HEALTHCARE DATA MINING LITERATURE REVIEW: 2012 TO 2019

[1]Mohnish Mahamune, [2]Pramod Deo,

[1]Assistant Professor, [2]Associate Professor
[1]School of Computational Sciences,
[1]S. R. T. M. University, Nanded, India

Abstract: In this paper we review the current research being utilized using the data mining methods for the analysis and forecast diagnosis of numerous diseases, underlining critical concerns, and reviewing the methods in a set of cultured practices. The motto of this review is to classify and evaluate the most frequently used by algorithms of data mining on healthcare databases. The following algorithms have been known as applications of data mining in healthcare: Decision Trees (DT's) C4.5 and C5, Support Vector Machine (SVM), artificial neural networks (ANNs) and their Multilayer Perceptron model, Naïve Bayes, Logistic Regression, Genetic Algorithms (GAs) / Evolutionary Programming (EP), Fuzzy Rules. Review present that it's challenging to claim one algorithm of data mining as the best appropriate for the healthcare diagnosis of diseases. At present certain algorithms achieve better than others, but there are some cases when a pipeline of the facts for said algorithms claims outcomes more operational.

Index Terms **- Data Mining, Knowledge Management, Healthcare Data Mining, Data Mining review.**

I. INTRODUCTION

Data Mining is useful process to mine huge amount of data and extract the hidden information or discover significant pattern. The application of data mining in healthcare domain is to discover reliable procedure for disease prediction system using clinical and diagnosis data [9].

Data mining is described by Fayyad as a "non-trivial method of extracting implicit, previously unknown, and potentially useful data from data stored in a database" [1]. Healthcare databases have a tremendous amount of data, but there is an absence of methods and resources for organizational investigation to evaluate secret or appropriate knowledge. In their attempt to prescribe less costly solutions and alternatives, suitable algorithms and computer-based information systems and/or decision support systems can support doctors/practitioners. The effective and precise application of computerized system requires comparative learning of various techniques accessible. This paper presents the review of the current research and applications is accepted and using the applications of data mining for the diagnosis and forecast of various diseases.

Today, healthcare institutions are able to produce and accumulate a tremendous amount of data. This increase in data size periodically enables the data to be retrieved when needed. The practice of data mining approaches means that there will be minimal interesting patterns. In order to enhance job efficiency and improve decision-making excellence, it is possible to use the information added to this approach in the correct order.

Over all, a great need for a new generation of theories and data analytics tools is required to help people extract valuable knowledge from the increasing amount of digital data [1]. Information technologies are applied progressively often in healthcare organizations to support doctors and practitioners in their decision making. Computer systems algorithms used in data mining can be appreciated to control humans and to offer assistance to decision-making processes [2]. The characteristics of data mining is to recognize relationships, patterns, and models that have provisions for predictions and decision-making practice for diagnosis and treatment. In hospitals, Data Mining models are used to support for complex decision making. In addition, applications of Data Mining in healthcare empowers the inclusive management of medical knowledge and its secure conversation among healthcare workers and beneficiaries [3]. Knowledge Obtained from data analytics improve the decision-making efficiency and help to avoid human error [4]. When there is a large amount of data that people need to process, decisions are normally of low quality. [4]. The process of analysing raw data using a computer and extracting its significance is data mining. The method is also defined as the discovery of previously unknown and potentially useful knowledge from large volumes of data (unstructured) [5]. It is likely to predict the developments and behaviour of patients or diseases. This is done by studying data from different perceptions and discovers relationships and connections among apparently distinct information. Previously unknown trends and patterns from a database are discovered using techniques of data mining and turn information into meaningful solutions. [6]. The paper is organized accordingly: First, express the methodology of research used in this study. Two, we review them with different benchmarks and applications in healthcare three, lastly identify the best algorithms for disease diagnosis and forecast, and finally we show the conclusions of our work.

II. DATA MINING AN OVERVIEW

Data mining is a kind of tool which individuals employ to extract valuable knowledge from huge set of information or even raw data. This data processing is known as Knowledge Discovery in Database (KDD). The KDD sequentially introduce, Data selection, Data cleaning, Data processing, Data integration, Data transformation, Data mining, Pattern evaluation, and Knowledge representation. The process of extracting information to spot patterns, trends, and useful data that might allow the business to require the decision based on data-driven methods from huge sets of knowledge is named data processing. Data mining is the act of automatically checking out vast data stores to look for trends and patterns that exceed basic procedures for study. For data segments, data mining utilizes complicated mathematical algorithms and determines the likelihood of future events. In addition, data mining is called knowledge discovery from databases (KDD). Data Mining is a method used by companies to collect specific knowledge to solve business challenges from large databases. It mainly transforms raw data into information that is useful.

