in Eq. (9).

$$f : G * G \rightarrow [1,0] \quad \ldots\ldots\ldots\ldots (9)$$

Where, f is a function of similarity, and G is the gene collection [65].

5.3. The Similarity function for gene expression

Actually, the similarity function between two genes for gene expression data is called the "Pearson correlation coefficient " of the two genes related expression arrays that could be computed by Eq. (10) [74].

$$r\,(x,y) = \frac{\sum(x_i - \bar{x}\,)\,(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \;\; \sum(y_i - \bar{y})^2}} \quad \ldots\ldots\ldots\ldots (10)$$

Where, $x_i$is a member from the gene expression vector x, while $\bar{x}$ is the average of vector x members, and y_i is a member from the gene expression vector, and $\bar{y}$ is average of vector y members. Then, the negative value cannot be applied to the term GO. Accordingly, if the coefficient of Pearson correlation receives a negative value, then it will be modified to $zero$ [75]. So, the similarity between two expression-based genes is computed by Eq. (11).

$$f_{expression}\,(g_i\,, g_j) = \max\big(r\big(v_i\,, v_j\big), 0\big) \ldots\ldots\ldots\ldots (11)$$

Where, v_i and v_j are genes expression vectors related to genes g_i and g_j.

5.4. The similarity functions of Textual Phenotype

Indeed, to make the computation simpler, it is necessary to change textual phenotype into another shape. Thus, the term frequency-inverse document frequency ($tf - idf$), which is the most common method for the conversion of text phenotype to array . The frequency of the term (tf) represents the appearance number of a term in a document, whereas the frequency of the inverse document (idf) term appears in more documents. The ($tf - idf$) measures the cosine distance between the two gene-associated ($td - idf$) arrays as shown in Eq. (12) [76].

$$f_{phynotype}\,(g_i\,, g_j) = \cos\big(v_i\,, v_j\big) = \frac{v_i^t\; v_j}{\|v_i\|\|v_j\|} \quad \ldots\ldots\ldots\ldots (12)$$

Where, v_i and v_j are vectors of ($tf - idf$) related to genes g_i andg_j. Thereafter, the similarity diagram could be measured after calculating the similarity of text phenotype and gene expression respectively. Even so, the graph consists of nodes describing genes, whilst weighted edges between nodes representing gene similarity. Hence, measuring the weight of edges either by finding total values of similarity function between two genes as in Eq. (13), or by taking the max value as in Eq. (14).

$$w_{(i,j)} = \sum_{k=1}^{n} f_k\left(g_i, g_j\right) \quad \ldots\ldots\ldots\ldots (13)$$

$$w_{(i,j)} = \max_k f_k\left(g_i, g_j\right) \quad \ldots\ldots\ldots\ldots (14)$$

Where, $w_{(i,j)}$ represents connections (weighted edges) among nodes (genes) g_i andg_j, and f_k represents a similarity function depending on a single dataset.

5.5. Functional annotations estimation

Firstly, in order to estimate functional annotations, then delete the gene from evaluation, then find two thresholds to determine annotation for a specific gene. As in Eq. (15), where the lower threshold is calculated by the gene h with annotation , which is the lowest cumulative similarity to other genes with this annotation. Hence, for all h where annotation (a, h) and $h_i \neq h_j$ by locating the h gene without annotation, so the upper threshold is very close to the annotation genes as in Eq. (16). However, this mechanism has been explored in Figure 4.

$$min_i \sum_j sim\left(h_i, h_j\right) \quad \ldots\ldots\ldots\ldots (15)$$

$$max_i \sum_j sim\left(h_i, h_j\right) \forall : \sim \text{annotation(a, h) and } h_i \neq h_j \ldots\ldots\ldots\ldots (16)$$

