# A Survey on importance of Data Mining in Health care

[1]AMEEFA P K, [2] Lilly Cheerotha
*Assistant Professors, IES College of Engineering,*
*Trissur, Kerala,*

***Abstract*-This survey mainly dealing with the health issues. Their related information is collected and how important data mining is in health care. Now-ever-days health issues are increasing and recording every information are becoming more complex. Gathering the health record by gathering the data sets related to insurance claims, health surveys and other sources including data quality and privacy issues. The analysis of new health conditions using data mining is necessary. Clinical Decision Making by diagnose their issues or disease rather than by the medical experts. Bio medicine and Genetics for finding the specific diseases that are effects of genetics are studied with the help of bio medical and molecular level. Population Health focused mainly on the patterns, trends and causes of specific disease across a population. Health Administration and Policies mainly focus on the insurance plans in the area of health administration. It is believed that by analyzing the related data and extracting the hidden information out of that data, many useful and applicable solutions can be developed. With the increase of implementing electronic systems, such as electronic health records systems, in the health sector, there are massive amount of data being captured every day.**

***Keywords***—*Data Mining; Health Data Analysis; Data Quality; Predictive Modeling; Health Big Data; Data Mining Applications; Classification; Clustering; Association*

## I. INTRODUCTION

Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity. Determining the disease causes, spreading from individual to individual and within community has becoming increasingly global [1]. Electronic system gathers a large amount of data on each day. Data mining has received a lot of attention due to its strong ability of extracting information from data. Due to this large amount of data, data mining algorithm are improved and new versions are adopted to improve reliability. In other words, the collected data in any organization is now seen as a new source of very important information that can directly affect the operational efficiency of that organization, provide higher quality outcomes, and even cut the unnecessary expenditures that waste the budget. Therefore, data mining has received a lot of attention due to its strong ability of extracting information from data.

However, most of these data sets are not very well structured and appropriate for analytically purposes. In addition, health data are usually very complex and not easy to analyze [1]. Therefore, health sector is still a very demanding domain in applying data mining techniques to find trends and hidden information to make the health organizations more cost efficient, and provide complementary clinical solutions.

Hemodialysis (HD) and peritoneal dialysis are the two modalities of dialysis treatment. HD is typically performed in a clinic setting and accounts for more than 80% of the dialysis population. various factors such as climate, environment, water quality and management, education, air pollution, natural disasters, social and many others which are the reasons for the emergence of diseases. To store such a large amount of data or information the sizes of databases are increased very rapidly. Such type of databases consist very useful information. This information may be very useful for decision making process in any field. It becomes possible with the help of data mining or Knowledge Discovery in Databases (KDD).

In this paper, we present the advantages of data mining and special characteristics of health data that makes data mining very important to be considered in health data analysis. In this section, we demonstrated some evidences and explanations about why data mining has received attention, especially in the health domain. In the next section, data mining techniques and concepts are briefly presented. In continue, the applications , necessities, and challenges of data mining in health care have been described and summarized. Finally, we conclude about the influences of applying data mining in health care.

## II. DATA MINING

In order to improve classification accuracy, insignificant parameters and patient data were removed from the data set. For example, all patients with less than 15 visits were removed from the data set. The data for patients who received transplants required special consideration. Transplant‘s success depends on factors such as age, prior time on dialysis, gender, primary cause of the disease, and so on[1].

### *A. Main Techniques*

In traditional data analysis, data for these patients would be censored following the transplants. However, in this dialysis data set, there is only one record per patient. Censorship would imply that data for patients who received transplants would have to be omitted from mining, thus making the data set too small for analysis. Therefore, the dialysis time for each transplant patient was calculated using the dialysis beginning date and the transplant date. The patients with dialysis time (before the transplant) greater than 3 years were classified as above-median while others were classified as below-median. There is no way to determine whether the below-median patients (with the transplant) would have survived on dialysis for more than 3 years. Due to the small data set, it was assumed that without transplant these patients would have died before the median survival time. Also, it was assumed as they were placed on the transplant list, their condition was deteriorating and the chances of them surviving above-median were small. The data set also contained data for four patients who had returned to dialysis treatment following a transplant. Since this situation is rather unusual their data records were removed from the data set. All initial data mining was conducted using a rough-set (RS) algorithm. In the RS theory lower and upper approximations of the concept are computed [11]. As a result, there are two types of A.

