Scientific Research Publishing

Int. J. Communications, Network and System Sciences, 2022, 15, 149-165
https://www.scirp.org/journal/ijcns
ISSN Online: 1913-3723
ISSN Print: 1913-3715

Data Mining as a Technique for Healthcare Approach

E. N. Ekwonwune[1], C. I. Ubochi[1], A. E. Duroha[2]

[1]Department of Computer Science, Imo State University, Owerri, Nigeria
[2]Department of Computer Science, Gregory University, Uturu, Nigeria
Email: ekwonwuneemmanuel@yahoo.com

How to cite this paper: Ekwonwune, E.N., Ubochi, C.I. and Duroha, A.E. (2022) Data Mining as a Technique for Healthcare Approach. *Int. J. Communications, Network and System Sciences*, **15**, 149-165.
https://doi.org/10.4236/ijcns.2022.159011

Received: August 17, 2022
Accepted: September 27, 2022
Published: September 30, 2022

Abstract

Data Mining, also known as knowledge discovery in data (KDC), is the process of uncovering patterns and other valuable information from large data sets. According to https://www.geeksforgeeks.org/data-mining/, it can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. With advance research in health sector, there is multitude of Data available in healthcare sector. The general problem then becomes how to use the existing information in a more useful targeted way. Data Mining therefore is the best available technique. The objective of this paper is to review and analyse some of the different Data Mining Techniques such as Application, Classification, Clustering, Regression, etc. applied in the Domain of Healthcare.

Keywords

Data Mining, Techniques, Relational Database, Knowledge, Clustering, Classification, Regression, Healthcare

1. Introduction

Nowadays, there is huge amount of data being collected and stored in databases everywhere across the globe running into terabytes of data and the tendency is to keep increasing year after year. Today, the healthcare industry which is one of the largest industries throughout the world includes medical industries having the large amounts of health-related and medical-related data. It also includes thousands of hospitals, clinics and other types of facilities that provide primary, secondary and tertiary levels of care. The delivery of healthcare service is the most visible part of any healthcare system, both to users and the general public. Accurate, early and error-free diagnosis and treatment given to patients has been a major issue highlighted in medical service nowadays. Quality service in health care field implies diagnosing patients correctly and administering treatments that are effective. To achieve a correct and cost-effective treatment, a system can be developed to fulfill the task.

Today, healthcare systems use hospital information systems to manage their healthcare or patient data. These systems generate huge amounts of medical data which may contain electronic patient records with their computer-readable entries like Magnetic Resonance Imaging (MRIs), signals like Electrocardiography (ECG), clinical information like blood sugar, blood pressure, cholesterol levels, etc. as well as the physician's interpretation. These data may comprise a lot of invaluable information and knowledge hidden in such databases, which can be used to support effective clinical decision-making. But the discovery of hidden patterns and knowledge from such databases often goes unexploited. This explosive growth of data requires an automated way to extract what is called "nuggets of knowledge" from the large sets of data for the diagnosis and treatment of disease from the database and this increasingly becomes necessary. Conventionally, the data is analyzed manually using retrospective tools typical of decision support systems and many hidden and potentially useful relationships may not be recognized by the analyst because it lies outside their expectations. The research purpose is to find alternatives to the solution of complex medical diagnosis in detection or prediction of heart disease where human knowledge is apprehended in a general fashion. Successful application examples shown previously, reveal that human diagnostic capabilities are significantly worse than computer-aided diagnostic systems.

Advanced data mining techniques in medicine can deal with this problem and remedy the situation. Data mining can improve the management level of hospital information, saving time and cost. Using data mining techniques, we can extract interesting knowledge and regularities. The discovered knowledge can then be applied to the medical data to increase the working efficiency and help medical practitioners in their decision-making. This provides near-endless opportunities for symptom trend detection, earlier detection of illness, and DNA trend analysis, etc. thus increasing efficiency in healthcare.

While developed countries like the United States have their systems automated already and Intelligent Systems are increasingly being deployed in medicine and healthcare, to practically aid the busy clinician and to improve the quality of patient care, developing countries like Nigeria still make use of traditional statistical methods., But with the help of data mining, we can also join the moving trend.

According to [1], Data Mining consists of five major elements viz:

- Extract, transform and load data onto data warehouse system.
- Store and manage the data in multidimensional database.
- Provide data access to analysts.
- Analyze the data by application software.

- Present the data in useful format.

Steps in a Data Mining Process are:

1) Data integration: first of all, data is collected from and integrated from all different sources.

2) Data selection: data is selected on behalf of some criteria.

3) Data preprocessing: data collected may contain errors, inconsistencies that need to be removed.

4) Data transformation: the data even after preprocessing may not be ready for mining so it needs to be transformed into a form appropriate for mining.

5) Knowledge discovery: meaningful patterns are extracted from large data. At last, meaningful patterns help in decision-making [1].