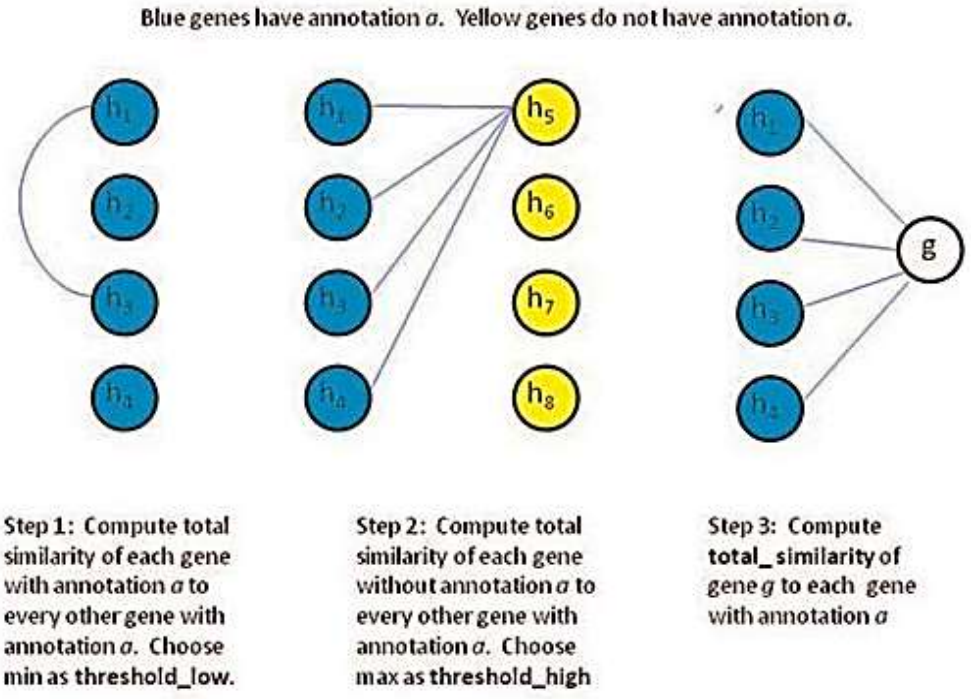

Figure 4-The similarity computation

Accordingly,

If sim $\left(h_i, h_j\right) > threshold\ G$, then still required to provide an annotation$a$,

If sim $\left(h_i, h_j\right) < threshold\ G$, then never expected to provide an annotation of a,

If sim $\left(h_i, h_j\right)$ $between\ upper\ and\ lower\ threshold$, then interpolation is used to determine the position of similarity according to two thresholds, and produce $number\ between\ 0\ and\ 1$, as in Eq. (17) [65].

$$int_{sim} = \frac{total_{similarity} - lower\ _{threshould}}{upper_{threshould} - lower_{threshould}} \dots\dots\dots\dots (17)$$

In this concern, $IF\ interpolated\ similarity\ (int_{sim}) > cutoff$, then assigned the annotation, this process has been explained in Figure 5. Hence, the $cutoff$ is a predefined value used to predict and assign the gene annotation or not. Over recent years, data mining classification algorithms have been successfully proposed to estimate cancer disease depending on gene expression data. Meanwhile, Microarray that is an effective screening method that can produce all genes in a cell simultaneously with a small amount of gene expression information. One of the main Microarray techniques is the Gene Expression Analysis. Nevertheless, a nucleic acid test (target) will be hybridization to a wide collection of oligo-nucleotide probes, which is a microarray for establishing the sequence or identifying differences in a gene sequence or the levels expression for gene mapping. Also, the tumor classification has been widely investigated in recent years by applying DM techniques to extract cancer gene expression microarray datasets, then to predict the cancer existence [13].