#### *1) Classification*

Classification in data mining defines as assigning an object to a certain class based on its similarity to previous examples of other objects. The classification process comes under the predictive method. With classification, new samples of data are classified into known classes. The classification is the initial process of data mining and use algorithms like decision trees, Bayesian classifiers. For classification the data required must be already labeled one.

Classification techniques are widely used in health data analyses, including: analyzing microarray data [10], diagnosing skin diseases [11], performance of different classifiers on cancer datasets [12], predicting cost of healthcare services [13][14], identifying significant factors in healthcare coverage and predicting the status [15].

Table 1-Training set And Predicting Set for Medical Database

Training Set

| Age | Heartbeat Rate | Blood Pressure | Heart Problem |
|---|---|---|---|
| 45 | 75 | 140/64 | YES |
| 28 | 85 | 101/60 | NO |
| 38 | 62 | 105/55 | NO |

Prediction Set

| Age | Heartbeat Rate | Blood Pressure | Heart Problem |
|---|---|---|---|
| 33 | 89 | 142/82 | ? |
| 45 | 52 | 102/56 | ? |
| 87 | 83 | 138/61 | ? |

#### *2) Clustering*

It is important to understand that data mining is a close relative, if not a direct part of data science. Data mining focuses using machine learning, pattern recognition and statistics to discover patterns in data.

Clustering would fall into the machine learning / pattern recognition realm.

Supervised Learning **-** These include machine learning algorithms that have variables used as predictors and a variable to predict. Unsupervised Learning – These algorithms have no variable to predict tied to the data. Instead of having an output, the data only has an input which would be multiple variables that describe the data. This is where clustering comes in.

*3) Association*

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

### *B. Advantages Over the Traditional Statistics*

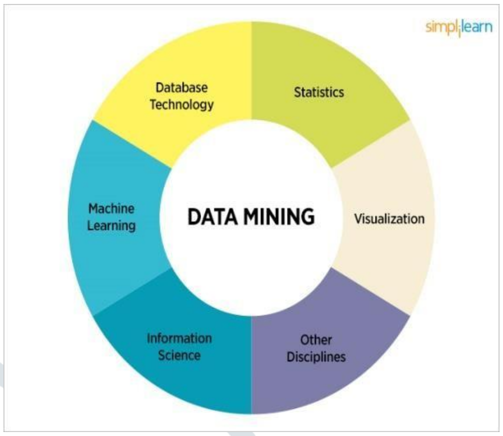
Data mining and statistics are related to learning from data. They are all about discovering and identifying structures in data, with the aim of turning data to information. And although the aims of both these techniques overlap, they have different approaches.

Statistics is only about quantifying data. While it uses tools to find relevant properties of data, it is a lot like math. It provides the tools necessary for data mining. Data mining, on the other hand, builds models to detect patterns and relationships in data, particularly from large data bases.

First, statistics prefers to use more conservative strategies in the first phases of analysis, and in general, employ concrete mathematical methods to run analysis. On the other hand, data mining is open to consider various approaches in regards to mine the data in different orders [17]. Due to this flexibility, data mining uses heuristics as well when facing with real-world issues, so that categorical (discrete) attributes are included in the analysis too [4].

Second, statistics runs analysis only on a sample of data, as this was probably the approach to handle large datasets for analysis in past, and it has retained in this method‘s nature. In contrast, data mining has the ability to consider the whole dataset for analysis which in return provides more reliable results by considering all details of the population.

Third, statistical methods can only work with numeric data [17]. However, there are a lot of categorical (discrete) attributes – e.g. race, gender, diagnosis code – in addition to numeric and even other types of data in the current databases. Most data mining techniques are capable of handling these types of data in addition to numeric data.



