**Vincent Limo**

# A REVIEW OF DATA MINING IN BIOINFORMATICS

**Thesis**
**CENTRIA UNIVERSITY OF APPLIED SCIENCES**
**Information Technology**
**December 2019**

## ABSTRACT

|  |  |  |
|---|---|---|
| **Centria University of Applied Sciences** | **Date**<br>December 2019 | **Author**<br>Vincent Limo |
| **Degree programme**<br>Information Technology |  |  |
| **Name of thesis**<br>A REVIEW OF DATA MINING IN BIOINFORMATICS |  |  |
| **Instructor**<br>Kauko Kolehmainen |  | **Pages**<br>31 |
| **Supervisor**<br>Kauko Kolehmainen |  |  |

In the beginning of the 20th century, commonly known as the information age, there has been a phenomenal growth of potentially deadly group of abnormal diseases such as cancer. Because of this, there is a need for cancer and other biomedical research in and around transcriptomics, genomic and genetics which have a direct application of computer science methods such as data analysis and mathematics.

The aim of this bachelor's thesis is to highlight and discuss in detail the application of data mining techniques in bioinformatics. It begins by discussing the interdisciplinary relationship between data mining, knowledge discovery and bioinformatics before a comprehensive descriptive research in data mining techniques and their application in bioinformatics. The results stablished that gene expression analysis and gene sequencing rely on the application of clustering techniques such hierarchical, fuzzy, graph, and distance clustering while classification techniques, such as machine vector learning, supervised learning, support vector machine and random forest are fundamental in genomic and proteomic synthesizing. It recommends data transformation, cleaning, and scalable statistical models as solutions to the prominent data quality and computational challenges in data mining. This thesis is divided into four main parts, Introduction, Data mining, Application of data mining in bioinformatics and a conclusion.

Pages : 33

**Key words**

Artificial Neural Networks, Bayesian System, Bioinformatics, Clustering, Data Mining, Decision Tree, Fuzzy C, K-Means, K-Nearest Neighbor, Knowledge Discovery, M-Means, P-DBSCAN, Self-Organizing Maps, Support Vector Machine.

## CONCEPT DEFINITIONS

DE - Data Extraction

DNA - Deoxyribonucleic Acid

DP - Data Preprocessing

HGP – Human genome Project

KD – Knowledge Discovery

KDD – Knowledge Discovery from Data

mRNA – Massager Ribonucleic Acid

OSTs – Open Source Tools

RNA - Ribonucleic Acid

ROI –Return on investment

SNP(s) - Single-Nucleotide Polymorphism(s)

CONTENTS

**GRAPHS**

**FIGURES**

**TABLES**

# 1 INTRODUCTION

Scientific boost in the technologies for data generation and collection has resulted in the immense growth of commercial and scientific databases. The new technologies argue in favor for the collection of data across enterprise operations and processes such as banking transactions, digital interaction, online transactions, reservations and diagnostics. The discovery across industries is how they can dissect their databases and key out data patterns and extract information for developing effective enterprise strategies. For example, healthcare databases store bulky data, but there are limited analysis tools to discover hidden knowledge which can assist in thwarting, treatment, and reclamation. Therefore, knowledge discovery and data mining techniques are absolute requisites in the analysis and drawing out of information since traditional tools for data analysis are becoming ineffective and unreliable because of the heterogeneous, complex, large scale and distributed data.

The work of (Chen & Tang & Bouguila & Wang & Du & Li, 2018) and (Aguiar-Pulido & Huang & Suarez-Ulloa & Cickovski & Mathee & Narasimhan, G. 2016) are part of thousands of researches that has been done on the specific disciplines of DA, Data mining and bioinformatics. These works thus, give the direction to the writing of this thesis in order to study the interdisciplinary relationships between the two topics which enhance their competitor advantage by addressing pungent aspects which include real-time processing, as well as enhanced product quality. In addition, market forces and dynamics have shifted towards custom-made services and individualized customer trials. Therefore, enterprises are aiming at techniques which bring forward meaningful and reliable information from the large data sets for purposes of customer satisfaction. Automated processes in data industries are held safely in the collection of diverse data sets and data mining is becoming successful in the analysis of bulky and large datasets.

Technological advancement has propagated knowledge extraction methods which have been adopted in wider ranges of organizations because organization data sets and their sizes have been increasing in recent times. The sizes of hospitals are growing with increasing big data formats of the health industry. People have become aware of their health status and embraced the health industry, therefore, contributing to the generation of massive data in x-ray, billing departments and other significant areas in the health and other big organizations. The organizations have developed systems and databases which store this data, but their challenge is vested in the discovery of useful information from these datasets for decision making.

This thesis has been divided into four parts, a general overview of literature review, an in deep discussion of data mining and data analysis, mining and bioinformatics theory of bioinformatics and clustering techniques used in bioinformatics and lastly a conclusion.

# 2 DATA MINING

The evolution of the Internet as well as furtherance in technologies has stoutly increased the amount of data processing and processed across organizations. Thousands of terabytes data produced yearly across the data enterprises requires safe, secure storage and easy accessibility. Additionally, the large amount of data produced has provided an opportunity for research and development in bioinformatics, tourism, marketing and other pungent organization departments. The large data has led to new fields such as data science where a data scientist processes the massive data and also provides solid and persistent solutions to world problems. Data scientists have developed algorithms and ways which have advanced the analyzing, processing, and presenting data for effective and efficient analysis of data. (Garca & Herrera 2015, 25).

## 2.1 Data analysis (DA)

Data analysis is the process of inspection, cleaning, and transformation of data into useful information for decision making. Enterprises generate voluminous data; therefore; it has become essential to use the data analysis techniques to capture valuable information which can aid enterprises to achieve their intended goals. (Sutton & Austin 2015, 224).

Sutton & Austin (2015, 227) argue that data analysis is anchored in the concept of utilizing technical skills and analytical software to drive insight from voluminous data. The ability of the data scientists to quickly analyze institution data avails an opportunity to understand businesses and projection of market trends. Companies comprehend customer’s purchasing behaviors; analyze their performance, document advertisement drivers, and market trends through computational analysis. There are different techniques and approaches to data analysis. These techniques include; exploratory data analysis, descriptive statistics, and confirmatory data analysis.

### 2.1.1 The exploratory analysis

The exploratory analysis uses the visual methods to summarize the main characteristics of the dataset. The exploratory analysis technique provides additional information about the datasets beyond the hypothesis and the formal modeling. Exploratory data analysis is critical in uncovering the underlying data structures, extracting significant variables, maximizing insight in the dataset, detecting anomalies and outliers in the massive data sets, developing parsimonious models and testing underlying assumptions in the dataset (Martinez & Solka 2017, 3-5.). Exploratory data analysis uses simple graphical techniques which incorporate the plotting of raw data such as histograms, block plots. Simple statistics such as the median, standard deviations, mean are using these plots to maximize the abilities to recognize natural patterns. (Rogalewicz & Sika 2016, 226)

### 2.1.2 Descriptive statistics

Descriptive statistics is a technique used to describe or summarize data patterns. Descriptive statistics limit us to the data we analyze, and therefore, we cannot reach a hypothesis about our problem statement. Descriptive statistics techniques are significant in aiding us to visualize the dataset and make a simple presentation and interpretation. There are two forms of descriptive statistics which include measures of central tendency and measures of spread. Measures of central tendency include mean, mode frequency, and median. Measures of spread include quartiles, range, variance, and absolute deviation (Aguiar-Pulido et al. 2016).

### 2.1.3 The confirmatory data analysis

The confirmatory data analysis technique is based on quantifying changes in case there is a deviation from your model. The evidence is evaluated using statistical tools such as inference and significance. The main themes in confirmatory data analysis involve the provision of evidence using regressions, variance analysis, and testing a hypothesis. The researcher findings and arguments are tried and confirmed (Chen et al. 2018).

According to (Barrett & Salzman 2016, 847) Data analysis requires technical insights to identify poignant information from massive data for decision making.  Data analysis involves phases which are fundamental in data cleaning and transformation to facilitate the actualization of organizational goals and objectives. The data analysis phases include; data cleaning, quality analysis exploratory confirmation, stability of the results, Statistical methods, and knowledge presentation. Poignant issues in data analysis include data preparation and the evaluation of classification methods. Data preparation involves data cleaning, relevance analysis and data transformation (Han, Pei & Kamber 2012, 84).  Data cleaning encompasses the preprocessing of data to identify missing values, incorrect data and deleting duplicated data. Relevance analysis is when the redundant attributes are removed from the data sets, and data transformation involves normalization of data and the data is converted into a cleaned format that can support decision making in the organizations (Han et al. 2012).