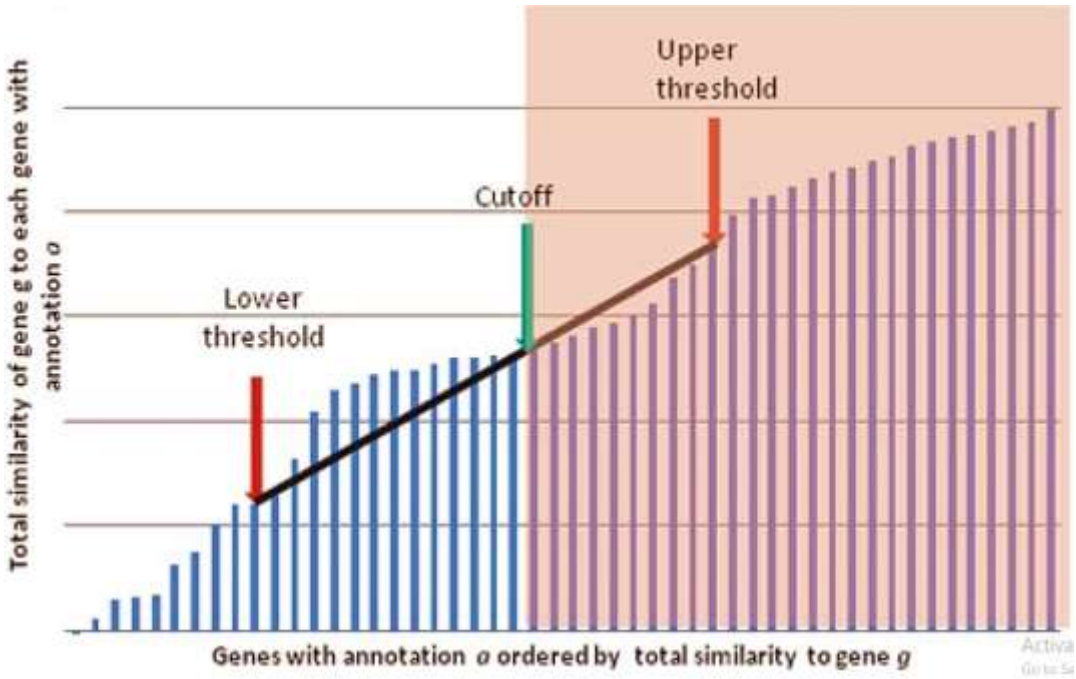

Figure 5-prediction transfer to annotations to gene [65]

6. The Discussion

As presented in the previous sections of this review, the DM played a great role in Bioinformatics domain. So, the investigated techniques have been explored in Table 4 according to their evolution along the years in a comparison among them according to their accuracy performance within the healthcare field.

Table 4-DM performance on Bioinformatics data

DM algorithm	Accuracy	Author name and year
Naïve Byes	78.56%	Andreeva, P [77]-2006
Decision Tree	75.73%	
Neural Network	82.77%	
Naïve Byes	62.03%	Sitar-Taut, et al [78] -2009
Decision Tree	60.40%	
Naïve Byes	52.33%	Rajkumar, et al [79]-2010
Decision Tree	52%	
KNN	45.67%	
Naïve Bayes	84.14%	Srinivas, et al [80]-2010
One Dependency Augmented Naïve Bayes classifier	80.46%	
Genetic with Decision Tee	92.2%	Anbarasi, et al [81]-2010
Genetic with Naïve Bayes	96.5%	
Genetic with Classification via Clustering	88.3%	
Naïve Bayes	76.3 %	Kaur, et al [82] -2014
MLP	73.39 %	
ADTree	72.91 %	
J48	73.82 %	
Support Vector Machine	98.1%	Obaid, et al [83]- 2018
Decision Tree	93.7%	
KNN	96.7%	
KNN	76.92%	Ansari, et al [84] -2019
Decision Tree	71.73%	
Naïve Bayes	74.51 %	
Support Vector machine	73.63 %	
Artificial Neural Network	82.16 %	

7. Conclusions

This review shows the role of KDD in the Bioinformatics field through extracting knowledge from huge amounts of information, and explores the relationship between KDD and DM, KDD has several steps including pre-processing, model building, and discovering knowledge . Thus, DM algorithms achieved successes in dealing with the biological data in spite of suffering from noise, missing value, multiple data types that may be text or image, also spread in a wide area. So, this review highlighted the OLAP as one of the common solutions for transforming big data into useful information. According to the Bioinformatics field, cancer datasets have been used to diagnose and treat this disease, as an effort to enhance human aging. Therefore, DM investigations that related to cancer are the most important for biomedical scientists. Furthermore, the gene expression had been highlighted here, since it provides a big support for DM algorithms to diagnose cancer disease. Whereas, the Microarray had been presented, because it had been considered as an efficient recognition mechanism that simultaneously produces a number of gene expression data from all the human genes in a cell. As well-known the breast cancer is the most spread disease among people specially women and may cause death, then DM presented as an effective approach that plays an important role in the healthcare field either generally or specially in prediction of breast cancer. Additionally, accuracy had been discussed as shown in Table 4 since it is a decisive metric to determine the most suitable algorithm and obtained results, and obvious accuracy increased when combined DM algorithms with GA. Finally, it is recommended for the future to enhance the contribution among physicians and informatics in order to reduce the cost and time of disease manipulation and their data management. As well as, it is commendable to establish a contribution bridge among the medical physicians and informatics researchers in Iraq with a view to support the patients as could as possible.