Data mining is vastly being used in solving numerous research problems. Data mining is becoming increasingly popular and essential in healthcare sector [2]. Data mining applications can provide advantage to all parties involved in the healthcare industry [3] [4]. For example, data mining can help healthcare insurer detect fraud and abuse, physicians identify effective treatments and best practices and patients receive better and more affordable healthcare services [5]. Huge amount of data generated by healthcare transactions are too complex and voluminous to be processed and analyze by traditional methods. Data mining provides the methodology and technology to transform these amounts of data into useful information for decision-making in healthcare sector.

Statement of Problem

Data Mining as an analytic process is designed to extract useful data, patterns and trends from a large amount of data (typically business, medical or market related data) by using techniques like clustering, classification, association and regression. The ultimate goal of data mining is prediction. Research statistics have shown that most healthcare-related diseases make use of data mining techniques that do not create optimal results.

Below are some of the various inconsistencies associated with the use of the wrong data mining techniques:

1) Not having very high accuracy in decision.

2) Shortage of expertise.

3) Difficulties in knowledge upgrade.

4) Time-dependent performance (very time-consuming).

Because of these problems, there is necessity to deploy data mining to provide the assistance mechanism in diagnosis procedure. The conclusion is clear: humans and their statistical methods cannot ad hoc analyze complex data without errors. In medicine and healthcare where safety is critical, it is important if data mining techniques are to be widely accepted in clinical practice [6].

The goal of the process is to take the medical data which contain many attributes and determine which ones are actually relevant to the diagnosis, symptoms and result of heart disease. Without automatic methods for extracting this information, it is practically impossible to mine for them, seeing that we are looking at a very huge amount of data running into terabytes of data.

2. Literature Review

2.1. Conceptual Framework

The healthcare industry battles with millions of digitally recorded data and patterns being collected at enormous speed due to the widespread usage of powerful computer devices nowadays [7]. The data collected are mostly unorganized and have not been used properly for appropriate applications, thus, imposing new challenges regarding their management including their modeling, storage, and retrieval capabilities. There is often interesting knowledge in the data that is not readily evident. The spread of electronic patient records, with their computer-readable entries e.g. Magnetic Resonance Imaging (MRI), signals like ECG (Electrocardiography), clinical information like blood sugar, blood pressure, cholesterol levels, etc. as well as the physician's interpretation is opening new possibilities for medical data mining and a world of virtual research [8].

Knowledge Discovery in Databases (KDD) and Data Mining (DM) provide a solution to the information flood problem by extracting valid, novel, potentially useful, and ultimately understandable patterns from data [9]. Patterns constitute compact and rich in semantics representations of raw data [10]; compact by means that they summarize, to some degree, the amount of information contained in the original raw data and rich in semantics by means that they reveal new knowledge hidden in the abundance of raw data.

Different data mining tasks achieve different insights over the data: classification captures the class of data or a new item, clusters reveal natural groups in data, decision trees detect characteristics that predict (with respect to a given class attribute) the behavior of future records, and so on [11]. This unorganized data requires processing to be done to generate meaningful and useful information from the large databases. In order to organize large amount of data, you implement the concept of Database Management Systems (DBMS) such as Oracle, and SQL Server. These Database Management Systems require you to use SQT, a specialized query language to retrieve data from a database. However, the use of SQT is not always adequate to meet the end user requirements of specialized and sophisticated information from an unorganized large data bank. Database researchers pay more attention to the issues related to the volume of data and also concerned with the effective use of the available database techniques such as efficient data retrieval mechanisms. This therefore necessitates you to look for certain alternative techniques to retrieve information from large and mostly unorganized sources of data.

Nowadays, data stored in medical databases are growing in an increasingly rapid way. Analyzing that data is crucial for medical decision-making and management [12]. It has been widely recognized that medical data analysis can lead to an enhancement of health care by improving the performance of patient management tasks. There are two main aspects that define the need for medical data analysis:

1) Support of specific knowledge-based problem solving activities through the analysis of patients' raw data collected in monitoring.

2) Discovery of new knowledge that can be extracted through the analysis of representative collections of example cases, described by symbolic or numeric descriptors. For these purposes, the increase in database size makes traditional manual data analysis to be insufficient. To fill this gap, new research fields such as knowledge discovery in databases (KDD) have rapidly grown in recent years. KDD is concerned with the efficient computer-aided acquisition of useful knowledge from large sets of data.

It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

Data Mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process [12]. The KDD process is often to be nontrivial; however, we take the larger view that KDD is an all-encompassing concept. KDD is a process that involves many different steps. The input to this process is the data, and the output is the useful information desired by the users. However, the objective may be unclear or inexact. The process itself is interactive and may require much elapsed time. To ensure the usefulness and accuracy of the results of the process, interaction throughout the process with both domain experts and technical experts might be needed.

Data mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships of interest in a particular representational form or a set of such representations as classification rules or trees, regression and clustering, to the interpretation/evaluation step of the KDD process. The definition clearly implies that what data mining (in this view) discovers are hypotheses about patterns and relationships. Those patterns and relationships are then subject to interpretation and evaluation before they can be called knowledge.