Predictive analysis tools provide solutions to organizations through the processing of their massive data and making predictions of the unknown future events. Predictive analysis uses predictive modeling to integrate information technology, business processes, and management to project future business operations (Han et al. 2012, 89). Transactions of data patterns and historical characteristics are used as a baseline to aid in the identification of future opportunities and risks. The predictive analysis operates by capturing relationships across different factors to assess risks among processes and assign a score, therefore, allowing the business to interpret, analyze and make a profound decision (Han et al. 2012).
Healthcare and other organizations always aim at avoiding risks which are likely to lead to greater loss therefore, predictive analytics have become integral in lawfully managing risks through insightful analysis of future data patterns (Rogalewicz & Sika 2016).

## 2.2 Knowledge Discovery

Knowledge discovery tools are an emerging trend in data analysis, they enables the business to extract big structured and unstructured data which is stored in multiple databases. Most of the organization file systems are based on different forms such as; database management system (DBMS), application program interfaces, and customized platforms (Pei & Kamber 2011, 26).

The advancement in technologies has boosted the adoption of approaches such as knowledge discovery and data mining, which are poignant in the discovering and extraction of information from massive data collected across enterprise operations and stored in the databases. Data mining is the development and application of the algorithms to identify and analyze data patterns. The fundamental objective of data mining is to identify peculiar data patterns and their relationship from large data sets and extract information for alignment and realignment of organization strategies. Data mining requires technical expertise in developing algorithms and models which can explore massive data sets (Pei, & Kamber 2011, 23).

### 2.2.1 Data Mining and Knowledge Discovery

Personalized customer experiences have become prominent due to a shift in the individual customer demand where customers expect an individual response to their purchasing and consumption behaviors, treatment and rehabilitative care, and other services across enterprises. In response to this paradigm shift, companies have changed their approaches in the provision of services to address personalized customer experiences. The advancement in technologies has boosted adoption of approaches such as knowledge discovery and data mining, which are poignant in the discovering and extraction of information from massive data collected across enterprise operations and stored in the databases (Pei, & Kamber 2011, 23).

Data mining is the development and application of the algorithms to identify and analyze data patterns. The fundamental objective of data mining is to identify peculiar data patterns and their relationship from large data sets and extract information for alignment and realignment of organization strategies. Data mining requires technical expertise in developing algorithms and models which can explore massive data sets. The key methods in data mining are prediction and description methods. Prediction methods are essential in using variables to predict future values or unknown concepts. Description methods are based on the concept where you identify patterns to describe data. (Aguiar-Pulido et al. 2016).

The techniques used in data include decision trees and rules, probabilistic models, example-based methods, nonlinear regression, and classification method, and relational learning models. (Aguiar-Pulido et al. 2016).Its challenge for the human to manual processing a large amount of data generated in the organizations, therefore, it has become necessary to adopt the automated system which supports real-time processes, higher processing rates, improved quality, and voluminous production. Data mining methods enhance the real-time access of data. The automated systems are rapid; therefore, information is generated within the stipulated organization time. The knowledge hidden in the voluminous data is accessrapidly, unlike in the manual method which takes a lot of time. Bibliomining has been significant in the tracking of the data pattern behaviors in the library system and discovering important library information in the historical data (Nicholson 2006, 10).

Data mining methods are dedicated to unlocking insights across a range of processes in different industries. For instance, data mining has become effective in criminal identification, detection of frauds in financial industries, preparation, and generation of prognoses and diagnosis in health care. Data mining modules like data preprocessing (DP), data extraction (DE), clustering, Google map representation, classification, and WEKA_ implementation are used for criminal identification (Tayal et al., 2015, 117). Additionally, data mining is essential for organizations to deliver the right products and services to their customers. Companies can identify those areas and regions where their products are highly consumed. Furthermore, data mining techniques are significant in the identification of the customers and their experiences. Organizations can create personalized experiences for their clients. Data scientists help the organizations to understand their customers at their granular level; therefore, develop marketing strategies based on the customer's experiences. The products experiences enable the customers to make request and demands of the products (Tayal et al., 2015, 123-124).

Knowledge is a form of information which is significant to the organization processes and operations. The process of extracting knowledge from massive data is based on the development of the sophisticated methods and techniques which process the voluminous data to make sense to the organization management. The knowledge discovery concept is essential because some of the data are complex and need to be analyze to generate a meaningful decision. Knowledge discovery is based on data mining techniques to extract meaningful information from data (Kumar & Chatterjee 2016, 24). Knowledge discovery is an interdisciplinary field, which means it works in collaboration with another field such as data mining to extract useful information for the organizations. Therefore, data mining is a necessity in knowledge discovery. In knowledge discovery, massive data is explored to identify patterns which can generate useful decisions for the organizations. Traditional approaches such as deductive databases were expensive and slow in the analysis and the interpretation of data. Disciplines such as medicine has complex data sets. Therefore, they require robust models and techniques to generate reliable information and knowledge for decision making.

Knowledge discovery is the science of extracting significant information which was previously unknown, while data mining is the actual steps which are used in knowledge discovery, the scientists apply algorithms to extract patterns from massive data for decision making. Data mining is a guarantee of the extraction of poignant knowledge and information from massive data. It aids in the finding of the new information or knowledge which is exciting for the organization. As illustrated in figure 1, data mining processes include five very important steps. Firstly, data cleaning and preprocessing which is done by removal of outliers, identification of the missing values and transformation. Secondly, data integration which is done by combing different data sources into not more than a fifth of the whole data sample. Thirdly, data selection; this involves retrieval of necessary data which is relevant to the task. Fourthly, data mining; this involves application of intelligent methods to extract relevant data patterns. Lastly, knowledge presentation; it involves the use of visualization technologies to present mined knowledge Han et al., 2012, 7).

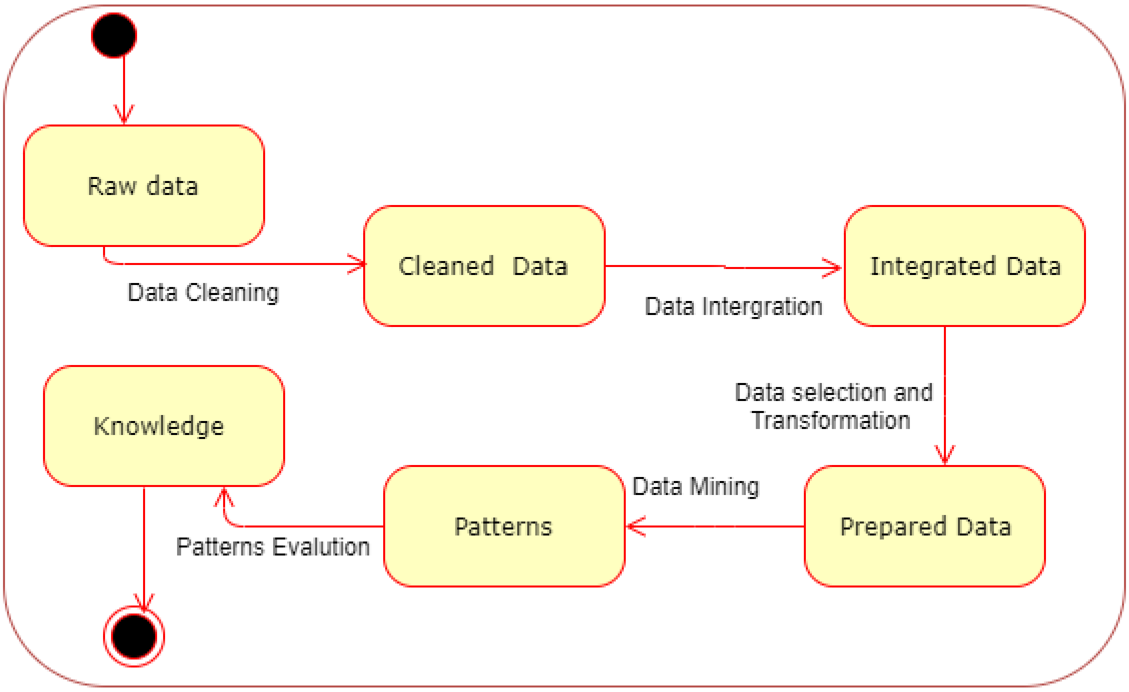


FIGURE 1: Data mining process (Han et al. 2011, 5)

According to Han et al., (2011), data mining utilizes techniques from many disciplines such as machine learning, statistics, database systems, data warehouse, visualization, and algorithm. Data mining has an inherent relationship with statistics. Data mining uses statistical models, statistical descriptions, and predictive statistics to identify missing values, to describe data patterns and draw inferences about organizations processes (Han et al., 2011, 23). Machine learning focuses on how computer programs automatically identify complex data patterns and establish intelligent decision beneficial to the enterprise. Data mining relates to machine learning by adopting supervised learning, semi-supervised learning, or unsupervised learning.

