IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Survey on Prediction and Analysis of Diabetic Data using Machine Learning Techniques

Monika[1], Pooja Sharma[2]

Department of Computer Science & Engineering,
IKG Punjab Technical University main campus Kapurthala, Punjab,India

Sainimonika038@gmail.com, sharma_pooja@live.com

Available online at: www.ijcseonline.org

Accepted: 08/Jun/2018, Published: 30/Jun/2018

Abstract - In the current era of technology, evolution of medical sciences becomes an active field of research as people have more curiosity towards their health. Different techniques of data mining are used to mine the information from various data patterns. Prior, PC was used to manufacture an information based clinical result which utilizes learning from therapeutic specialists and moves this information into PC calculations physically. This process takes lot of time and gives subjective results as this information only depends on medical professional only. To overcome these type of problems various techniques of machine learning are used to extract important medical patterns from the raw data. In this paper, we have critically analyzed various data mining techniques to gather informative patterns from data sets in medical sciences.

Keywords – Informative Patterns, Clinical Databases, Data Mining, Prediction, Machine Learning.

I. Introduction

Computational health informatics is an emerging research topic which involving various sciences such as biomedical, medical, nursing, information technology, computer science, and statistics [1]. To extract hidden patterns and relationships from large databases, data mining merges statistical analysis, machine learning and database technology. In several areas of medical services, including prediction of the effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data, data mining techniques have been applied [2]. In medical science, doctor’s facilities introduced different data frameworks with a lot of information to manage medical insurance and patient information but unfortunately, data are not mined to discover hidden information for effective decision [2][3]. Clinical test outcomes are regularly made on the basis of doctors’ perception and experience rather than on the knowledge enrich data masked in the database and sometimes this procedure prompts inadvertent predispositions, doctor’s expertise may not be capable to diagnose it accurately which affects the disease diagnosis system [2] [3].

Data warehousing and Data Mining offers a comprehensive support for gathering, analyzing and presenting medical data. Clinical decisions are often made using the doctor’s prescription and experience in his or her field rather than the knowledge base which is rich in data hidden in the database [4]. Usually such practices results in errors, wrong advice to the patients in case if the doctors are in fatigue and stress, unwanted biases and also leads to the extravagant medical price which directly affects the quality of services provided to patients.[4]

Clinical decision support integrated with computer generated patient records could enhance patient safety, diminish medical errors, improve victim outcome and reduce unfavorable practice variation [5].

Rest of the paper is organized as follows. Section 2 describe data mining, section 3 presents basic model of machine learning, section 4 presents related work, finally section 5 conclude the paper.

II. Data Mining

The term data mining by and large alludes to the procedure of naturally looking at expansive databases to separate valuable and shrouded designs. A similar way information revelation in wide region computerized reasoning which is otherwise called measurable investigation or machine taking in, the idea of information mining develops to find concealed guidelines and examples from the gigantic measure of information. The term Data mining is not quite the same as machine learning and measurements in the way that it oversees extensive volumes of information, which is fundamentally put away on plate. That is, information mining manages "learning disclosure in databases." The extraction of non inconsequential, suggested, concealed and significant learning from huge volume of datasets is performed by information mining. Disclosure of the helpful learning that can enhance capability of procedures can't be taken care of physically.

Data mining typically actualizes two sorts of methodologies:

- Supervised Learning Strategy
- Unsupervised Learning Strategy
- In a supervised learning method, a training set is already available which is used to learn parameters. Classification algorithm uses supervised learning method approach. Every one of these information mining strategies utilizes an alternate approach contingent on the motivation behind demonstrating objective [6]. There are generally two basic displaying goals viz. Grouping and Prediction. Order demonstrate predicts the clear cut information that is in discrete and unordered frame though expectation show predicts the persistent esteemed information [6]. In unsupervised learning method, no training set is available to learn the parameters. Clustering algorithm uses
- unsupervised learning method approach. There are various clustering algorithms like K-mean clustering algorithm, K-mediod algorithm, DBSCAN and OPTICS, hidden markov algorithm. Unsupervised learning provides the capability to learn more larger and complex models. In unsupervised learning strategy, the learning can be preceded in hierarchical fashion from the observations resulting into ever more deeper and abstract levels of representation.