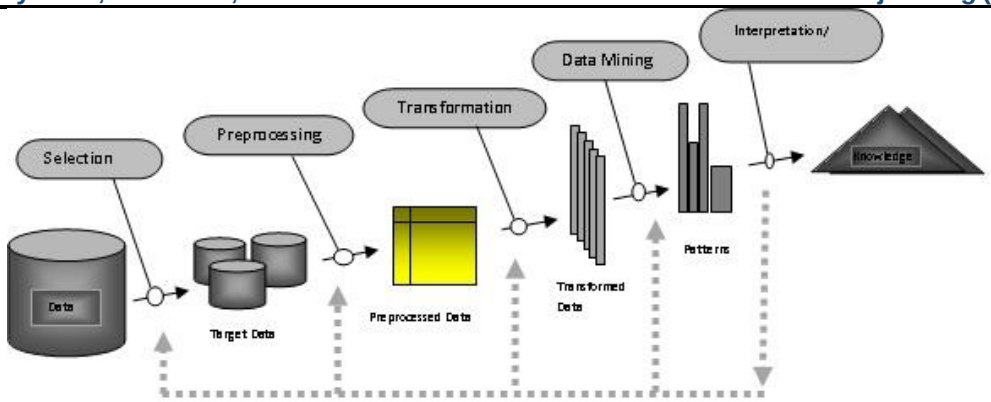


Fig. 1 Data mining process for Knowledge discovery from databases. (Source: Fayyad, et.al., 1996)

III. Data Mining Techniques

1. Classification: This analysis is used for the classification of essential and specific metadata and data. These techniques and methods of data mining help to classify data into various modules.
2. Clustering: The discovery and study of clustering is a technique of data mining to identify similar data. The modifications and similarities between the data are recognized by this method.
3. Regression: Regression exploration is the data mining technique of recognizing and investigating the association among variables. It is used to identify the possibility of a particular variable, given the presence of other variables.

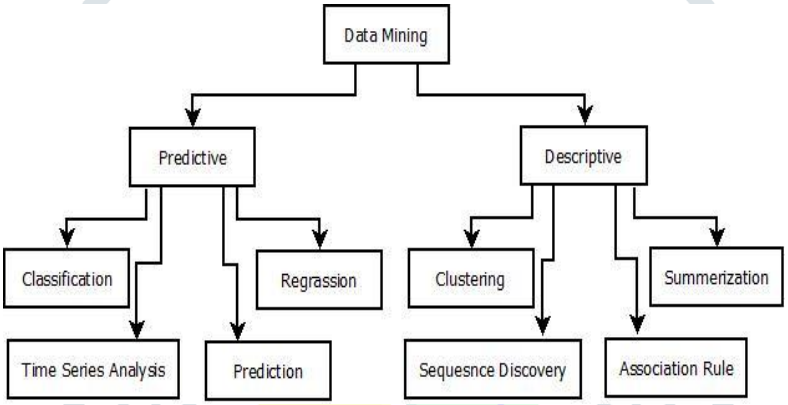


Fig. 2 Data Mining Techniques.

4. Association Rules: The Association Rule method helps to define the relation between two or more objects. In the data set, it learns a hidden pattern.

5. Outer detection: Outer detection technique brings up observation of data items in the data sets which unable to match a predictable pattern or expected activities. In a number of domains, such as interference, detection of fault or fraud, etc., this approach may be used. Outer identification is also referred to as Outlier or Outlier Analysis.

6. Sequential Patterns: This approach helps to evaluate or identify related series patterns or trends for the presumed duration in transaction data sets.

7. Prediction: Prediction is used to group other data mining techniques, such as trends, sequential patterns, clustering, classification, etc. For predicting a potential incident or incidents, it analyses past data events or occurrences in the correct structure sequence.

IV. Advantages of Data Mining

- Methods and techniques for data mining allow organizations to find knowledge-based data.
- Data mining enables efficient organizational and output improvements to be produced.
- Associated with other statistical and numerical data applications, data mining is cost-efficient.
- Data Mining supports the decision-making process of an organization.
- It makes it possible to discover hidden patterns automatically, as well as to predict trends and behaviours.
- It can be induced in the innovative system as well as the existing platforms.
- It is a rapid method that makes it easy to investigate large volumes of data in a short time for new users.