A simple data mining process model includes the following steps [13]:

1) Select a target data set.

2) Data preprocessing.

3) Data transformation.

4) Data mining.

5) Interpretation/evaluation.

6) Presentation.

7) Documentation: Simply the documentation and reporting it to interested parties are done at this last step.

Whereas in unsupervised learning no training set is used. Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [14]. Several data mining algorithms are used in IQ diagnosis of heart disease such as Naive Bayes, Decision Tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies. Alternative names to data mining are: knowledge discovery mining) in databases (KDD), knowledge extraction, data/patterns analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

2.2. Techniques of Data Mining

According to https://www.geeksforgeeks.org/data-mining/, Data Mining has been integrated with many other techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, etc. to gather more information about the data and to help predict hidden patterns, future trends, and behaviours and allows businesses to make decisions (see **Figure 1** below). Over the past two decades, it is clear that we have been able to develop systems that collect massive amounts of data in health care, but now what do we do with it?

Data mining methods use powerful computer software tools and large clinical databases, sometimes in the form of data repositories and data warehouses, to detect patterns in data. Within data mining methodologies, one may select from an extensive array of techniques or patterns types that include, among many others, classification, clustering, and association rules [12].

Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.

Data Mining can be applied to any type of data e.g. Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-series Databases, World Wide Web.

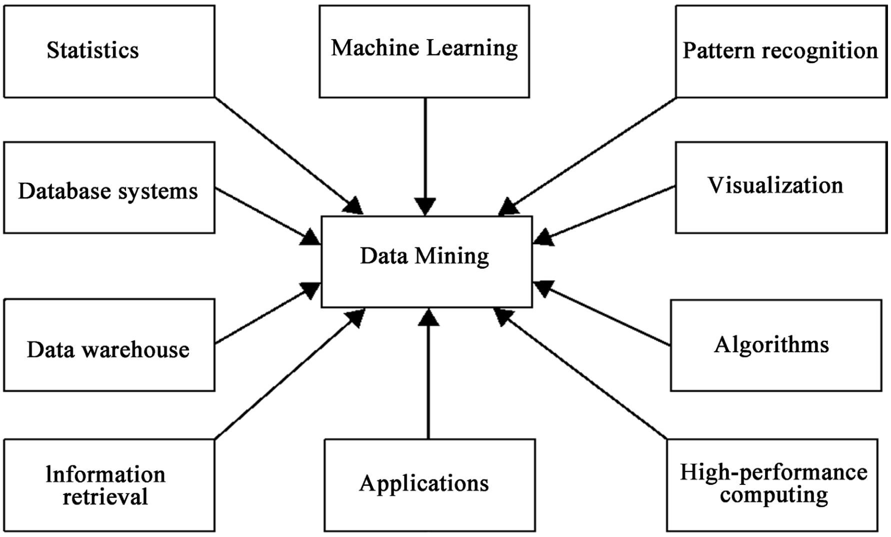

Figure 1. Data mining techniques. Source: https://www.geeksforgeeks.org/data-mining/.

The whole process of Data Mining consists of three main phases. This is shown in Figure 2 below:

1) Data Pre-processing—Data cleaning, integration, selection, and transformation take place.

2) Data Extraction—Occurrence of exact data mining.

3) Data Evaluation and Presentation—Analyzing and presenting results.

2.3. Factors Driving Data Mining

Data mining has come of age because of the confluence of three factors:

- The first is the ability to inexpensively capture, store and process tremendous amounts of data.
- The second is advances in database technology that allow the stored data to be organized and stored in ways that facilitate speedy answers to complex queries.
- Finally, there are developments and improvements in analysis methods that allow them to be effectively applied to these very large and complex databases [15].

2.3.1. Advantages of Data Mining in Medicine

According to [15], there are numerous advantages associated with data mining in medicine some of which are listed below:

- Earlier detection of illness.
- Symptom trends.
- Data analysis.
- Improved drug reactions.
- Provides new knowledge from existing data.
- Public databases.