In this approach, the aim is to identify the structures, patterns, and knowledge in the unlabeled data. There is an input data in the unsupervised learning, but there are no corresponding output variables; therefore, there is an in-depth understanding of the data structure to generate meaningful information (Buczak & Guven, 2015). Unsupervised learning is digested into clustering and association methods. In clustering, the scientist aims at understanding the inherent data groupings, which involve maybe a grouping of the patients according to their diagnosis or the customers according to their purchasing behaviors. Association methods are anchored in the discovering of the rules and procedures which describe large data portions, for instance, clients buying product X with the possibilities of purchasing product Y. or multiple information of having a possibilities of ailment X and Y. Unsupervised learning algorithms which are popular include K, which is predominately used for clustering problems and apriori for the association. (García, Luengo& Herrera 2015, 7-8.) Unsupervised data is essential in learning the structures and variables in the large data sets.

In this method, there is the labeling of the portion of data during the acquisition by human experts. The data mining experts have a large amount of data X and labeled portion Y. for instances, using archive data where a portion is labeled, and over half of the data is unlabeled. Labeling data in organizations is expensive and time-consuming (García, Luengo & Herrera 2015, 9). Additionally, the process requires data scientists to have access to the domain experts. Unlabeled data is flexible to access, collect, and store.

All the dataset in the organization database is labeled, and the chosen algorithms predict the outcome from the input data (García, Luengo& Herrera 2015, 6). Additionally, the supervised techniques can be used to make predictions for the unlabeled data and input the data in the supervised learning algorithm as training data to design prediction of the unseen data.

As shown in figure 2 below, database system and data warehouse are significant disciplines which relate to data mining. The concept behind the database system is to create, maintain and use the database for enterprises and their end-users. The database features such as scalability to accommodate large datasets, structured datasets, and techniques such as query languages, data store, and indexing are significant components during data mining. Data mining takes advantage of the scalability of the databases technologies to achieve efficiency. A data warehouse aids in the consolidation of data with different formats, therefore, aiding in the multidimensional data mining.

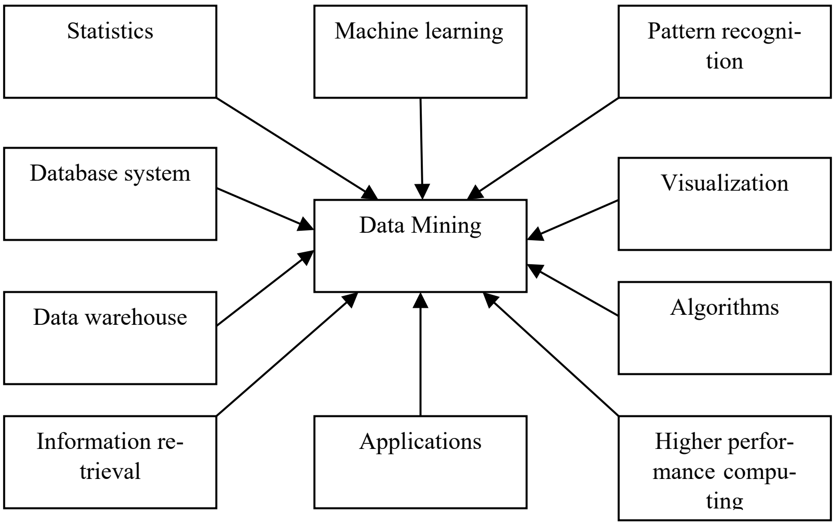


FIGURE 2: Data mining adopts techniques from many domains (Adapted from Han et al., 2011).

Extensibility is a paramount feature of a data mining systems since it is essential in helping the data mining system to keep up with the variability of the task involved in the data mining process. The basic feature in extensibility is adding features in the data mining system without reprogramming the core components. This allows the other developers or the third parties to extend the existing system without prior understanding of the internal process of the system. Achieving extensibility in data mining requires having design API s, and description of the declarative tools to allow the kernel support extensions (Petermann et al., 2016 1316). Additionally, a task manager is an essential component of the data mining system because of the different tasks and the variety of methods used in data mining.

### 2.2.2 Uses of Knowledge Discovery Tools

Knowledge discovery tools isolate deep perception of information from these platforms for facilitation in the decision-making process for the business. For instance, knowledge discovery tools such as virtual analysis are used in the health industry to exploit clinic-genomic data. Objectives of the knowledge discovery tools include accomplishment, extensibility, and serviceability. Accomplishment; the analysis system to be robust enough to facilitate and support large analysis and optimization which requires an effective software architecture extensibility the tools should provide a plan where they can be extended to other pungent analysis problems in the organization and serviceability; the system must be user-friendly to the new users and experts in the organizations.

Data virtualization allows the retrieval of data and manipulation without considering the technical details of the datasets, such as the format, location, and source (Katragadda et al., 2019, 3). Data virtualization supports real-time processing; therefore, minimizing data errors, heavy workload and data virtualization also allows for the documentation of the transaction data updates to the source system. It's essential in the retrieval of data without the consideration of the technical restrictions which might include the physical location of data and data formats.

Data analysis has an enormous benefit to the business as it poses power to transform enterprise to realize their objectives and goals. Fundamental areas which data analysis has shown its benefits encompass the prediction how the trends and behaviors of the customers increase the productivity of the companies since they can identify meaningful information which enhances the organization's competitive advantage. In business, data analysis is an important aspect of the prediction of customer's preferences and market analysis. Technologies such as predictive analytics, knowledge discovery tools, stream analytics, NoSQL databases, data preprocessing, data virtualization, data integration, and data quality are essential, and they have provided an opportunity for easy analysis of massive data.

## 2.3 Application of Data Science

Data mining, data analysis, and knowledge discovery techniques have been developed by data scientist and are being applied in our day to day engagements as follows; Health industry uses an electronic health record system in form of application software across the hospitals for capturing, recording, manipulation and management of the patient medical progress records. Digitization of the patient data has assisted doctors in diagnosis, prescription, and dispensation of drugs. Recent space missions in data explorations have resulted in the generation of weighty data. (Raghupathi, 2014, 7).

Data scientists have been helpful in storing data and analysis to facilitate favorably successful future space explorations. Prediction algorithms have been used to facilitate the implementation of new business strategies and logics. Energy companies are using the data algorithms to manage the demand and supply, and to ensure that the supply and demand curve goes in handy with the organizational needs resulting in effective use of resources, maximum utilization of resources, enhanced productivity and profitability. The new technologies are changing significantly due to elevation in the prediction algorithms and data mining means. They are like artificial intelligence (Raghupathi, 2014, 23-25).

## 2.4 Data Mining Examples

The predictive capacity of data mining has changed the design of business strategies. Now, one can understand the present to anticipate the future. These are some examples of data mining in the current industry.

### 2.4.1 Marketing

Data mining is used to explore increasingly large databases and to improve market segmentation. By analyzing the relationships between parameters such as customer age, gender, tastes, etc., it is possible to guess their behavior in order to direct personalized loyalty campaigns. Data mining in marketing also predicts which users are likely to unsubscribe from a service, what interests them based on their searches, or what a mailing list should include to achieve a higher response rate (D’Souza & Minczuk 2018)

Also consider a marketing head of a telecom service provider who wants to increase revenues of long-distance services. For high ROI on his sales and marketing efforts, customer profiling is important. He has a vast data pool of customer information like age, gender, income, credit history, etc. But it is impossible to determine characteristics of people who prefer long distance calls with manual analysis. Using data mining techniques, one may uncover patterns between frequent long-distance call users and their characteristics. For instance, he might learn that his best customers are married females between the age of 45 and 54 who make more than $80,000 per year. Marketing efforts can be targeted to such demographic (Han et al. 2012).