V. Data Mining Applications

Data Mining is primarily used by retail, communication, marketing, financial, price, product positioning, consumer desires, and effect on customer loyalty, sales, and corporate profits organizations with penetrating customer demands. Data mining enables a retailer to use consumer consumption point-of-sale criteria to create goods and promotions that allow the company to attract customers.

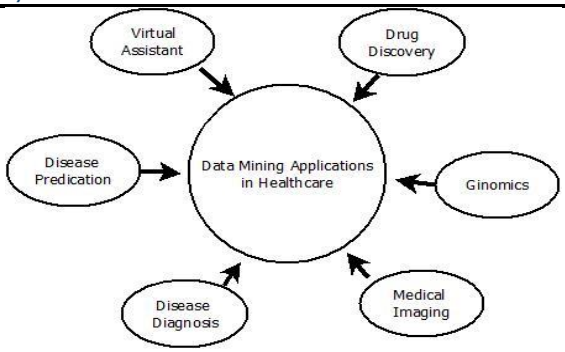


Fig. 3 Data mining applications in various sectors.

VI. Data Mining in Healthcare

Data mining in healthcare has brilliant potential to improve the health system. It uses data and analytics for improved insights and to recognize best practices that will improve health care services and reduce costs. Data mining techniques are used by researchers, such as multi-dimensional databases, machine learning, soft computing, visualization of data, and statistics. It is possible to use Data Mining to predict patients of each kind. The steps ensure that patients receive ICU treatment in the right place and at the right time when appropriate. Data mining also allows insurers of healthcare facilities to detect fraud and abuse.

VII. Applications of Data Mining in Healthcare, Literature Review

Due to the rapid increase in the number of electronic health records, data mining holds enormous potential for healthcare services. Earlier physicians and Doctors stores patient data in hard copy where the data was quite difficult to hold. Digitization and invention in technology and techniques decrease human efforts and ensures data easily assessable. For example, computer systems save a large patient information with accuracy, and it helps to improve the quality of the whole data management system. Still, the main challenge is what should healthcare facilities providers do to filter all the data professionally? This is the place where data mining has proven to be very useful. Researchers are utilizing different approaches like classification, decision trees, clusters, neural networks, and time series to publish research. Healthcare, however, has historically been unable to integrate the new findings into daily practice.

A different approach to mine the data in healthcare:

A three-system method is the best technique for bringing data mining beyond the principle of academic science. The introduction of all three frameworks is the way to push a real-world change of any healthcare analytics initiative. Unluckily, all three of these programs are run by unusual healthcare organizations.

These are the following three systems:

Fig. 4 approach to mine the data in healthcare.

- The analytics system: To collect knowledge, understand it, and standardize measurements, the analytics system combines technology and expertise. The basis of the framework is to integrate health, patient satisfaction, financial, and other data into an enterprise data warehouse (EDW).
- The content system: The content method requires the standardization of work on knowledge. This extends evidence-created best results to the delivery of treatment. Every year, researchers make important discoveries about scientific practice, but it has been mentioned earlier that it takes a very long time to integrate these results with clinical practice. A comprehensive infrastructure of content enables organizations to easily bring up-to-date medical conformation into daily practice.
- The deployment system: The implementation framework involves pushing change management over new systems of hierarchy. It mainly involves enforcing group processes that consistently enable best practices to be adopted by an enterprise-wide organization. To push the application of best practices in an organization, it needs a true ordered modification.

Application of Data Mining in Healthcare:

Numerous organizations have used data mining intensively and widely. Data mining is desirable and more popular nowadays in healthcare. All sorts of parties that are interested in the healthcare industry will greatly benefit from data mining applications. In customer relationship management, fraud identification and exploitation, productive patient care, and best practices, fair healthcare facilities, data mining will assist the healthcare industry. Huge quantities of data produced by transactions in healthcare are too complicated and too broad for traditional methods to manage and analyze.

Data mining provides the framework and techniques to transform raw data into valuable information for data-driven decision purposes.