References

[1] V. Brusic and J. Zeleznikow, “Knowledge discovery and data mining in biological databases,” *Knowl. Eng. Rev.*, vol. 14, no. 3, pp. 257–277, 1999.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, vol. 17, no. 3, pp. 37–53, 1996.

[3] O. Maimon and L. Rokach, “Data Mining and Knowledge Discovery Handbook,” *Data Min. Knowl. Discov. Handb.*, pp. 1–2, 2010, doi: 10.1007/978-0-387-09823-4.

[4] Z. W. Raś and J. M. Z. ytkow, “Discovery of equations and the shared operational semantics in distributed autonomous databases,” 1999, doi: 10.1007/3-540-48912-6_60.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “The KDD Process for Extracting Useful Knowledge from Volumes of Data,” *Commun. ACM*, 1996, doi: 10.1145/240455.240464.

[6] A. Krogh, I. S. Mian, and D. Haussler, “A Hidden Markov Model That Finds Genes in Escherichia-Coli Dna,” *Nucleic Acids Res.*, 1994.

[7] A. Brazma, J. Vilo, E. Ukkonen, and K. Valtonen, “Data mining for regulatory elements in yeast genome.,” *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1997.

[8] T. Shoudai *et al.*, “BONSAI Garden: parallel knowledge discovery system for amino acid sequences.,” *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1995.

[9] D. Mining, “Datasets for data mining.” pp. 1–12, 2016.

[10] T. K. Mustafa and M. S. Abd, “Proposed approach for analysing general hygiene information using various data mining algorithms,” *Iraqi J. Sci.*, vol. 58, no. 1B, pp. 337–344, 2017.

[11] A. Dwivedi, K. Rehman, M. Ghosh, and R. Raman, “Data Mining Algorithms in Healthcare,” *Int. J. Comput. Appl.*, 2018, doi: 10.5120/ijca2018916901.

[12] R. Agarwal and M. V. Joshi, “PNrule: A New Framework for Learning Classifier Models in Data Mining (A Case-Study in Network Intrusion Detection),” 2001, doi: 10.1137/1.9781611972719.29.

[13] G. Krishna, B. Kumar, N. Orsu, and S. B., “Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification,” *Int. J. Adv. Res. Artif. Intell.*, 2013, doi: 10.14569/ijarai.2013.020508.

[14] Y. kumar and G. Sahoo, “Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA,” *Int. J. Inf. Technol. Comput. Sci.*, 2012, doi: 10.5815/ijitcs.2012.07.06.

[15] P. Domingos and M. Pazzani, “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss,” *Mach. Learn.*, 1997, doi: 10.1023/a:1007413511361.

[16] N. Landwehr, M. Hall, and E. Frank, “Logistic model trees,” *Mach. Learn.*, 2005, doi: 10.1007/s10994-005-0466-3.

[17] B. W. White and F. Rosenblatt, “Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms,” *Am. J. Psychol.*, 1963, doi: 10.2307/1419730.

[18] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to Platt’s SMO algorithm for SVM classifier design,” *Neural Comput.*, 2001, doi: 10.1162/089976601300014493.

[19] D. Coomans and D. L. Massart, “Alternative k-nearest neighbour rules in supervised pattern recognition. Part 1. k-Nearest neighbour classification by using alternative voting rules,” *Anal. Chim. Acta*, vol. 136, no. C, pp. 15–27, 1982, doi: 10.1016/S0003-2670(01)95359-0.

[20] A. Patasius and G. Smailyte, “Re: MaryBeth B. Culp, Isabelle Soerjomataram, Jason A. Efstathiou, Freddie Bray, Ahmedin Jemal. Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates. Eur Urol 2020;77:38–52,” *European Urology*. 2020, doi: 10.1016/j.eururo.2019.11.030.