III. Machine Learning

Machine learning is the logical field managing the manners by which machines gain as a matter of fact. For some researchers, the expression "machine learning" is indistinguishable to the expression "counterfeit consciousness", given that the likelihood of learning is the principle normal for a substance called savvy in the broadest feeling of the word. The reason for machine learning is the development of PC frameworks that can adjust and gain from their experience. A more formal meaning of machine learning is given by Mitchel [7]: A PC program is said to gain for a fact E regarding some class of assignments T and execution measure P, if its execution at errands in T, as estimated by P, enhances with encounter E.

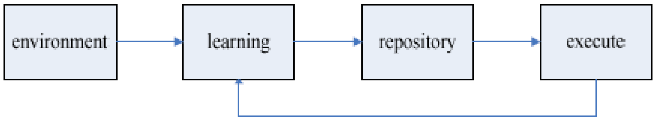

Figure 1. Basic Model of Machine Learning

Machine learning has evolved from study of pattern recognition and computational learning theory. It is most efficient way utilized in the area of data analytics for prediction purposes through devising some algorithms and models. These analytical models let engineers, researchers, data scientists, & analysts to produce valid and reliable discussions and results. It also assists in searching some hidden features or patterns by historical trends and learning in data. Machine learning's most important task is feature selection. Model is developed on the basis of results obtained from training data; this is the reason behind non-interactive nature of machine learning algorithms. To make precise prediction, it studies past observations. It is hard to make precise prediction rule, on the basis of these rules algorithm can be developed [8].

IV. Related Work

Shouman et al. determines the gap resulting from the previous theories of research on diagnosis and treatment of heart disease and introduces a model to systematically close those occurring gaps to discover if we apply techniques of data mining to the heart disease treatment data then it can provide a reliable performance than it is achieved in diagnosing heart disease. In this paper creator has utilized half breed information mining strategies which incorporates credulous thickness, stowing calculation and bolster vector machine. In this writing study, the distinctive datasets which are being utilized as a part of the earlier year papers utilizing diverse systems are being characterized and their precision is estimated so as to determine the various algorithms of data mining used in the diagnosis of heart disease. The author uses single type algorithm like Naïve Bayes, Decision Tree, Bagging algorithm and hybrid type algorithm like Fuzzy-AIRS-K-nearest neighbor, Neural network ensembles and then determine the accuracy. The results show that the hybrid type approach is having more accuracy than the single type as the maximum accuracy achieved by using single data mining technique is 84.14% by naïve bayes while the accuracy achieved by using hybrid data mining technique is 89.01% by neural network ensemble. This paper results in the output that heart disease can be predicted with higher accuracy with hybrid data mining techniques [6].

Yuvaraj et al. proposes the novel usage of machine learning calculations in hadoop based groups for diabetes expectation. Human services frameworks are simply intended to address the issues of expanding populace all around. Individuals around the world are influenced with various kinds of deadliest maladies. Among the diverse kinds of ordinarily existing illnesses, diabetes is a noteworthy reason for visual deficiency, kidney disappointment, heart assaults, and so on. Medicinal services observing frameworks for various illnesses and indications are accessible all around the globe. The quick advancement in the fields of Information and Communication Technologies influenced wonderful upgrades in wellbeing to mind frameworks. Different Machine Learning calculations are proposed which computerizes the working model of human services frameworks and improves the exactness of ailment expectation. Hadoop bunch based appropriated figuring structure bolsters in effective handling and putting away of amazingly vast datasets in cloud condition. The outcomes demonstrate that the machine learning calculations can ready to deliver profoundly precise diabetes prescient medicinal services frameworks. Pima Indians Diabetes Database from National Institute of Diabetes and Digestive Diseases is utilized to assess the working of calculation [9].

MacDougall et al. brings the use of research based evidence into practice so as to develop clinical guidelines into practice. This paper provides a review of current research on the integration of Health Information Technology (HIT) into clinical guidelines so as to achieve more accurate results [10].