### 2.4.2 Banking

Banks use data mining to better understand market risks. It is commonly applied to credit ratings and to intelligent anti-fraud systems to analyze transactions, card transactions, purchasing patterns and customer financial data. Data mining also allows banks to learn more about our online preferences or habits to optimize the return on their marketing campaigns, study the performance of sales channels or manage regulatory compliance obligations. (Sindhu & Sindhu 2017).

Another example is bank wants to search new ways to increase revenues from its credit card operations. They want to check whether usage would double if fees were halved. In this case, the bank has a multiple year of record on average credit card balances, payment amounts, credit limit usage, and other key parameters. They create a model to check the impact of the proposed new business policy. The data results show that cutting fees in half for a targeted customer base could increase revenues by $10 million (Sindhu & Sindhu 2017).

### 2.4.3 Medicine

Data mining enables more accurate diagnostics. Having all of the patient's information, such as medical records, physical examinations, and treatment patterns, allows more effective treatments to be prescribed. It also enables more effective, efficient and cost-effective management of health resources by identifying risks, predicting illnesses in certain segments of the population or forecasting the length of hospital admission. Detecting fraud and irregularities and strengthening ties with patients with an enhanced knowledge of their needs are also advantages of using data mining in medicine (Dua & Chowriappa 2012).

### 2.4.4 Telecommunication

There are networks that apply real time data mining to measure their online television (IPTV) and radio audiences. These systems collect and analyze, on the fly, anonymous information from channel views, broadcasts and programming. Data mining allows networks to make personalized recommendations to radio listeners and TV viewers, as well as get to know their interests and activities in real time and better understand their behavior. Networks also gain valuable knowledge for their advertisers, who use this data to target their potential customers more accurately (Dedić & Stanier 2016).

# 3 DATA MINING AND BIOINFORMATICS

This chapter explores different data mining techniques and how they are applied in bioinformatics. The data mining technique discussed includes clustering, classification, sequencing, and analysis. The first section is an introduction to and history of bioinformatics, the second section examines clustering techniques such as fuzzy C, K-means, M-means, and how they are being applied in bioinformatics. The first section ends with an in-depth examination of clustering in gene analysis. The second part focuses on classification techniques such as supervised learning, support vector machine and their applications in proteomics and genomics. The chapter concludes biological database and integration.

## 3.1 Bioinformatics

Bioinformatics is a field of science that utilizes computer methods in the collection, analysis, and management of the biological data. Progress made in genomic technology and molecular biology has skyrocketed the quality and quantity of biological information. Bioinformatics is essential in the creation of the human genome databases, indexing, and organization of the biological information to aid in the analysis. Technology adoption in analyzing and extracting significant information from the biological molecules and sequences is vital since biomedical data is derived from a multi-dimensional and structural data unit which includes microscopic unit, laboratory research, field research, and the omics world that is proteomics, lipidomics, metabolomics, genomics, transcriptomic, fluxomic, and phenomics (Raza 2012, 114). Macroscopic information include data on public health informatics. The healthcare information is complex; for instance; the glucose molecule has the approximate size of 900 pm = 900 × 10-12m. Additionally, bioinformatics databases contain patient's health informatics retrieved from their health record as displayed in figure 3 below.

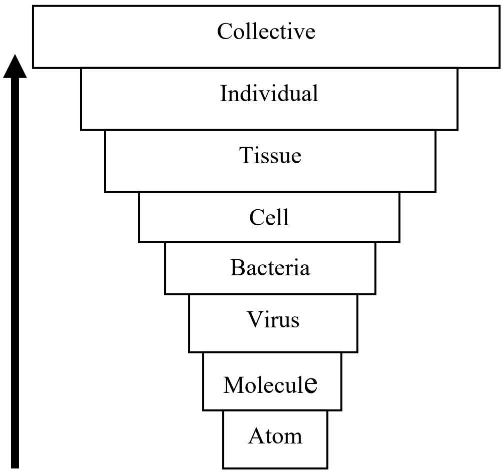


FIGURE 3: Complex healthcare data (Raza 2012).

Protein plays a significant role in the body organisms. Scientists use bioinformatics to determine the interaction of various proteins in the human body, which is essential in determining how diseases develop within the human body. The adoption of the sophisticated computerized technologies has become vital in the location of genes and their manifestations, which has greater benefits in the scientific advancements, hence, providing scientists with adequate time to test the hypothesis (Raza 2012, 115). Biomedical data is heterogeneous, inconsistent, and distributed. Furthermore, many errors are generated during the manual analysis of the complex data units. Data mining in the bioinformatics aims to ensure significant information is generated from the complex health. Currently, biomedical engineers are challenged by the complexity of the data. Therefore, they seeking data mining and knowledge discovery techniques to perform end-user analysis and derive biological solutions. (Raza 2012, 114-115).

### 3.1.1 Human Genome Project (HGP)

The Human Genome Project (HGP) was a global scientific research project with the objective of determining the base pairs that constitute human DNA, and of identifying and mapping all of the genes of the human genome from a physical and a genetic standpoint. The main agenda of the Human Genome Project were first articulated in 1988 by a committee of the U.S. National Academy of Sciences, and was later adopted through a detailed series of five-year plans written jointly by the National Institutes of Health and the Department of Energy of USA (D'Souza & Minczuk 2018).

Congress funded both the NIH and the DOE to work on further exploration of this concept and to further try to conceptualize it, the two government agencies made an agreement by signing a Memorandum of Understanding to coordinate research and technical activities related to the human genome (Raza 2012, 117.) The HGP has shown that there are about 20,500 human genes which determine characteristics in a human. This product of the HGP has given the world a resource of detailed information about the structure, organization and function of the complete set of human genes. This information can be thought of as the basic set of inheritable instructions for the development and function of a human being. The tools created through the HGP also tell more about efforts to characterize the entire genomes of other organisms used in biological research. These efforts support each other, because most organisms have many similar genes with similar functions (D'Souza & Minczuk 2018). Advanced methods for disseminating the information generated by the HGP to scientists, physicians and others, is necessary in order to ensure the most rapid application of research results for the benefit of humanity. Biomedical technology such as bioinformatics and research are beneficiaries of the HGP.

### 3.1.2 Transcription and Translation

The two main processes in gene expression are transcription and translation. Transcription is the first level functional control that occurs within non-protein genes. The process involves the initiation of the mRNA biosynthesis by binding an RNA polymerase on the DNA sequence (D'Souza & Minczuk 2018, 310). The variation of the DNA sequence affects the strength of transcription. Technology has provided an opportunity for the development of the synthetic promoters that can produce transcriptional strength desired by the bioinformatics.

On the other hand, translation involves the converging of mRNA molecule into a protein. Translation occurs in the three steps that include initiation, elongation, and termination (D'Souza & Minczuk, 2018, 311). Technology has provided an opportunity to design and control the individual gene expression level at the transcription and translation levels. Measuring the mRNA level using tools such as serial analysis of the gene expression, microarrays are used in determining the expression of a gene (D'Souza & Minczuk 2018, 314).

## 3.2 Clustering Techniques

Clustering is the statistical technique that ensures items are assigned to clusters where items assigned in the same cluster are similar while those assigned to the different clusters are as different as possible (Manikandan et al., 2018). A similarity measure is an effective technique used to identify different items in clusters.

Clustering in bioinformatics is determined using the Ben-Hur methods that advocates four steps for effective in determining the number of clusters in bioinformatics; estimating cluster numbers, partitioning of the sampling into K clusters, identifying the subgroups and placing them into the k clusters and calculating the correlation for each of the subset then mapping the correlation coefficients of each the subsets to determine the actual number of the clusters in the biological data (D'Souza & Minczuk 2018).

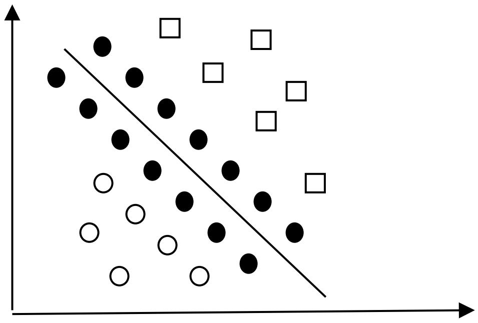
The distance metric is used in incorporating known gene functions and establishes whether genes can share common gene functions. The gene expression-based distance is shrunk towards zero by the distance metric when it establishes that the common gene functions can be shared. The mutual information measure of different data sets is taken to avail significant information, which can help in finding negative, positive and nonlinear correction on gene data. (Dua & Chowriappa 2012, 183-190.) Fuzzy, lso called non-hard clustering, is the clustering technique where the objects are assigned to more than one cluster. The data point in the non-hard clustering is where the data points are assigned to multiple clusters according to their degree association (Manikandan et al., 2018, 1814)

### 3.2.1 Fuzzy C

The Gath Geva (GG) algorithm expansion has been proposed to deal with information with numerical characteristics. The trial appraisal of the plan acknowledgment execution of the anticipated model with different applications has uncovered that it beats different algorithms for data clustering with changed numeric as well as downright characteristics. This radical algorithm should have capacity to process any sort of datasets either numerical, graphical or literary (Han et al. 2012).