- Treatment effectiveness: Data Mining applications can be used to assess the effectiveness of medical treatments. Data mining analyzes effective course of action by comparing symptoms, causes, and developments of treatments.
- Healthcare management: Data mining tools can be used to classify and monitor patients with chronic diseases and ICUs, decrease the number of hospital admissions, and assist in the administration of health care. In order to analyze large data sets and statistics, data mining is used to look for trends that can verify fraud.
- Customer relationship management: Customer and administration relations are very important for any organization to reach business goals. Customer relationship management is the main tactic to managing relations between commercial administrations for example retail sectors and banks, with their customers. Customer contact takes place via call centers, billing offices, and settings for outpatient treatment.
- Fraud and abuse: Data mining fraud and abuse claims can focus on unfortunate or wrong prescriptions and fraud insurance and medical claims.

VIII. Outcomes of relative analysis of various disease in Healthcare:

A relative study of data mining applications in the healthcare area by many experts as follows.

Table-1: Comparative table of mining tools with the accuracy level.

Sr. No.	Author	Data mining tool	Algorithm / Technique	Type of Disease
1	Yeh et al. 2009	Swarm Intelligence Method	Particle Swarm Optimization	Breast Cancer Diagnosis Model
2	Shouman M. et al. 2012	k-Nearest Neighbor	Cluster Modeling	Heart Disease
3	Boris Milovic et. al. 2012	Data Mining Methods	Decision Making	Disease Diagnoses
4	Mookiah, M.R.K, et al. 2012	SVM classifier	Wavelet Energy Features and Kernel Function	Glaucoma Diagnosis Process
5	Arvind Sharma et al. 2012	WEKA 3.6	Rule Mining and Classification	HIV/ AIDS
6	ShwetaKharya et al. 2012	K-means Clustering	MAFIA	Brain Cancer
7	B.Renuka Devi et al. 2013	SPSS	C 5.0 Statistics	Dengue Diagnosis
8	Bhatla et al. 2013	Neural Network	Genetic Algorithm	Heart Disease
9	Taranath NL et. al., 2013.	Machine-Learning	Prediction Model	Heart Diseases Diagnosis
10	Matthew Herland et. al., 2014.	Big Data Analysis	Prediction Model	Occurrence of Diseases
11	KasraMadadipouya, 2015.	C4.5	Decision Tree	Classification of Disease
12	H. M. Zolbanin et al. 2015	Predictive Model	Decision Support Systems,	Cancer
13	Anna L. Buczak et al. 2015	Fuzzy Logic	Rule Mining and Classification	Malaria Fever
14	Hogenboom, F., et al. 2016	Decision Support Systems	TEXT Extraction Method	Classification of Disease
15	Murphy, D.R. et al. 2016	Big Data	Computerized Triggers	Chest Imaging Results
16	S. Chidambaranathan et al. 2016	Machine Learning Algorithms	Feature Extraction by Hybrid of K-Means	Breast Cancer
17	MaS. Durga, et.al., 2017	SVM	Predictive Models	Lung Disease Prediction
18	Fadil Maulana et. al. 2017	KNN	Feature Selection	Diabetes Disease Type 2
19	Jiang F. et al. 2017	Artificial intelligence	Support Vector Machine	Neurology and Cardiology
20	Voyant, C., et al. 2017	Machine learning	Prediction Model	Solar Radiation
21	Ekin, T., et al. 2017	Concentration Function		Medical Fraud Assessment
22	Bardy, G.H. et al. 2018	Fuzzy Logic	Configurable Arrhythmia Analysis	Health Monitoring Apparatus
23	Shakuntala Jatav, 2018	SVM	RF algorithm	Diabetes, Kidney, And Liver Diagnosis
24	McFee, B. et al. 2018	Statistical Method	Computational Analysis	Sound Scenes
25	Mona Nasr et al. 2019	K Means Clustering	Cluster Modeling	Diabetes Disease
26	Indra Boruah et al. 2019	Linear Programming	Predictive Models	Malaria Fever
27	Md Imran Alam et al. 2019	Bayes Classification	C4.5	Diabetes Disease

Mostly data mining methods and tools are used to forecast the outcomes from the recorded information on healthcare issues. Numerous data mining tools are used to predict the accuracy level of various healthcare problems.