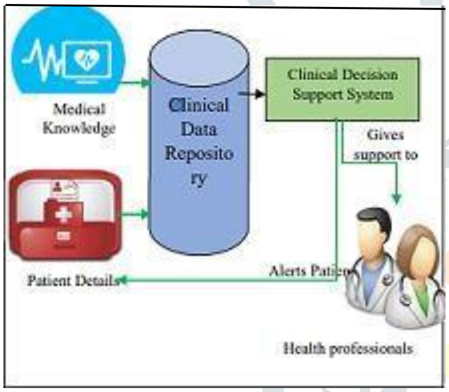
Finally, in statistics, a hypothesis is first created and then the data gets analyzed to prove or reject the hypothesis (hypothetico-deductive analysis). On the other hand, data mining does not consider any clear hypotheses. It starts exploring the data and tries finding knowledge out of the data (inductive analysis) [17]. This can be very useful when studying the prevalence of new diseases that their causing factors are unknown.

## *III. APPLICATION OF DATA MINING IN HEALTHCARE*

Data mining provides a variety of latest method for data analysis and discovers new useful knowledge. In different research areas, data mining is gaining popularity due to its infinite methodologies and applications to mine the information. Data mining techniques plays a vital role for uncovering new trends in healthcare organization which is also for all the parties associated with this field. Usage of such data mining techniques on medical data determines useful trends and patterns that are used in analysis and decision making. This survey features various data mining techniques such as classification, clustering, association, regression in health domain. It can find out some useful knowledge from large databases. Now-a-days large volume of data is being collected and stored at high speed. Traditional data analysis techniques have limitations. Human analyst may take time to discover useful information and much of the data is never analyzed at all. Automated analysis of massive data sets is considered as data mining. Compared with other data mining areas, medical data mining has some unique characteristics [37]. Since medical files are related to human subjects, privacy concern is taken more seriously than other data mining tasks.

### *A. Clinical Decision Making*

Clinical Decision Support System (C_DSS) is interactive application, which performs automatic decision making for clinical activities. This is designed to assist physicians and other health professionals with decision-making tasks by determining the diagnosis of patient data. Fig 1.0 shows the Clinical decision C_DSS helps to review and filter the physician‘s preliminary diagnostic choices to improve the treatments. The Post diagnosis in C_DSS systems is used to mine data to derive associations between patients and their past medical history. Using the patient medical history and clinical research, the application will predict future events. Clinical decision support systems are broadly classified into two main types namely Knowledge based C_DSS and Non-knowledge based C_DSS, which relies on machine learning approaches.



### *B. Clinical and public health implications*

The unchanged global HIV incidence may be related to ignoring acute HIV infection (AHI). This scoping review examines diagnostic, clinical, and public health implications of identifying and treating persons with AHI. AHI among individuals with behavioral and clinical characteristics more often associated with AHI. However, algorithms have not been implemented outside research settings. From a clinical perspective, there are substantial immunological and virological benefits to identifying and treating persons with AHI – evading the irreversible damage to host immune systems and seeding of viral reservoirs that occur during untreated acute infection. The therapeutic benefits require rapid initiation of antiretroviral, a logistical challenge in the absence of point-of-care testing. From a public health perspective, AHI diagnosis and treatment is critical to: decrease transmission via viral load reduction and behavioral interventions; improve pre-exposure prophylaxis outcomes by avoiding treatment initiation for HIV-seronegative persons with AHI; and, enhance partner services via notification for persons recently exposed or likely transmitting.

### *C. Health Administration and Policies*

Were able to discover patterns among health centers that led to policy recommendations to their Institute of Public Health. They concluded that ―data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making.‖ Data Mining in Healthcare‖. The preceding factors remind us of an incident in the Philippines at the Rizal Medical Center in Pasig City in October 2006. Failing to implement strict sanitation and sterilization measures the hospital contributed to the death of several new- born babies due to neonatal sepsis (bacterial infection). No one really knew what was going on until the deaths became more frequent. Upon examining hospital records, the Department of Health (DOH) found that 12 out of 28 babies born on October 4, for example, died of sepsis (Tandoc 2006). With an integrated database and the application of data mining the DOH could detect such unusual events and curtail them before they worsen.