[21] K. P. Byrne, C. L. Smith, J. Termaat, and B. C. H. Tsui, “Reversing the effects of a peripheral nerve block with normal saline: A randomised controlled trial,” *Turkish J. Anaesthesiol. Reanim.*, 2020, doi: 10.5152/TJAR.2019.09076.

[22] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2011.

[23] H. Yang and S. Fong, “Improving the accuracy of incremental decision tree learning algorithm via loss function,” 2013, doi: 10.1109/CSE.2013.136.

[24] S. L. Salzberg, “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann

Publishers, Inc., 1993," *Mach. Learn.*, 1994, doi: 10.1007/bf00993309.
[25] "Encyclopedia of Machine Learning and Data Mining | Claude Sammut | Springer." [Online]. Available: http://www.springer.com/us/book/9781489976871.
[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. 2017.
[27] S. Faisal Behadili, M. S. Abd, I. Kamil Mohammed, and M. M. Al-Sayyid, "Breast Cancer Decisive Parameters for Iraqi Women via Data Mining Techniques," no. May, 2019.
[28] K. R. Gandhi, M. Karnan, and S. Kannan, "Classification rule construction using particle swarm optimization algorithm for breast cancer data sets," 2010, doi: 10.1109/ICSAP.2010.58.
[29] C. A. Pe and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," *Artif. Intell. Med.*, 1999, doi: 10.1016/S0933-3657(99)00019-6.
[30] A. Rai, J. Bhati, and S. B. Lal, "Software Tools and Resources for Bioinformatics Research," *Appl. Comput. Biol. Stat. Biotechnol. Bioinforma.*, vol. 1, no. January, 2012.
[31] H. Saad and N. Nagarur, "Data Mining Techniques in Predicting Breast Cancer," *J. Appl. Sci.*, vol. 20, no. 3, pp. 124–133, 2020, doi: 10.3923/jas.2020.124.133.
[32] K. Rajendran, M. Jayabalan, V. Thiruchelvam, and V. Sivakumar, "Feasibility study on data mining techniques in diagnosis of breast cancer," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 328–333, 2019, doi: 10.18178/ijmlc.2019.9.3.806.
[33] D. Delen, "Analysis of cancer data: A data mining approach," *Expert Syst.*, 2009, doi: 10.1111/j.1468-0394.2008.00480.x.
[34] Vijay, "Top 15 Best Free Data Mining Tools: The Most Comprehensive List," *Software Testing Help*. pp. 1–11, 2018, [Online]. Available: https://www.softwaretestinghelp.com/data-mining-tools/%0Ahttp://www.softwaretestinghelp.com/penetration-testing-tools/.
[35] Anjali UJ, "The Top 10 Data Mining Tools of 2018 Analytics Insight." 2018.
[36] A. Rajput, "KDD Process in Data Mining - GeeksforGeeks." 2018, [Online]. Available: https://www.geeksforgeeks.org/kdd-process-in-data-mining/.
[37] F. Fatima, R. Talib, M. K. Hanif, and M. Awais, "A Paradigm-shifting from Domain-Driven Data Mining Frameworks to Process-based Domain-Driven Data Mining-Actionable Knowledge Discovery Framework," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3039111.
[38] E. F. Codd, "OLAP On-Line Analytical Processing," in *Controlling-Praxis erfolgreicher Unternehmen*, 1998.
[39] P. Shrishrimal, R. Deshmukh, and V. Waghmare, "Multimedia Data Mining: A Review," *Researchgate.Net*, no. January 2015, 2014, doi: 10.13140/2.1.1511.7129.
[40] E. R. Ziegel, U. M. Fayyad, G. Piatetski-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining," *Technometrics*, 1998, doi: 10.2307/1271414.
[41] A. E. Renshaw, D. M. Titterington, A. F. M. Smith, and H. E. Makov, "Statistical Analysis of Finite Mixture Distributions.," *J. R. Stat. Soc. Ser. A*, 1987, doi: 10.2307/2981482.
[42] Z. Wang, G. I. Webb, and F. Zheng, "Selective augmented bayesian network classifiers based on rough set theory," 2004, doi: 10.1007/978-3-540-24775-3_40.
[43] B. W. Silverman, *Density estimation: For statistics and data analysis*. 2018.
[44] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.
[45] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 4, pp. 140–147, 2020, doi: 10.38094/jastt1457.
[46] S. M. Weiss and C. A. Kulikowski, *Computer systems that learn: classification and prediction methods from statistics Nets, Machine Learning, and Expert Systems*. 1991.
[47] C. Apté and S. J. Hong, "Predicting Equity Returns from Securities Data with Minimal Rule Generation," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. Work.*, 1994.
[48] S. Djoko, D. J. Cook, and L. B. Holder, "An empirical study of domain knowledge and its benefits to substructure discovery," *IEEE Trans. Knowl. Data Eng.*, 1997, doi: 10.1109/69.617051.
[49] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *Int. J. Bio-Science Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013, doi: 10.14257/ijbsbt.2013.5.5.25.
[50] K. Madadipouya, "a New Decision Tree Method for Data Mining in Medicine," *Adv. Comput. Intell. An Int. J.*, vol. 2, no. 3, 2015, doi: 10.5121/acii.2015.2304.