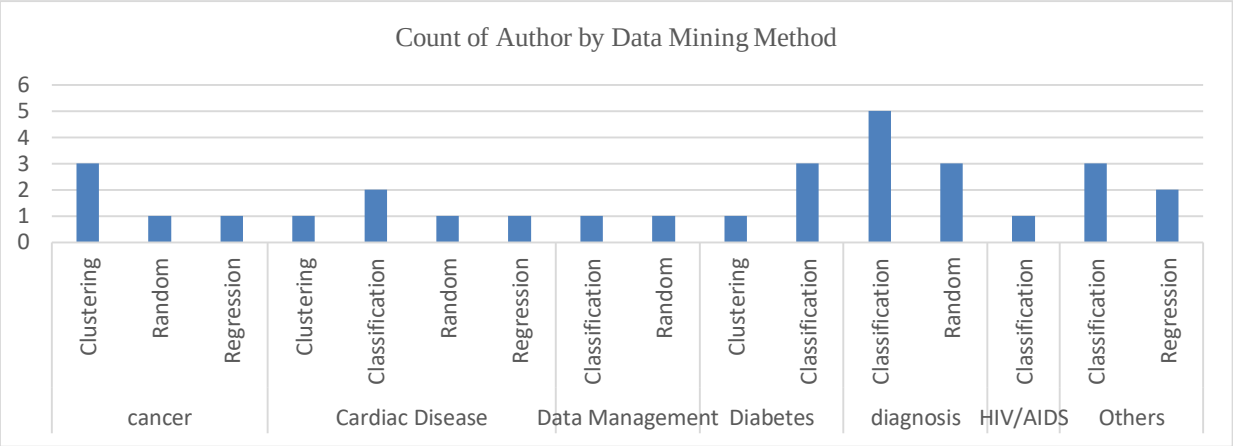
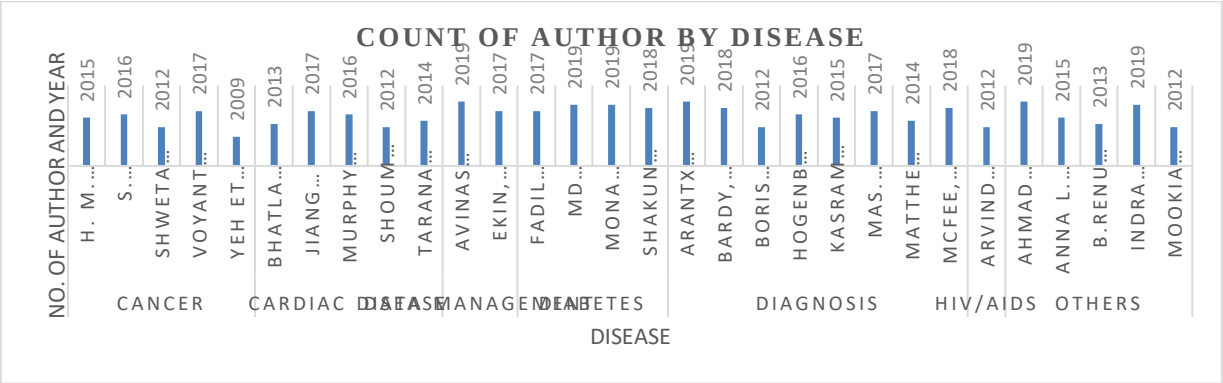


Fig 7: Demographics of publications according to Data mining method and algorithms used



IX. Advantages of Data Mining in Healthcare

The data organizations streamline and powers the workflow of health care organizations. Healthcare organizations reduce decision-making efforts and deliver new valued medical knowledge through the application of data mining popular in data frameworks. Predictive models give the best information support and knowledge to healthcare workers. The goal of predictive data mining in medicine is to create a predictive model that is transparent, offers accurate forecasts, and helps physicians strengthen their processes of diagnosis and treatment planning. An important application of data mining is for bio-medical processing connected by inner strategies and reactions to rise the condition, whenever there is an absence of knowledge around the connection between several subsystems, and when the typical examination methods are unproductive, as it is often in the situation of nonlinear associations.

X. Discussion

The accuracy of the data mining methods varies depending on the features of the data sets and the size of the data set between the training and testing sets from the papers examined and discussed. Highly imbalanced data sets are the common characteristics of healthcare data sets, whereby the majority and minority classifiers are not balanced, resulting in inaccurate prediction when run by the classifiers. The missing values are other features of healthcare data sets. As the data available is typically on a small scale, the sample size of the data is also seen as other features. To solve all these problems, there is no one suitable method of data mining.

XI. Conclusion

Today, there have been many efforts with the goal of successful application of data mining in healthcare institutions. The primary potential of this technique lies in the possibility for research of hidden patterns in data sets in the healthcare domain. These patterns may be used for clinical diagnosis. However, by default, available raw medical knowledge is widely dispersed, varied, and voluminous. Such data must be collected and stored in structured formats in data repositories, and can be incorporated into the hospital information system. Data mining technology provides a customer-oriented approach towards new and hidden patterns in data, from which the knowledge is being generated, the knowledge that can help in providing medical and other services to the patients.