Datasets should be managed by this clustering algorithm making it an inclusive algorithm that can manage all sort of information types. GG algorithm is a well-known method for numerical data clustering in fuzzy c means system which are dependent on the supposition that GG created groups are more adaptable than firebase cloud messaging (FCM) produced round groups. To deal with blended classification and numerical characteristics information; customarily fuzzy k models algorithm is utilized which is an encompassing form of FCM but it doesn't utilize the same disparity of work (Sutton & Austin 2015).

FCM clustering algorithm execution in a productive way is exhibited. FCM algorithm has shown an enhanced proficiency for Uniform appropriation of information focuses on Execution of results for manual appropriation and uniform circulation have uncovered negligible contrast along these lines productive mean of uniform dispersion should be investigated. They are demonstrating a good utilization of FCM clustering algorithm on three sorts of data sources for instance, information focuses which are first circulated, and statistical dispersions of ordinary information that focuses on utilizing the Box Muller equation and statistical disseminations of uniform information focusing on utilizing the Box Muller recipe on a given algorithm (Dedić & Stanier 2016).



GRAPH 1: Fuzzy clustering (Adapted from Chen et al. 2018).

### 3.2.2 K-Means

PCA has been connected on the dataset preceding the utilization of clustering technique to obtain the underlying centroid and clustering data into lower measurements. Use of three important parts alongside use of PCA strategy scrutinized about 99.48% of prepared information consistency causing absolute minimum loss of information with part of measurement decrease (Pei & Kamber 2011).

The proposed method are connected to numerous sorts of informational collections to evaluate the genuine potential. It is recommended that the proposed system might be tried/tested/connected to an assortment of datasets to explore new roads and potential outcomes. K-means algorithm application results rely upon the underlying estimation of cancroids. To discover beginning centroid for k-means the creator has proposed the use of Principal Component Analysis (PCA) for dimensional decrease of the datasets and heuristics way to deal with (Sindhu & Sindhu 2017).

### 3.2.3 M-Means

M-Means is a methodology which proposed Solution for K-Means clustering algorithm has been connected as an effective and straightforward instrument to screen execution of understudies. M-means Strength Proposed utilization of K-Mean clustering algorithm. Its weakness is the proposed use of the two methods don't present any alteration to decrease exertion repetition and assets required for its application. Suggestive improvements for the utilization of this method, efficient centroid assurance system to reduce redundant efforts required for random sampling technique should be incorporated. The procedure proposed isn't just a model for scholastic figures, however, its enhanced adaptation of the current models by evacuating their confinements (Rogalewicz & Sika 2016).

The current techniques portrayed in this paper are fuzzy models which utilizes the dataset of just two course results to anticipate understudies' scholarly practices. Another methodology depicted is harsh set hypothesis to investigate understudy information utilizing the Rosetta toolbox. The reason for utilizing this toolbox is to survey information in connection to recognizing relationship between the influencing components and understudy review (Rogalewicz & Sika 2016, 232-233)

### 3.2.4 P-DBSCAN

The P-DBSCAN algorithm has been exhibited as an enhancement of the BSCAN clustering algorithm for and preparing gathering of geo-labeled photos. Specialized for the issue of investigation of spots and occasions utilizing a wide accumulation of geo-labeled photographs (Nicholson 2006, 786) Distinctive parts of the methods that were proposed were not specified in this paper as it is a continuous progressing research. Effort needs be centered on appraisal approaches, and database joining. It led to the proposition of another clustering algorithm P-DBSCAN which is based on a unique DBSCAN clustering algorithm to process and investigate geo-labeled pictorial information of occasions and places of Washington, D.C (Nicholson 2006).

The two upgrades in unique meaning of DBSCAN presented an adaptive thickness way to deal with upgrade look for thick zones and fast connection of calculation with high thickness groups. All characterized a thick benchmark dependent on the measurable figures of individuals taking picture in the area. The perception can be executed on strategies empowering the client to see the various kinds of data in one necessity of looking at figures. The proposed method isn't just ready to uncover comparable information and question gatherings. Additionally, it encourages representation of comparable information protests in chart arrangement of indistinguishable datasets by applying least spreading over trees. This method is well able to recognize the comparable property bunches based on charts drawn utilizing either input information or the SOM nodes (Chen et al. 2018).

### 3.2.5 Self- Organizing Maps (SOM)

SOM based component clustering technique was exhibited to help uses and procedures in investors' investigation portfolio to recognize comparative restocks and listing the stock on the stipulated time. The principle point of building a portfolio is to widen the financial specialist's profile by restricting the buy of unnecessary stock because it is dangerous to put resources into the weight of comparable conduct (Sindhu & Sindhu 2017).

K-Means is one of the well-known clustering procedures because it is straightforward and productive. The point when K-means is used to cluster expansive datasets is exact to the issue at hand. K-Means and Self-sorting out maps to suit for a wide range of information constrained K-Means Clustering has been proposed as an enhanced adaptation of K-means (Aguiar-Pulido et al. 2016.) Assessment consequences of the proposed method show  a critical upgrade in productivity and nature of bunching when contrasted with varied systems. The proposed procedure requires to improve its instructions for clamor decrease. As such, proper imperatives and instruments should be characterized for commotion decrease in the Constrained K-Means.

## 3.3 Machine Learning Algorithms for Bioinformatics

This subtopic introduces machine learning classification algorithms and how they can be used in bioinformatics. In brief, techniques such as deep learning enables the algorithm to use the already existing algorithms to combine several input data to a more important set of features. As shown in figure 4 below, this section discusses the decision tree, Bayesian network, Support Vector Machine, K-Nearest neighbor and neural Network (Rogalewicz & Sika 2016).

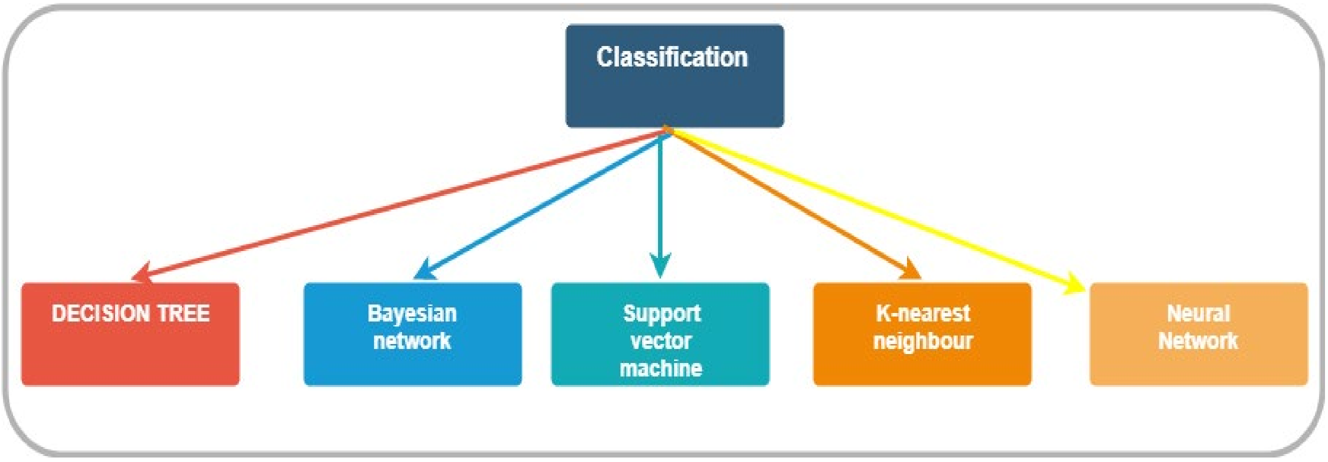


FIGURE 4: Bioinformatics Clustering Techniques (Rogalewicz & Sika 2016).