### *D. Heart Disease Prediction*

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "; mined"; to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex "; what if"; queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. IHDPS is Web-based, user-friendly, scalable, reliable and expandable. It is implemented on the .NET platform.

## IV. CONCLUSION

Data mining has great importance for area of medicine, and it represents comprehensive process that demands thorough understanding of needs of the healthcare organizations. Healthcare is one of the major sectors which can highly benefit from the implementation and use of information system. We have provided an overview of applications of data mining in infrastructure, administrative, financial and clinical Health care system. Knowledge gained with the use of techniques of data mining can be used to make successful decisions that will improve success of healthcare organization and health of the patients. Data mining requires appropriate technology and analytical techniques, as well as systems for reporting and tracking which can enable measuring of results. Data mining, once started, represents continuous cycle of knowledge discovery.

*References*

[1] D. Tomar and S. Agarwal, _A survey on Data Mining approaches for Healthcare‘, Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241–266, 2013.

[2] J. Natale, _Leveraging Technology to Revolutionize Canadian Health Care‘, Policy: Canadian Politics and Public Policy, vol. 2, no. 6, pp. 27–30, Dec-2014.

[3] D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining. MIT Press, 2001. [4] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, _Data Mining in Healthcare and Biomedicine: A Survey of the Literature‘, J. Med. Syst., vol. 36, no. 4, pp. 2431– 2448, May 2011.

[5] _The Technology Review Ten‘, MIT Technology Review, Feb-2001. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 1061

[6] U. Fayyad, G. Piatetsky- Shapiro, and P. Smyth,
_The KDD Process for Extracting Useful Knowledge from Volumes of Data‘, Commun ACM, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[7] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques . Morgan Kaufmann, 2006.

[8] U. Fayyad, G. Piatetsky- Shapiro, and P. Smyth,
_From Data Mining to Knowledge Discovery in Databases‘, Commun ACM, vol. 39, no. 11, pp. 24–
26, 1996.

[9] S. Velickov and D. Solomatine, _Predictive Data Mining: Practical Examples‘, in 2nd Joint Workshop on Applied AI in Civil Engineering, Cottbus, Germany, 2000.

[10] H. Hu, J. Li, A. Plank, H. Wang, and G. Daggard, _A Comparative Study of Classification Methods for Microarray Data Analysis‘, in Proceedings of the Fifth Australasian Conference on Data Mining and Analystics, Darlinghurst, Australia, Australia, 2006, vol. 61, pp. 33–37.

[11] H. Cataloluk and M. Kesler, _A diagnostic software tool for skin diseases with basic and weighted K-NN‘, in 2012 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 2012, pp. 1–4.

[12] R. Potter, _Comparison of classification algorithms applied to breast cancer diagnosis and prognosis‘, presented at the 7th Industrial Conference on Data Mining, ICDM 2007, Leipzig, Germany, 2007, pp. 40–49.

[13] G. A. Beller, _The rising cost of health care in the United States: Is it making the United States globally noncompetitive?‘, J. Nucl. Cardiol., vol. 15, no. 4, pp. 481–482, Jul. 2008.

[14] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J.
C. Kryder, R. Pandey, S. Vempala, and G. Wang,
_Algorithmic Prediction of Health-Care Costs‘, Oper. Res., vol. 56, no. 6, pp. 1382–1392, Dec. 2008.

[15] M. H. Tekieh, B. Raahemi, and S. A. Izad Shenas, _Analysing healthcare coverage with data mining techniques‘, Int. J. Soc. Syst. Sci., vol. 7, no. 3, pp. 198–221, 2015.

[16] D. J. Hand, _Data Mining: Statistics and More?‘, Am. Stat., vol. 52, no. 2, pp. 112–118, May 1998.

[17] D. J. Hand, _Statistics and Data Mining: Intersecting Disciplines‘, SIGKDD Explor Newsl, vol. 1, no. 1, pp. 16–19, Jun. 1999.

[18] _Highmark maximizes Medicare revenues with SAS.‘ SAS, 2006.

[19] _Healthways Heads Off Increased Costs with SAS.‘ SAS, 2009.