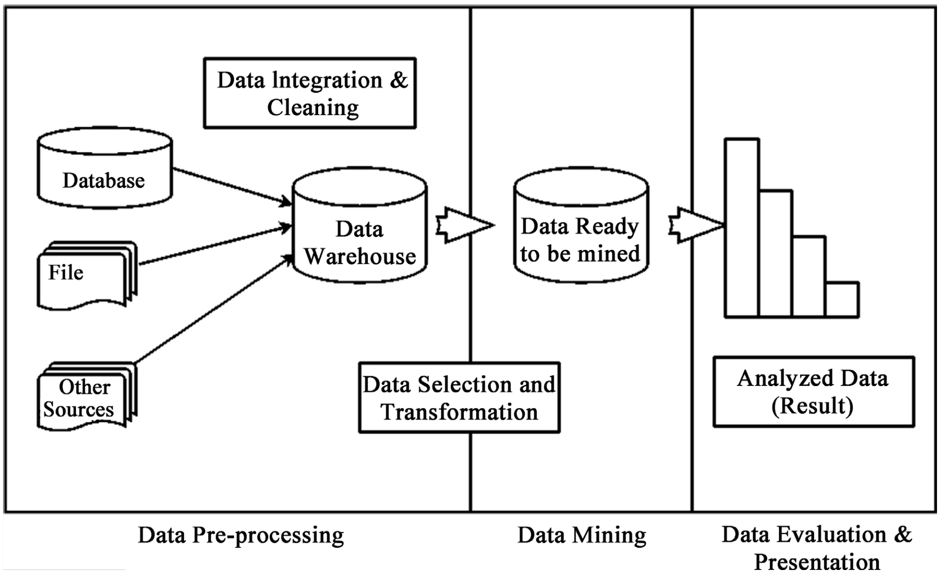

Figure 2. Data evaluation and presentation. Source: https://www.geeksforgeeks.org/data-mining/.

- Government sources.
- Company or healthcare Databases.
- Old data can be used to develop new knowledge.
- New knowledge can be used to improve services or products.
- Improvements lead to:

  Bigger profits.

  More efficient service.

2.3.2. Disadvantages of Data Mining in Medicine

- No uniform language-Medical.
- Incomplete records.
- Privacy, etc.

2.4. Methodological Studies

Data mining has played an important role in the intelligent medical systems [16]. The relationships of disorders and the real causes of the disorders and the effects of symptoms that are spontaneously seen in patients can be evaluated by the users via the constructed software easily. Large databases can be applied as the input data to the software by using the extendibility of the software. Knowledge of the risk factors associated with heart disease helps healthcare professionals to identify patients at high risk of having heart disease. Statistical analysis and data mining techniques help healthcare professionals in the diagnosis of heart disease. Throughout the years, many algorithms were created to extract what is called nuggets of knowledge from large sets of data. There are several different methodologies to approach this problem like Classification, Regression Trees, Decision Trees, Support Vector Machines, etc.

2.4.1. Classification

In their opinion, [1] stated that Classification is the process of predicting output based on some given input data. The goal of classification is to accurately predict the target class for each case in the data [17]. In order to predict the data, it processes the training set and predictive set. It first develops relationships between the attributes of training data set. Then it is provided with the predictive data set, which contains similar attributes but with different data values. Then it analyzes the given data and produces prediction by placing the different data sets in different classes based on the relationship of attributes [18] [19]. For example, in a medical database; the training set would have relevant patient information based on its previous records, whereas the prediction attribute is whether the patient has chances of heart attack as shown in **Table 1** and **Table 2**.

Classification uses predictive rules expressed in the form of IF-THEN rules where the first part (IF part) consists of conjunction of conditions and the second part (THEN part) predicts a certain prediction attribute value that satisfies the first part. Using the above example, a rule predicting the first row in the training set may be represented as follows: IF (age = 62 and heart rate > 72) or (age > 60 and

Table 1. Training set. Source: Kamna Solanki *et al.*, 2016.

AGE	HEART RATE	BLOOD PRESSURE	HEART PROBLEM
62	79	145/70	YES
35	82	115/75	YES
79	65	110/68	NO

Table 2. Predictive set. Source: Kamna Solanki *et al.*, 2016.

AGE	HEART RATE	BLOOD PRESSURE	HEART PROBLEM
45	96	143/69	?
63	54	108/73	?
83	95	115/68	?

blood pressure > 140/70) then Heart problem = yes. This technique provide 80% prediction rate, but the optimal solution is a rule with 100% prediction rate; which is very hard to achieve. Following are the classification techniques used in health care.

2.4.2. Decision Trees

Decision tree is similar to flow chart in which every non-leaf node denotes a test on a particular attribute and every branch represents an outcome of the test. Root node is the topmost node in the decision tree. For example, with the help of readmission tree, we can decide whether a patient needs to be readmitted or not. Using Decision Tree, a decision maker can choose the best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [20] [21]. Decision tree are self-explanatory and easy to follow. Set of rules can also be constructed with the help of decision tree. Decision Tree can be considered as nonparametric method because there is no need to make assumptions regarding distribution of space and structure of classifier. Decision tree have several disadvantages. These are: Most of the algorithm like ID# and C4.5 require target attributes to have discrete values as decision tree use divide and conquer strategy. More the complex relationship among attributes lesser is the performance.

2.4.3. Support Vector Machines (SVM)

Vladimir Vapnik first introduced idea of Support Vector Machine [22]. Its accuracy is better than all other available techniques. It was first introduced for binary classification problems; but it can be further extended to multi class problems. It creates hyper-planes to separate data points [23].

It can be implemented in 2 ways:

1) Mathematical programming.

2) Using kernel functions.

With the help of training data sets, non-linear functions can be easily mapped to high dimensional space. This can only be possible using kernel functions like Gaussian, sigmoid, etc.