### 3.3.1 Support Vector Machine (SVM)

Support vector machines (SVM) are a sort of machine learning apparatus, a help vector machine builds a plane in a large dimensional space, which can be utilized for classification, relapse, or different specifications. SVMs were first connected to protein arrangement order and have been connected to remote homology recognition too. SVMs are supervised parallel classifiers used to locate a straight partition between various classes of focuses in 3-D space (Aguiar-Pulido et al. 2016).

This locates an ideal isolating plan among individuals and non-individuals from a given class in a conceptual space. SVM'S are connected to quality articulation information starts with a collection of known classifications of genes (Aguiar-Pulido et al. 2016.) A property of SVM at the same time cost the experimental arrangement mistake and boost the geometric edge. So SVM is also called Maximum Margin Classifiers. The condition appeared underneath is the hyper plane: aX + bY = C , it’s applied as shown below,

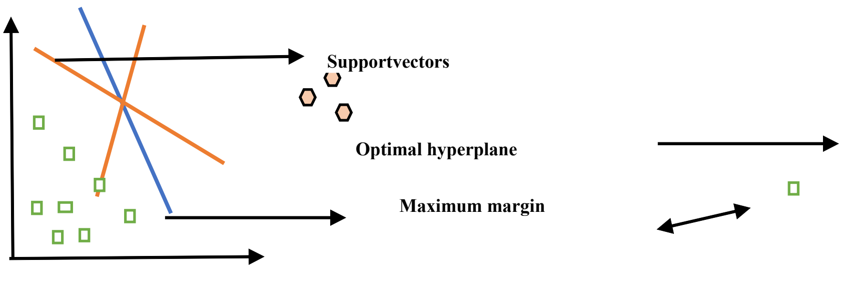


Figure 5. Application of SVMs (Adapted from Han et al. 2012).

### 3.3.2 K – Nearest Neighbor (KNN)

The KNN Algorithm usually depends on closeness measure and is used to store every single open case and used to know the obscure information point dependent on the closest neighbor. It is straight and precise as it gives very good work in fields and practice, particularly in classification. The weighted nearest neighbor classifier (wk-NNC) is a strategy which adds a support to every one of the neighbors in a classification. K-Nearest Neighbors utilizes distance function as shown in the equations below (Han et al. 2012).

$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$Encliden

$\sum_{i=1}^{k}|x_i - y_i|$Manhattan

$[\sum_{i=1}^{k}(|x_i - y_i|)^q]^{1\backslash q}$——— Minkowski

Where n represents the number of training patterns, K-Nearest Neighbor Mean Classifier (k-NNMC) and the Hamming Distance $D_H = \sum_{i=1}^{k}|x_i - y_i|$

X=Y→ D=0 X≠Y→ D=1

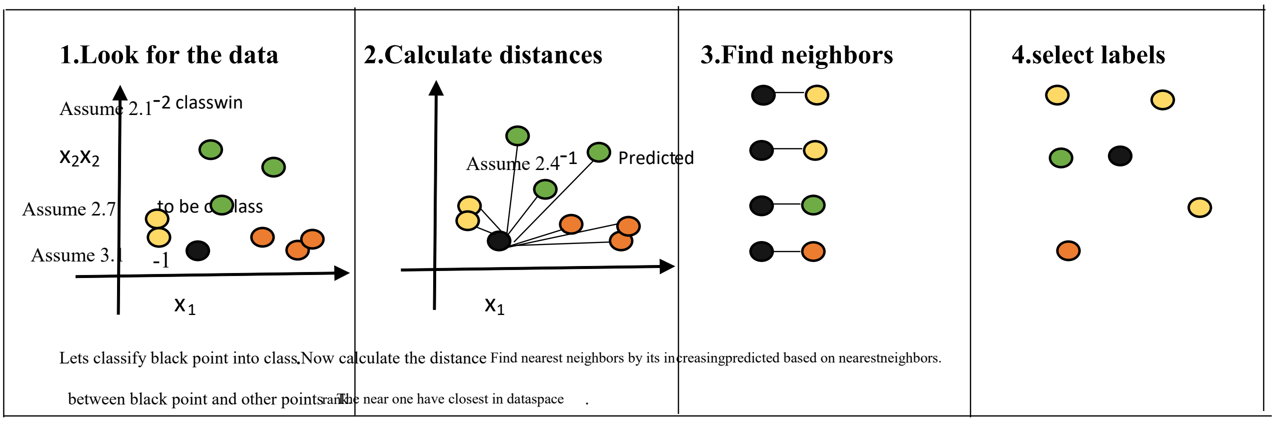

1.Look for the data
2.Calculate distances
3.Find neighbors
4.select labels
Assume 2.1-2 classwin
X2X2
Assume 2.7 to be c lass
Assume 3.1 -1
X1
Assume 2.4-1 Predicted
X1
Lets classify black point into class.Now calculate the distance Find nearest neighbors by its increasingpredicted based on nearestneighbors.
between black point and other pointsrankThe near one have closest in dataspace .

FIGURE 6: KNN Algorithm (Sindhu & Sindhu 2017)

### 3.3.3 Decision Tree

The decision tree is the most utilized data mining methods because of its straightforwardness to comprehend and utilize. The base of the decision tree is a condition that has varied solutions where each solution gives an arrangement of conditions to help process the data so that the conclusion can be made. In addition, decision tree point to a various leveled model of choices and their expense (Dua & Chowriappa 2012).

At the point when a tree is utilized for arrangement, at that point it is said to be as a classification tree. ID3 (Iterative Dichotomiser3), C4.5 Algorithm, CART (Classification and Regression Tree) ID3 are a standout amongst the most important decision tree algorithms (Rogalewicz & Sika 2016, 101). Data gain ahead of time and for the most part to decide appropriate property for every hub of a produced decision tree. One can choose the characteristics with the bewildering data as the test property dependent on current node.

### 3.3.4 Bayesian System

The Naive Bayes algorithm is an honest classifier that is being utilized to ascertain an arrangement of probabilities by utilizing mixes of qualities in an informational collection (Dedić & Stanier 2016.) It is a graphical model for likelihood connections among an arrangement of factors. This consists of two segments, the first segment is principally a coordinated non-cyclic which contains nodes known as the random factors and the edges between the nodes.

The second part which contains an arrangement of parameters that portray the contingent likelihood of every factor given its parents. Naive Bayes classifiers can be prepared exceptionally well in learning and this strategy is vital for a few reasons (Sindhu & Sindhu 2017)

The formulae is applied as shown below,

$(|) = \frac{\text{ь}(x)}{\text{ь}(x|c)\text{ь}(c)}$ Where, P(x/c) - Likelihood,
P(c) - Class Prior Probability,
P(c/x) - Posterior Probability,
P(x) - Predictor Prior Probability.

$P(C|X) = P(X_1|C) \times P(X_2|C) \times \ldots\ldots.. \times P(Xn|C) \times P(C)$

POSTERIOR = PRIOR×LIKELIHOOD/EVIDENCE

Where posterior is where the anticipated occasion will happen, Prior is past understanding, likelihood is conceivable of shot and evidence sums up to the number of occasions that will happen.

### 3.3.5 Artificial Neural Networks

A neural system is a blend of hubs related in a topology with every hub that contains info and yields associations with different nodes and hence Neural Networks are used in model acknowledgment and classification of the neural systems working with straightforward individual handling components are able to perform complex techniques. When given the relating input vector, the perception in a solitary layer neural system whose weights and inclinations generates a right yield (D'Souza & Minczuk 2018).

ANN is used in comprehending the organic neural systems and in tackling man-made consciousness issues. The issues can be explained without utilizing an organic framework in light of the fact that the genuine, natural apprehensive are exceedingly entangled. The ANN algorithm strivers to outline this uncertainty and spotlight on actual yet a large portion of the data are from preparing perspective (Barrett & Salzman 2016)

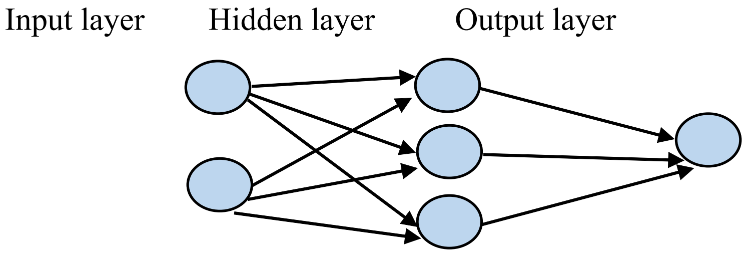


FIGURE 7: Artificial neural network (Chen et al. 2018).