[20] Y. Zhang, S. Fong, S. Fiaidhi, and S. Mohammed, _Real-time clinical decision support system with data stream mining‘, J. Biomed. Biotechnol., vol. 2012, p. 8, 2012.

[21] T. Haferlach, A. Kohlmann, L. Wieczorek, G. Basso, G. T. Kronnie, M.-C. Bene, J. De Vos, J. M. Hernandez, W.-K. Hofmann, K. I. Mills, A. Gilkes,
S. Chiaretti, S. A. Shurtleff, T. J. Kipps, L. Z. Rassenti, A. E. Yeoh, P. R. Papenhausen, W. -m. Liu,
P. M. Williams, and R. Fo, _Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: Report from the international microarray innovations in leukemia study group‘, J. Clin. Oncol., vol. 28, no. 15, pp. 2529–2537, 2010.

[22] R. Salazar, P. Roepman, G. Capella, V. Moreno,
I. Simon, C. Dreezen, A. Lopez-Doriga, C. Santos, C. Marijnen, J. Westerga, S. Bruin, D. Kerr, P. Kuppen,

C. van de Velde, H. Morreau, L. Van Velthuysen, A. M. Glas, and R. Tollenaar, _Gene expression signature to improve prognosis prediction of stage ii and iii colorectal cancer‘, J. Clin. Oncol., vol. 29, no. 1, pp. 17–24, 2011.

[23] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, _Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring‘, Science , vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[24] L. J. van ‘t Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, _Gene expression profiling predicts clinical outcome of breast cancer‘, Nature, vol. 415, no. 6871, pp. 530–536, Jan. 2002.

[25] R. Kandwal, P. K. Garg, and R. D. Garg, _Health GIS and HIV/AIDS studies: Perspective and retrospective‘, J. Biomed. Inform., vol. 42, no. 4, pp. 748–755, Aug. 2009.

[26] D. Delen, G. Walker, and A. Kadam, _Predicting breast cancer survivability: a comparison of three data mining methods‘, Artif. Intell. Med., vol. 34, no. 2, pp. 113–127, Jun. 2005.

[27] S. Shah, A. Kusiak, and B. Dixon, _Data mining in predicting survival of kidney dialysis patients‘, in Proceedings of Photonics West—Bios 2003, Belingham, 2003, vol. 4949, pp. 73–79.

[28] _First Things First—Highmark makes healthcare-fraud prevention top priority with SAS.‘ SAS, 2006.

[29] Y. M. Chae, S. H. Ho, K. W. Cho, D. H. Lee, and S. H. Ji, _Data mining approach to policy analysis in a health insurance domain‘, Int. J. Med. Inf., vol. 62, no. 2–3, pp. 103–111, Jul. 2001.

[30] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, _Health big data analytics: current perspectives, challenges and potential solutions‘, Int. J. Big Data Intell., vol. 1, no. 1/2, pp. 114–126, 2014.

[31] R. D. Canlas Jr, _Data mining in healthcare: Cu rrent Applications and Issues‘, Carnegie Mellon University, Australia, 2009.

[32] F. Hosseinkhah, H. Ashktorab, R. Veen, and M. M. Owrang O., _Challenges in Data Mining on Medical Databases‘, IGI Glob., pp. 502–511, 2009.

[33] S. Hoffman and A. Podgurski, _Big Bad Data: Law, Public Health, and Biomedical Databases‘,J. Law. Med. Ethics, vol. 41, pp. 56–60, Mar. 2013.

[34] C. Violán, Q. Foguet-Boreu, E. Hermosilla-Pérez, J. M. Valderas, B. Bolíbar, M. FàbregasEscurriola, P. Brugulat-Guiteras, and M. Á. Muñoz-Pérez, _Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity‘, BMC Public Health, vol. 13, no. 1, p. 251, Mar. 2013.

[35] K. El Emam, Guide to the De-Identification of Personal Health Information. CRC Press, 2013.

[36] ESRI White Paper. Enterprise GIS in health and social service agencies; 1999

[37] HianChyeKoh and Gerald Tan, ―Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Vol 19, No 2.