2.4.4. Neural Network (NN)

It was developed in 20th century. Neural network was regarded as the best classification algorithm before the introduction of decision tree and SVM which has far better results. This was the reason that encouraged NN as the most widely used classification algorithm in various bio-medicine and healthcare fields. For example, NN has been used as the algorithm supporting the diagnosis of diseases like cancer and predicting outcomes. In NN, basic elements are nodes or neurons. These neurons are interconnected and within the network they work together to produce the output functions. They are fault tolerant as they are capable of producing new observations from the existing observations in those situations where some neurons within the network fail. An activation number is associated with each neuron and a weight is assigned to each edge within the NN. The basic property of NN is that it can minimize the error by adjusting its weights and by making changes in its structure as it is adaptive in its nature. One major advantage of NN is that it can properly handle noisy data for training and can reasonably classify the new type of data which is different from training data. There are also various disadvantages of NN. First, it requires many parameters including the optimum no of hidden layer nodes that are empirically determined and its classification performance is very sensitive to the parameters selected. Second, its training or learning process is very slow and expensive.

Table 3 depicts the usage of Classification techniques in healthcare sector.

2.4.5. Regression

Regression is a data mining technique that helps in identifying those functions that are useful in order to demonstrate the correlation among different variables. It is a mathematical tool and can be easily constructed using training data sets. Regression can be classified into linear and non-linear based on certain count of

Table 3. Usage history of classification techniques in HealthCare Sector. Source: Kamna Solanki *et al.*

Researcher	Technique used	Purpose
1. Hu *et al.* [12]	SVM, decision tree, bagging and boosting.	To analyze micro array data.
2. Huang *et al.* [13]	Hybrid SVM based diagnosis model	For breast cancer.
3. Khan *et al.* [14]	Decision tree	For breast cancer.
4. Chang *et al.* [15]	Integrated Decision tree model.	For skin diseases in adults and children.
5. Curiac *et al.* [16]	Bayesian method	For psychiatric disease.
6. Moon *et al.* [17]	Decision tree algorithm	To characterize the smoking behaviour among smokers by assessing their psychological health conditions and consumption of alcohol.
7. Chien *et al.* [18]	Hybrid decision tree classifier.	For chronic disease.
8. Shouman *et al.* [19]	K-NN classifier.	For heart disease.
9. Liu *et al.* [20]	Fuzzy-NN classifier.	For thyroid disease.
10. Er *et al.* [21]	Artificial Neural network	For chest disease.

independent variables. In order to estimate association between two types of variable in which one is dependent variable and another one is independent variable, linear regression is used. One of the disadvantages of this technique is that it cannot be used for categorized data. The categorical data can be used with the help of logistic regression. Usage of Regression for health sector has been summarized in Table 4.

2.4.6. Clustering

It is an unsupervised learning technique which is different from classification technique (supervised learning method). It is best suited for large amount of data. It works by observing independent variables. The main task is to form clusters from large databases on the basis of similarity measure. Different types of clustering algorithms are defined in Table 5 and various clustering algorithms used in health care are described in Table 6.

2.5. Applications of Data Mining

According to https://www.geeksforgeeks.org/data-mining/, Data Mining can be

Table 4. Usage history of regression techniques in HealthCare sector. Source: Kamna Solanki *et al.*

Researcher	Technique used	Purpose
1. Divya *et al.* [22]	Weighted SV Regression	To provide better healthcare services by continuously monitoring patients.
2. Xie *et al.* [23]	Regression decision tree algorithm	To study number of hospitalization days.
3. Alapont *et al.* [24]	Linear regression	For effective utilization of hospital resources.

Table 5. Types of clustering algorithms in Healthcare sector. Source: Source: Kamna Solanki *et al.*

Technique	Description
1. Partitioned Clustering	With the help of "n" data points maximum possible of "k" clusters is obtained by relocating objects to "k" clusters.
2. Hierarchical Clustering	Data points are partitioned in tree form either top-down or bottom-up.
3. Density-based Clustering	It can handle cluster of any arbitrary shape whereas above two can handle only spherical shape clusters.

Table 6. Usage history of clustering techniques in Healthcare sector. Source: Kamna Solanki *et al.*

Researcher	Technique used	Purpose
1. Chen *et al.* [25]	Hierarchical clustering	To analyze micro-array data.
2. Chipman *et al.* [26]	Hybrid Hierarchical clustering.	To analyze micro-array data.
3. Bertsimas *et al.* [27]	Clustering algorithm	To predict health care cost.
4. Peng Y *et al.* [28]	Clustering algorithm	To detect healthcare frauds.
5. Belciug *et al.* [29]	Hierarchical, partitioned and density-based clustering.	Efficient utilization of healthcare resources.

applied in the following Areas as shown in Figure 3 below:

1. Financial Analysis
2. Biological Analysis
3. Scientific Analysis
4. Intrusion Detection
5. Fraud Detection
6. Research Analysis

Applications of Data Mining in Healthcare Domain

Applications of data mining in healthcare have already led to some measurable improvements in patient care [24]. New York-Presbyterian Hospital has reduced the rate of potentially fatal blood clotting in patients by using analytics software, Dr Nicholas Morrissey, a surgeon involved in the effort, said on an interview. As patients were being admitted, the hospital started using Microsoft software in 2010 to scan records for risk factors, such as cancer, smoking and the amount of time patients were bed-bound. By letting the software rather than the hospital staffs make the assessment, doctors saved time and made better evaluations, Morrissey said. The clotting now happens at a rate of 0.23 incidents per 1000 patient days, as against 0.33 incidents per 1000 patient days before implementation of the software, Morrissey said. "I won't be out there saying that we've solved the problem, but we're definitely making progress-that was a significant drop", he said.