An interconnected gathering of neural or counterfeit neurons that utilizes a numerical or computational model for data preparing dependent on a connectionist way to deal with algorithm is on account of artificial neurons called (ANN). ANN is a versatile framework that changes its structure dependent on outer or inside data moving through the system (Han et al. 2012.) Neural systems are customized to store, see and recover examples or database sections for taking care of poorly characterized issues, to channel dissatisfaction from data estimated (Masood & Khan 2015).

### 3.4 Perl and Python Libraries

Bioperl and Biopython are OST which has been developed to solve problems in life sciences. The two high level languages; perl and python are used widely in research, educational purposes and business. On choosing the best programming language to be used in laboratory research and bioinformatics methods genomic sequence, 3-dimensional structure of protein and the suitable database to be used are normally considered. In comparison Perl is betterin I/O operations than python although python works better in string operations (Han et al. 2012).

These are Python libraries for DA: numpy is used for its N-dimensional array objects, pandas is a data analysis library that includes data frames, matplotlib is 2D plotting library for creating graphs and plots, scikit-learn the algorithms used for data analysis and data mining tasks, and seaborne is a data visualization library based on matplotlib (Dua & Chowriappa 2012)

### 3.5 Google Cloud Platform - Cloud Life Sciences

Annually, life sciences are generating a tremendous amount of data on patients, research, and laboratory practices. In practice, it is estimated that healthcare data volumes is increasing by 48% percent annually. Among this data, is raw and unprocessed data and a smaller portion of it is electronic health data that produce a bigger opportunism for research and reused by other institution. Google life sciences, therefore, provides a platform for parties involved to speed up the discovery of new research methods and more efficient processing of new treatment medicine. This section discusses in details the three main dimensions of cloud life sciences (Google Cloud 2019).

Firstly, Life Sciences from genomics.  Life Sciences enhance the community of life sciences in processing biomedical data in large quantities. It is mainly considered to be effective in terms of cost and gains supported from a favorable ecosystem of growing partners. Cloud Life Sciences enables you to focus on data analysis and results reproduction while the Google Cloud takes care of the rest of the task. Institutions undertaking academic researches about life sciences also utilize google cloud. The rationale behind this is that bioinformatics professionals, create not just what you need but also what you want, using open standards. Researchers fasten the research process, develop new questions, and transfer data in a safe and secure online environment. Professionals in the IT field rest easy considering that you have the resources you need to meet the demand for computation, secure data, and ensure the systems are reliable  (Google Cloud 2019).

Secondly, Workflow engines support and data transfer. Most of the renowned workflow engines are run on Google cloud for continuity in the use of the renowned tools, including Cromwell, NetFlow, and Galaxy. Google Cloud is the central control for several life science platforms such that one can focus on the work and leave the rest to the experts, including platform service, and the support partners.Data transfer within and without medical organizations is valuable to the life sciences community. In such a case, one can better manage access and usage of data by hosting it in a storage bucket within which operations, network, and retrieval costs can be easily billed to your clients (Google Cloud 2019).

In addition to Workflow engines support, security of information and compliance give reliable information security structured to attain or go past the requirements of the Health Insurance Profitability and accountability Act (HIPAA) and protected health information (PHI). They are Covered by the Business Associates from the HIPAA Agreement and available for the National Cancer Institute via Cancer Cloud

Pilots. Table 1 illustrates the restrictions and roles played by different individuals or groups on accessing the databases and their respective permissions (Chen et al. 2018).

| Role | Permission |
|---|---|
| role/lifesciences | • lifesciences.workflows.run<br>• lifesciences.workflows.cancel<br>• lifesciences.workflows.get<br>• lifesciences.workflows.list |
| roles/lifesciences.viewer | • lifesciences.workflows.get<br>• lifesciences.workflows.list |
| Workflows role | Permission |
| roles/lifesciences.workflowsRunner | • lifesciences.workflows.run<br>• lifesciences.operations.cancel<br>• lifesciences.operations.get<br>• lifesciences.operations.list |

TABLE 1. Cloud life access control (Adapted from Google Life Science)

### 3.6 Biological Database and Integration

A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of information. A few popular databases are Gen Bank from NCBI (National Center for Biotechnology Information), SwissProt from the Swiss Institute of Bioinformatics and PIR from the Protein Information Resource (Barrett & Salzman 2016)

Gen Bank (Genetic Sequence Databank) is one of the fastest growing repositories of known genetic sequences. EMBL, The EMBL Nucleotide Sequence Database is a comprehensive database of DNA and RNA sequences collected from the scientific literature and patent applications and directly submitted from researchers and sequencing groups. SwissProt is a protein sequence database that provides a high level of integration with other databases and also has a very low level of redundancy (Sutton & Austin 2015).

## 3.7 Common Example

This subsection discusses the five most common biological medicines that rely on bioinformatics and data mining techniques. These include Genomic, Proteomics, Micro-Array, Phylogenetic and population genetics.

### 3.7.1 Genomics

Genomics is a branch in biology that deals with mapping, structure, evolution and function of genomes which is normally done using genome referencing. It involves data types like genomic regulations, variant interpretation, splice-site prediction, DNA. RNA, methylation, chromatin accessibility, histone modifications, and chromosome interactions (Sindhu & Sindhu 2017).

Data mining has helped in coming up with models that help researchers in predicting regulatory elements and non-coding variant effects from a DNA/RNA sequence which can be tested for their contribution to gene regulation and observable traits (Sindhu & Sindhu 2017). Data mining has also helped researchers to be even clearer and provide accurate predictions of genomics features and functions.

### 3.7.2 Proteomics

Proteomics is the scientific study of proteomics. Proteome is a complement that can be expressed by a cell, tissue or organism. It’s mostly applied in checking protein function, protein modification, protein localization, protein-protein modification. It is normally done by checking the functionally modified protein (Sutton & Austin 2015, 126) Data mining in proteomics involves the use of Pyteomics package which has the ability to calculate basic physio-chemical properties of polypeptides, mass and isotopic, the Pi and charge, chromatographic retention, access to proteomics data, FASTA databases, search engines output, easy manipulation of consequences of modified peptides and proteins. Some of the commonly used libraries in data mining are Numpy and Matplotlib

Machine learning in proteomics helps in identifying the main proteins found in a certain sample e.g. comparing a tissue with a disease and others without disease. It's also used to check rates of protein production proteins modification, proteins interaction, proteins movements in cells and proteins expression. Data mining has also helped in identifying the functions of newly discovered genes showing where drugs bind with proteins and how they can interact with each other (Aguiar-Pulido et al. 2016, 117-118)

### 3.7.3 Micro-Arrays

Micro-Arrays are mostly used for automatically collecting datasets about large amounts of biological materials and evidences. This involves testing of a large set of genetic samples which help in identifying families of genes important in processes, this also helps in identification of mutations and single nucleotide polymorphisms (SNP), classification of tumors and drug discovery. Data mining has also led to developing of new molecular taxonomy of cancer, help in seeing action of drugs and assessing its sensitivity and toxicity, used in early detection of oral precancerous lesions. It's also used in gene hybridization where it determines loss presence of gene loss (Dedić & Stanier 2016).

### 3.7.4 Phylogenetic

It deals with the comparison of relationships between groups of organisms using phylogenetic trees. It is used to study the evolutionary relationship among organisms. Data mining in phylogenetic is mostly done using Dendpro, Bython and Bio. Phylo packages which have Numpy, pandas, matplotlib, PygraphviZ and Pydot libraries preinstalled. Numpy, pandas libraries are used for extraction of datasets for data analysis. Matplotlib, Pygraphviz and Pydot libraries are mostly used in constructing and drawing of phylogenetic trees (they are the ones used for visualizations) (Masood & Khan 2015).

Data mining in phylogenetic has helped the researchers in understanding how species, genes and genomes undergo evolution. It also used in courts to solve cases where DNA evidence is required. It helps also in learning more about new pathogen attack and in classification of new species (Rogalewicz & Sika, 2016, 104)

### 3.7.5 Population Genetics

Population genetics is basically studying the changes in the frequency of genes in population over a certain period of time, space selection and migration. Its' mostly based on data sequences. Its data mining mostly rely on analyzing SNP data e.g fixation index, pop structure and principal component analysis. It involves analysis of different datasets so as to make the analysis easier. Population genetics helps researchers in finding how genetic differences respond to different medications (Nicholson, 2006).