This paper aimed to compare the different data mining applications in the healthcare sector for extracting useful information. It is a difficult task to predict diseases using data mining applications, but it significantly reduces human effort and improves diagnostic accuracy. In terms of human capital and skills, designing successful data mining tools for an application could minimize costs and time constraints. It is such a dangerous job to explore knowledge from medical data as the information found is noisy, meaningless, and huge too. In this case, the methods of data mining are useful in exploring medical data information and it is very interesting. This study shows that a combination of more than one data mining technique is a single technique in which diseases are detected or predicted.

REFERENCES

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. G. R, 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, CA.

[2] Candelieri, A., Dolce, G., Riganello, F., & Sannita, W. G., 2011. Data Mining in Neurology. In Knowledge- Oriented Applications in Data Mining, pp. 261-276.

[3] Bushinak, H., AbdelGaber, S., & AlSharif, F. K., 2011. Recognizing the Electronic Medical Record Data from Unstructured Medical Data Using Visual Text Mining Techniques Prof. Hussain Bushinak., (IJCSIS), Vol. 9, No. 6, 25-35.

[4] Eapen, A. G., 2004. Application of Data mining in Medical Applications. Ontario, Canada, University of Waterloo.

[5] Milovic, B., 2012. Usage of Data Mining in Making Business Decision. YU Info 2012 & ICIST, (pp. 153- 157).

[6] boirefillergroup.com, Data Mining Methodology,2012, from Boire-Filler Group: http://boirefillergroup.com/methodology.php

[7] M. Durairaj, V. Rajani, 2013. Data Mining Applications in Healthcare, International Journal of Scientific & Technology Research Volume 2, Issue 10.

[8] Ameera M. Almasoud, Hend S., Al-Khalifa, Abdulmalik Al-Salman, 2015. Recent developments in data mining applications and techniques, Tenth International Conference on Digital Information Management (ICDIM), 21-23.

[9] Neesha JothiNur'Aini Abdul Rashid Wahidah Husain, 2015.The Third Information Systems International Conference Published by Elsevier.

[10] Umair Shafique, Fiaz Majeed, 2015. Data Mining in Healthcare for Heart Diseases, Innovative Space of Scientific Research Journals, Vol. 10 No. 4, pp. 1312-1322.

[11] Mousa Albashrawi, 2016. Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015, Journal of Data Science 14, 553-570.

[12] Sheenal Patel, Hardik Patel, 2016. Survey of Data Mining Techniques Used in Healthcare Domain, International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2.

[13] Ionuț Țăranu, 2015. Data mining in healthcare: decision making and precision, Database Systems Journal vol. 6, no. 4.

[14] Sajida Perveen, Muhammad Shahbaz, 2016.Performance Analysis of Data Mining Classification Techniques to Predict Diabetes, Procedia Computer Science, Volume 82, Pages 115-121.

[15] Hoa Hong Nguyen, Farhaan Mirza, 2017. A review on IoT healthcare monitoring applications and a vision for transforming sensor data into real-time clinical feedback, IEEE 21st (CSCWD)

[16] Ahsan Humayun, Adeel Waqar, 2017. A Comparative Study on Usage of Data Mining Techniques in Healthcare Sector, International Journal of Computer Applications (0975 – 8887) Volume 162 – No 6.

[17] Pasquale Pace; Gianluca Aloi, 2019. An Edge-Based Architecture to Support Efficient Applications for Healthcare Industry 4.0, IEEE Transactions on Industrial Informatics, Vol. 15, Issue 1.

[18] Bruno Samwaysdos Santos, Maria Teresinha ArnsSteiner, 2018. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018, https://doi.org/10.1016/j.cie.2019.106120.

[19] Nandini Nayar, Sachin Ahuja, 2019. Swarm Intelligence And Data Mining: a review of literature and applications in healthcare, ICAICR '19:Article No.: 12 Pages 1–7 https://doi.org/10.1145/3339311.3339323

[20] Ramin Ghorbani, Rouzbeh Ghousi, 2019. Predictive data mining approaches in medical diagnosis: A review of some disease's prediction, International Journal of Data and Network Science, Vol. 3 Issue 2 pp. 47-70.