In the opinion of [25], the successful application of data mining in various domains such as marketing, retail, engineering or banking has led to the expansion of its horizon to new fields, namely medicine and public health. Nowadays, an increasing number of data mining applications focus on analyzing health care centers for a better health management, or to detect disease outbreaks in hospitals, to prevent patients' deaths, and, obviously, to detect fraudulent insurance claims [21] [26] [27] [28].

Data mining provides the methodology and technology to discover knowledge from the huge amount of data and furthermore, this knowledge is used for decision making. In [27] the author discusses the capability of data mining to improve

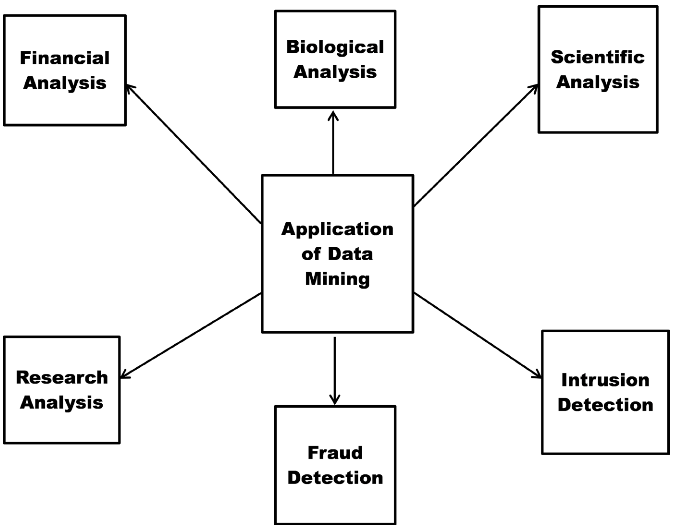

Figure 3. Application of data mining. Source: https://www.geeksforgeeks.org/data-mining/.

the quality of the decision-making process in pharma industry. One of the major problems with pharmaceutical data is actually the lack of information. Predicting drug behaviour is essential to find out if the treatment helps the patients or their health status gets worse. Data mining can help experts in healthcare management [29] to make decisions in the sector of customer relationship management. Patients will receive better and more affordable healthcare services if large amount of data about the degree of other patients' satisfaction regarding medical sector will be analyzed and adequately interpreted. Biological databases may be considered the raw material for multi-relational data mining techniques [30], due to their wide variety of data types, often with complex relational structure.

At the University of Alabama [31], there was implemented a surveillance system that uses data mining techniques (association rules) in order to identify new and interesting patterns in the infection control data. Data collected over one year (1996) were analyzed and three separate analyses were conducted, each one using a different size of data partition.

In the research [32], it is presented the case study of American Healthways which provides diabetes management services to hospitals and health plans so that to enhance the quality and lower the cost of treating patients with diabetes.

The authors of the present article focus their research on applying data mining techniques in order to classify patients with thyroid disorders. In the literature existing on the diagnosis of thyroid diseases, the authors have identified the following data mining algorithms: decision trees, artificial neural networks, support vector machine, expert systems, etc. For example, the diagnosis of thyroid disorders by using ANN's is discussed in [33] [34] [35]. In [33], authors used data related to UCI site, collected in 1992 by James Cook University, Townsville of Australia. The total number of laboratory samples was 215. Data mining algorithm used five attributes as predictors and one attribute as a target. By selecting a hidden layer, the Logsig activation function for the hidden layer and 6 neurons from this layer, the level of classification accuracy was 98.6% in case of thyroid disease. The software used for testing the model was MATLAB 2012. In [34], authors present their work with respect to three ANN algorithms for the diagnosis of thyroid disease: the Back propagation algorithm (BPA), the radial basis function (RBF) Networks and the learning vector quantization (LVQ) networks. After the model evaluation, LVQ network had the best accuracy rate, i.e. 98%.

The classification of thyroid nodules was performed with support vector machines in [36]. In [37] there is presented a comparison study on data mining classification algorithms (C 4.5, C5.0) for the thyroid cancer. The authors of [37] used a database with 400 records extracted from the UCI thyroid database and 29 attributes. The study indicated that the confidence level for the rule set generated by C5.0 was higher than 95%. In [37], C4.5 approach was implemented in java platform by using Eclipse and XP operating system. A diagnosis expert system based on fuzzy rules is described in [38], while a three-stage expert system based on support vector machines is discussed in [39]. The system proposed in [39] reached the highest accuracy reported so far in the classification of thyroid disorders, by using a 10-fold cross-validation method, with the mean accuracy of 97.49% and with the maximum accuracy of 98.59%.