## 3.8 Challenges Facing Data Mining in Bioinformatics

There are many challenges hampering data mining in bioinformatics. Some of these challenges include the large size of the biological databases, heterogeneous biological data, and diversity of the databases with no standard ontology that can be essential in the querying of the data. Most of the biological data in the databases end up being created in hierarchical nature. (Wang Li & Perrizo 2015, 64-66; Tetko et al., 2016, 618). These factors, therefore, contribute to large inconsistencies in the biological data with greater impacts.

Firstly, data quality problems consist of two categories, single-source, and multisource problems. These two categories are further divided into schema and instance-related problems. Instance-related problems refer to errors and inconsistencies in the actual data contents that are not visible at the schema level and their primary focus is data cleaning. Schema-level problems are found at the schema level, and they can be addressed by incorporating changes into the schema design, i.e., evolving the schema by performing schema translation and schema integration operations (Paige et al., 2016, 10). The data quality problems of violations are related to the design of the schema. There are data quality issues when we have poor schema design and lack of adoption to proper data constraints (Chu et al., 2016, 3).

Secondly, integrity constraints, they are applied to the biological data during the design of the schema and the field of the databases and their preferred data types are controlled by the integrity constraints.

Biological databases are designed on the file systems with the relational schemes that are loosely defined. Therefore, the databases have limited restriction on the type of data which can be entered and stored the limited level of restrictions results in inherent inconsistencies that contribute to errors (Chu et al., 2016, 6).

Thirdly, biological databases utilize data from more than two sources. The inherent problem of cleaning data for a single data source is magnified when using multiple databases. In the multiple integrations of data, the problems of each independent source are combined. Data from different sources can contradict, overlap, and represented differently, heterogeneous resulting in complexities in designing schemes and data mining (Chen et al. 2018).

Fourthly, Unbalanced datasets. Unbalanced datasets are aspects which contributes to the challenges are the unbalanced datasets found in the field of bioinformatics. In the detection of the splice sites in bioinformatics the negatives sample are 100 times more than the positive samples. The costs of the classes in the SVMs is calculated through the association of the soft margin constant to the individual or each class about the samples in the class (Dua & Chowriappa 2012, 295).

Last but not least, over fitting problem is the second challenge in supervised learning in bioinformatics. In supervised learning, a model is regarded over fit when it's closely tied to the train set and the noise in the data (Dua & Chowriappa 2012 295). The results of the model which is tied close to its train are often regarded as biased and such models cannot perform tests on the selected sample or sample test.

Lastly, Computational challenges. Computational challenges arise during gene expression analysis. For instance, when using the Bayesian networks, there are possibilities of treating each of the genes found in the microarray as variables. Additionally, other challenges might arise from the other attributes such as the random variables. The random variables include the experimental conditions, temporal indicators and exogenous conditions. A Bayesian network enables one to arise a wide array of queries, whether there is dependence between the experimental conditions under study and the levels of expression of the gene. (Finotello, & Di Camillo 2015 131-134.)

However, these inferences are connected to the statistical constraints and interpretation of results obtained. When a complex system of genes is modeled, it entails processing complexity and a degree of algorithms. Dimensionality is the greatest challenge in the modeling process that is brought about by the few samples of analysis and thousands of genes. (Barrett & Salzman 2016).

## 3.9 The Future of Bioinformatics and Data Analysis

The application of knowledge discovery and data mining in bioinformatics is a growing research field. Therefore, it is of significance to explore research issues, which are significant in bioinformatics and design those data mining techniques, which are scalable, comprehensive and effective in their analysis. The data mining techniques must provide a room for an upgrade in the models to handle growing and dynamic biological data (García, Luengo& Herrera 2015, 19).

The data transformation and cleaning technique should be fully utilized to solve data problems associated with the biological databases (Malik et al., 2016, 12738). For instance, the data transformation techniques are based on the solving of the problems associated with the dynamic data types and summarization of the schemes, therefore, influencing and enhancing the representation of data and the structure of the designs. Additionally, the data transformation technique aid in the mapping of data from their formats and transforms them into formats as expected by the applications. Data transformations are effective and essential in handling evolving databases

# 4 CONCLUSION

In conclusion, industries such as health care, business, and research experience transformations due to growing interdisciplinary fields such as knowledge discovery, data mining and bioinformatics. Bioinformatics comprises multisource databases, rich in data with different formats. Therefore, data mining techniques have shown to be useful in the field of bioinformatics since techniques are suited for classifying, clustering, visualization, and prediction of the bioinformatics data to aid in drug discovery, cancer treatment, gene expression, gene translation, disease identification, protein function prediction, and microarray analysis.

Clustering techniques such as hierarchical, distance-based, fuzzy and graph-based are used in bioinformatics to enhance gene expression analysis. Hierarchical clustering is applied in gene clustering to distinguish overlapping and irregular genetic data. Distance metrics is used in incorporating known gene functions and establishes whether genes can share common gene functions. Fuzzy clustering allows the classification or placing of the genes in more than one cluster. Genes are placed in different clusters. Thus, allowing the bioinformatics to identify those genes, which are conditionally co-expressed or co-regulated cluster (Manikandan et al., 2018, 1814). Graph-based clustering is used in gene expression through the application of shortest path technique in analyzing gene expression data and using a minimum spanning tree of a graph to construct the genetic linkage maps classification technique is applied in proteomics and genomics. The aim of the bioinformatics in proteomics is to classify the protein sequence into structural and functional families according to homology sequence (Dua & Chowriappa, 2012, 219).

## REFERENCES

Aguiar-Pulido 2016. Metagenomics, metssatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. Evolutionary Bioinformatics, 12, EBO-S36436. And criminal identification in India using data mining techniques. AI & society, 30(1), 117-127.

Barrett, S. P., & Salzman, J. 2016. Circular RNAs: analysis, expression and potential functions. Development, 143(11), 1838-1847.

Chen, Y., Tang, S., Bouguila, N., Wang, C., Du, J., & Li, H. 2018. A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data. Pattern Recognition, 83, 375-387.

D'Souza, A. R., & Minczuk, M. 2018. Mitochondrial transcription and translation: overview. Es-says in biochemistry, 62(3), 309-320.

Dedić, N., & Stanier, C. 2016, November). Towards differentiating business intelligence, big data, data analytics and knowledge discovery. In International Conference on Enterprise Resource Planning Systems (pp. 114-122). Springer, Cham.

Dua, S., & Chowriappa, P. 2012. Data mining for bioinformatics. CRC Press.

García, S., Luengo, J., & Herrera, F. 2015. Data preprocessing in data mining (pp. 59-139). New York: Springer.

Google Cloud life Sciences, 2019. Google Cloud. Available: https://cloud.google.com/solutions/life-sciences/, Accessed 15 December 2019.

Han J, Kamber M, Pei J 2012 Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann Publishers Inc, San Francisco, CAMasood, M. A., & Khan, M. N. A. (2015). Clustering techniques in bioinformatics. IJ Modern Edu-cation and Computer Science, 1, 38-46.

Katragadda, S., Gottumukkala, R., Venna, S., Lipari, N., Gaikwad, S., Pusala, M. & Bayoumi, M. 2019. VAStream: A Visual Analytics System for Fast Data Streams. In Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) (p. 76). ACM.

Kumar, A., & Chatterjee, I. 2016. Data Mining: An experimental approach with WEKA on UCI Dataset. International journal of computer applications, 138(13).

Martinez, W. L., Martinez, A. R., & Solka, J. 2017. Exploratory data analysis with MATLAB. Chapman and Hall/CRC.

Nicholson, S. 2006. The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. Information processing & management, 42(3), 785-804.

Petermann, A., Junghanns, M., Kemper, S., Gómez, K., Teichmann, N., & Rahm, E. 2016. Graph mining for complex data analytics. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (pp. 1316-1319). IEEE.

Rogalewicz, M., & Sika, R. 2016. Methodologies of knowledge discovery from data and data min-ing methods in mechanical engineering. Management and Production Engineering Review, 7(4), 97-108.

Sindhu, S., & Sindhu, D. 2017. Data Mining and Gene Expression Analysis in Bioinformatics. In-ternational Journal of Computer Science and Mobile Computing, 6(5), 72-83.

Sutton, J., & Austin, Z. 2015. Qualitative research: Data collection, analysis, and management. *The Canadian journal of hospital pharmacy*, *68*(3), 226.