[51] M. Chiu, Y. Yu, and L. C. Hao, “THE USE OF FACIAL MICRO-EXPRESSION STATE AND TREE-FOREST MODEL THE USE OF FACIAL MICRO-EXPRESSION STATE AND TREE-FOREST MODEL FOR PREDICTING CONCEPTUAL-CONFLICT BASED CONCEPTUAL CHANGE,” no. January, 2016.
[52] B. de Ville, “Introduction to Data Mining,” *Microsoft Data Mining*. pp. 1–21, 2001, doi: 10.1016/b978-155558242-5/50003-6.
[53] N. Qader and H. K. Al-Khafaji, “Motif Discovery and Data Mining in Bioinformatics,” *Int. J. Comput. Technol.*, 2014, doi: 10.24297/ijct.v13i1.2932.
[54] X. Hu, “Data mining in bioinformatics: challenges and opportunities,” *DTMBIO 09 Proceeding third Int. Work. Data text Min. Bioinforma.*, 2009, doi: 10.1145/1651318.1651320.
[55] P. Groth, B. Weiss, H. D. Pohlenz, and U. Leser, “Mining phenotypes for gene function prediction,” *BMC Bioinformatics*, 2008, doi: 10.1186/1471-2105-9-136.
[56] L. Conde, Á. Mateos, J. Herrero, and J. Dopazo, “Improved Class Prediction in DNA Microarray Gene Expression Data by Unsupervised Reduction of the Dimensionality followed by Supervised Learning with a Perceptron,” 2003, doi: 10.1023/B:VLSI.0000003023.90210.c8.
[57] Q. Zou, X. Bin Li, W. R. Jiang, Z. Y. Lin, G. L. Li, and K. Chen, “Survey of MapReduce frame operation in bioinformatics,” *Brief. Bioinform.*, 2014, doi: 10.1093/bib/bbs088.
[58] K. Raza, “Application of Data Mining in Bioinformatics,” *Indian J. Comput. Sci. Eng.*, vol. 1, no. 2, pp. 114–118, 2010.
[59] M. V. Schneider *et al.*, “Bioinformatics training: A review of challenges, actions and support requirements,” *Brief. Bioinform.*, vol. 11, no. 6, pp. 544–551, 2010, doi: 10.1093/bib/bbq021.
[60] S. Das, A. Abraham, and A. Konar, “Swarm intelligence algorithms in bioinformatics,” *Stud. Comput. Intell.*, 2008, doi: 10.1007/978-3-540-76803-6_4.
[61] G. Rose and D. J. Barker, “Epidemiology for the uninitiated. Repeatability and validity.,” *BMJ*, 1978, doi: 10.1136/bmj.2.6144.1070.
[62] “251461347_Chapter_5_Case-based_reasoning_driven_gene_annotation.” .
[63] H. Ji and W. H. Wong, “Computational biology: Toward deciphering gene regulatory information in mammalian genomes,” *Biometrics*. 2006, doi: 10.1111/j.1541-0420.2006.00625.x.
[64] F. S. Collins and V. A. McKusick, “Implications of the human genome project for medical science,” *J. Am. Med. Assoc.*, 2001, doi: 10.1001/jama.285.5.540.
[65] B. M. Malone, A. D. Perkins, and S. M. Bridges, “Integrating phenotype and gene expression data for predicting gene function,” *BMC Bioinformatics*, vol. 10, no. SUPPL. 11, 2009, doi: 10.1186/1471-2105-10-S11-S20.
[66] M. Ashburner *et al.*, “Gene ontology: Tool for the unification of biology,” *Nature Genetics*. 2000, doi: 10.1038/75556.
[67] E. C. Dimmer *et al.*, “The UniProt-GO Annotation database in 2011,” *Nucleic Acids Res.*, vol. 40, no. D1, 2012, doi: 10.1093/nar/gkr1048.
[68] T. Z. Berardini *et al.*, “The Gene Ontology in 2010: Extensions and refinements,” *Nucleic Acids Res.*, 2010, doi: 10.1093/nar/gkp1018.
[69] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat. Protoc.*, 2009, doi: 10.1038/nprot.2008.211.
[70] I. Y. Lee, J. M. Ho, and M. S. Chen, “CLUGO: A clustering algorithm for automated functional annotations based on gene ontology,” 2005, doi: 10.1109/ICDM.2005.42.
[71] N. Daraselia, A. Yuryev, S. Egorov, I. Mazo, and I. Ispolatov, “Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks,” *BMC Bioinformatics*, 2007, doi: 10.1186/1471-2105-8-243.
[72] P. Groth *et al.*, “PhenomicDB: A new cross-species genotype/phenotype resource,” *Nucleic Acids Res.*, 2007, doi: 10.1093/nar/gkl662.
[73] P. Groth and B. Weiss, “Phenotype Data: A Neglected Resource in Biomedical Research?,” *Curr. Bioinform.*, 2008, doi: 10.2174/157489306777828008.
[74] J. L. Rodgers and W. A. Nicewander, “Thirteen Ways to Look at the Correlation Coefficient,” *Am. Stat.*, 1988, doi: 10.2307/2685263.
[75] J. D. Wren, “A global meta-analysis of microarray expression data to predict unknown gene