In the following section of the paper, the authors shall present a case study on the classification models applied to a database containing records about individuals with thyroid diseases. The data set consisted in 756 records extracted from UCI Machine Learning Repository [40]. We used 21 attributes as predictors and a class attribute. In the experiments described below, there were used CART and TreeNet models. By comparing the obtained results with those existing in the above-mentioned studies, in most of our experiments, the accuracy of CART model was over 93%, the highest value being 96.86%, for the following settings: Priors = Equal, Costs = 0.5, Parent node min cases = 10, Terminal node min cases = 1, Partition = 0.6. The accuracy of TreeNet model was 94.97%.

3. Summary and Conclusion

This research work discussed vital issues bothering on Data Mining as a Technique applied in Healthcare Industry. Some of these concepts discussed as Techniques include but are not limited to the following: Classification, Regression, Decision Trees, Support Vector Machines, etc. This work also concludes that data mining is of great importance in the solution of healthcare problems. Data mining, however, is not a “silver bullet” capable of solving all of Cardiovascular Diseases, but rather it aims at providing possible prevention methods, remedies and symptoms.

Some diseases pose a serious life-threatening risk globally across all races and age groups. But most of them could be controlled and risk reduced through a proper medical checkup, proper dieting, nutrition, healthy eating habits and exercising all of which are incorporated into our data mining software for easy access.

This work also highlighted that the main goal of achieving high accuracy and efficiency which is very important in healthcare sector still remains an open research issue.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Kamma, S., Parul, B., Sandeep, D. and Sudluv (2016) Analysis of Application of Data Mining Techniques in Healthcare. *International Journal of Computer Applications*, **148**, 16-21. https://doi.org/10.5120/ijca2016911011

[2] Koh, C.H. and Tan, G. (2011) Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, **19**, 64-72.

[3] Kaur, H. and Wasan, S.K. (2006) Empirical Study on Applications of Data Mining

Techniques in Healthcare. *Journal of Computer Science*, **2**, 194-200. https://doi.org/10.3844/jcssp.2006.194.200

[4] Obenshain, M.K. (2004) Application of Data Mining Techniques to Healthcare Data. *Infection Control & Hospital Epidemiology*, **25**, 690-695. https://doi.org/10.1086/502460

[5] Liao, S.H., Chu, P.-H. and Hsiao, P.-Y. (2012) Data Mining Techniques and Applications—A Decade Review from 2000 to 2011. *Expert Systems with Applications*, **39**, 11303-11311. https://doi.org/10.1016/j.eswa.2012.02.063

[6] Ubochi, C.I. (2017) Data Mining Technique for Detecting Cardiovascular Diseases. An Unpublished Msc Thesis.

[7] Savage, N. (2011) Mining Data for Better Medicine. *MIT Technology Review*, **38**, 235-237.

[8] Joshi, S., Deepashenoy, P., Venugopal, K.R. and Patnaik, L.M. (2010) Data Analysis and Classification of Various Stages of Dementia Adopting Rough Sets Theory. *International Journal on Information Processing*, **4**, 86-89.

[9] Fayyad, U.M., *et al.* (1996) From Data Mining to Knowledge Discovery in Databases. AAAI Press/The MIT Press, Cambridge.

[10] Rizzi, S., Bertino, E., Catania, B., Golfarelli, M., Halkidi, M., Terrovitis, M., Vassiliadis, P., Vazirgiannis, M. and Vrachnos, E. (2003) Towards a Logical Model for Patterns in ER. Springer, Chicago, Vol. 2813, 77-90. https://doi.org/10.1007/978-3-540-39648-2_9

[11] Ntoutsi, I. (2008) Similarity Issues in Data Mining—Methodologies and Techniques. University of Piraeus, Piraeus, 31-32.

[12] Li, J.-S., Yu, H.-Y. and Zhang, X.-G. (2011) Data Mining in Hospital Information System. In: Funatsu, K., Ed., *New Fundamental Technologies in Data Mining*, InTech, Shanghai, Vol. 1, 143-156.

[13] Balasunda, V., Devi, T. and Saravanan, N. (2012) Development of a Data Clustering Algorithm for Predicting Heart. *International Journals of Computer Applications*, **48**, 8-13. https://doi.org/10.5120/7358-0095

[14] Han, J. and Kamber, M. (2006) Data Mining Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers Inc., Burlington, 55-86.

[15] Chauraisa, V. and Pal, S. (2013) Data Mining Approach to Detect Heart Diseases. *International Journal of Advanced Computer Science and Information Technology*, **2**, 56-66.

[16] Aflori, C. and Craus, M. (2007) Grid Implementation of the Apriori Algorithm. *Advances in Engineering Software*, **38**, 295-300. https://doi.org/10.1016/j.advengsoft.2006.08.011

[17] Deulkar, D.S. and Deshmukh, R.R. (2016) Data Mining Classification. *Imperial Journal of Interdisciplinary Research*, **2**.

[18] Palaniappan, S. and Awang, R. (2008) Intelligent Heart Disease Prediction System Using Data Mining Techniques. *IEEE Conference on Computer Systems and Applications*, Doha, 31 March 2008-4 April 2008, 108-115. https://doi.org/10.1109/AICCSA.2008.4493524

[19] Srinivas, K.B., Rani, K. and Govrdhan, A. (2010) Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering*, **2**, 250-255.

[20] Ahmed, P., Qamar, S. and Rizvi, S.Q.A. (2015) Techniques of Data Mining in Healthcare: A Review. *International Journal of Computer Applications*, **120**, 38-50.

https://doi.org/10.5120/21307-4126

[21] Durairaj, M. and Ranjani, V. (2013) Data Mining Applications in Healthcare Sector: A Study. *International Journal of Scientific and Technology Research*, **2**, 29-35.

[22] Vapnik, V. (1998) The Support Vector Method of Function Estimation. In: Suykens, J.A.K. and Vandewalle, J., Eds., *Nonlinear Modeling*, Springer, Berlin, 55-85. https://doi.org/10.1007/978-1-4615-5703-6_3

[23] Burges, C.J.C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2**, 121-167. https://doi.org/10.1023/A:1009715923555

[24] Robertson, J. (2012) Data Mining in Doctor's Office Helps Solve Medical Mysteries. Vol. 1, Wal-Mart or Western Union United Healthcare Corp., New York.

[25] Ionita, I. and Ionita, L. (2016) Applying Data Mining Techniques in Healthcare. *Studies in Informatics and Control*, **25**, 385-394. https://doi.org/10.24846/v25i3y201612

[26] Canlas Jr., R.D. (2015) Data Mining in Healthcare: Current Applications and Issues.

[27] Ranjan, J. (2007) Application of Data Mining Techniques in Pharmaceutical Industry. *Journal of Theoretical and Applied Information Technology*, **3**, 61-67.

[28] Diwani, S., Mishol, S., Kayange, D.S., Machuve, D. and Sam, A. (2013) Overview Applications of Data Mining in Health Care: The Case Study of Arusha Region. *International Journal of Computational Engineering Research*, **3**, 73-77.

[29] Desikan, P., Hsu, K.W. and Srivastava, J. (2011) Data Mining for Healthcare Management. *SIAM International Conference on Data Mining*, Arizona.

[30] Page, D. and Craven, M. (2016) Biological Applications of Multi-Relational Data Mining. http://www.kdd.org/exploration_files/Page.pdf

[31] Brossette, S.E., Sprague, A.P., Hardin, M.K., Waites, B., Jones, W.T. and Moser, S.A. (1998) Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. *Journal of the American Medical Informatics Association*, **5**, 373-381. https://doi.org/10.1136/jamia.1998.0050373

[32] Ridinger, M. (2002) American Healthways Uses SAS to Improve Patient Care. *DM Review*, **12**, Article No. 139.

[33] Gharehchopogh, F.S., Molany, M. and Mokri, F.D. (2013) Using Artificial Neural Network in Diagnosis of Thyroid Disease: A Case Study. *International Journal on Computational Sciences & Applications*, **3**, 49-61.

[34] Shukla, A. and Kaur, P. (2009) Diagnosis of Thyroid Disorders Using Artificial Neural Networks. *IEEE International Advance Computing Conference*, Patiala, 6-7 March 2009, 1016-1020. https://doi.org/10.1109/IADCC.2009.4809154

[35] Prerana, E., Sehgal, P. and Taneja, K. (2015) Predictive Data Mining for Diagnosis of Thyroid Disease Using Neural Network. *International Journal of Research in Management*, *Science & Technology*, **3**, 75-80.

[36] Chang, C.Y., Tsai, M.F. and Chen, S.J. (2008) Classification of the Thyroid Nodules Using Support Vector Machines. *International Joint Conference on Neural Networks*, Hong Kong, 1-8 June 2008, 3093-3098. https://doi.org/10.1109/IJCNN.2008.4634235

[37] Upadhayay, A., Shukla, S. and Kumar, S. (2013) Empirical Comparison by Data Mining Classification Algorithms (C 4.5 & C 5.0) for Thyroid Cancer Data Set. *International Journal of Computer Science & Communication Networks*, **3**, 64-68.

[38] Keleş, A. and Keleş, A. (2008) ESTDD: Expert System for Thyroid Diseases Diagnosis. *Expert Systems with Applications*, **34**, 242-246.

https://doi.org/10.1016/j.eswa.2006.09.028

[39] Chen, H.L., Yang, B., Wang, G., Liu, J., Chen, Y.D. and Liu, D.Y. (2012) A Three Stage Expert System Based on Support Vector Machines for Thyroid Disease Diagnosis. *Journal of Medical Systems*, **36**, 1953-1963. https://doi.org/10.1007/s10916-011-9655-8

[40] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/machinelearning-databases/thyroid-disease