functions and estimate the literature-data divide," *Bioinformatics*, 2009, doi: 10.1093/bioinformatics/btp290.
[76] Y. Zhao and G. Karypis, "Data clustering in life sciences," *Molecular Biotechnology*. 2005, doi: 10.1385/mb:31:1:055.
[77] P. Andreeva, "Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech.," *Int. Conf. Comput. Syst. Technol. - CompSysTech' 2006*, pp. 1–6, 2006.
[78] A. V. S. Ă. Ut, D. Zdrenghea, D. Pop, and D. A. S. Ă. Ut, "Using machine learning algorithms in cardiovascular disease risk evaluation," *J. Appl. Comput. Sci. Math.*, no. April 2020, 2009.
[79] A. Rajkumar and G. S. Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm," *Glob. J. Comput. Sci. Technol.*, vol. 10, no. 10, pp. 38–43, 2010.
[80] K. K, M. N. M, and S. R, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," *Int. J. Data Min. Tech. Appl.*, vol. 7, no. 1, pp. 172–176, 2018, doi: 10.20894/ijdmta.102.007.001.027.
[81] N. C. S. N. I. M Anbarasi, E Anupriya, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370–5376, 2010.
[82] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," *Int. J. Comput. Appl.*, vol. 98, no. 22, pp. 13–17, 2014, doi: 10.5120/17314-7433.
[83] O. I. Obaid, M. A. Mohammed, M. K. Abd Ghani, S. A. Mostafa, and F. T. Al-Dhief, "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," *Int. J. Eng. Technol.*, vol. 7, no. 4.36 Special Issue 36, pp. 160–166, 2018, doi: 10.14419/ijet.v7i4.36.23737.
[84] Z. Ansari, H. Quazi Mateenuddin, and A. Abdullah, "Performance research on medical data classification using traditional and soft computing techniques," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special issue 3, pp. 990–995, 2019, doi: 10.35940/ijrte.B1185.0782S319.