Srinivas et al. uses data mining application techniques in the Healthcare and the prediction of Heart Attacks. The author has deeply examined the use of data mining techniques in classification such as Rule based technique, Decision Tree technique, Naïve Bayes Classification technique and Artificial Neural Network technique for the extraction of huge amount of patterns from the abundant data which is not mined so as to discover the hidden information from the data. The author implemented the data preprocessing tasks and decision making one dependency augmented Naïve Bayes classifier (ODANB) and Naïve credal classifier 2 (NCC2) and then the ODANB is equated with the existing methods that improve the Naïve Bayes with the Naïve Bayes itself. It can predict whether the patient is suffering from a heart disease or not using several medical parameters such as patient's age, sex, blood pressure and blood sugar etc. In the rule based technique the rules were piled in the database in the form of IF-THEN rules that is the antecedent part resulting in the conclusion part. Decision Tree includes Classification and Regression Tree (CART), Iterative Dichotomized 3 (ID3) and C4.5. The results when compared show that the ODANB is better than the other methods likes Naïve Bayes. The author also concluded that there are several problems and constraints of using different algorithms of data mining. The hidden patterns can be extracted regarding the prediction of heart attack from data warehouses [11].

Sundar et al. finds out the accuracy of the result by using the K-means Clustering Algorithmic Technique for the prediction and diagnosis of Heart disease. It uses two datasets – real and artificial datasets. The real dataset is the dataset taken from the real life patients of hospitals and patients of laboratory tests whereas the artificial dataset is the dataset taken from the UCI machine learning Databases, 2004.The author in this research mainly focuses on the prediction of Heart disease using K-means Clustering in context of Data Mining. The author first pre-processes the dataset which includes the various steps like eviction of duplicate records, finding out the missing values from the dataset, removing the outliers and noise and normalizing the various values which are used to portray the information in the databases. Then the preprocessed heart disease data is taken and clustered using the K-means algorithm. 13 attributes were taken in the dataset and the performance and analysis of the various algorithms like Decision Trees, Naïve Bayes, Neural network and K-means is done. Each algorithm is compared with another algorithm in terms of its accuracy and time taken to predict .According to this research been taken place the highest accuracy is of the K-means with 66.00% and the time taken is 8 second. The second highest accuracy is 39.96% with the shortest time taken that is of 4sec by the neural network. The lowest accuracy is of Decision Tree with 24.73% with time taken of 10 seconds [12].

Devi et al. presents the development of an amalgam model for classifying Pima Indian diabetic database (PIDD). This amalgam model combines K-means with K-Nearest Neighbour (KNN). They compare the results of simple KNN with cascaded K-means and KNN for the same k-values. It uses the following algorithms: KNN Classifier, K-means partitioning and Amalgam KNN. The dataset is taken from UCI Machine learning data repository for diabetes mellitus that is PIDD. This dataset is from Indian Pima Diabetes Datasets. The results are then compared by measuring the statistical measures such as accuracy, sensitivity and specificity and calculated using WEKA tool. For k=5, K-means and KNN has accuracy of 97% while the simple KNN has the accuracy of 73.17% and Amalgam KNN has accuracy of 97.4%. For k=3, Amalgam KNN has accuracy of 96.87% and simple KNN has accuracy of 72.65%.The author concluded that performance of the algorithm increases if the value of K increases [13].

Thangarasu et al. predicts the diabetic disease from clinical database by using Neural Network algorithm. This research was being conducted with the various objectives like it identifies the various complications that cause diabetes from clinical databases through Fuzzy logic Techniques. It develops a Hybrid Genetic Algorithm that computes the best fitness value which is used for evaluating the prediction accuracy of diabetes from clinical databases. It also identifies the type of diabetes the patient is suffering from through data clustering algorithms from the clinical database. Then, it will analyse the performance of projected algorithms. It uses hybrid system model to identify diabetes mellitus, its types and complications which uses certain algorithms like Neural-network algorithm, Fuzzy logic techniques, Hybrid Genetic Algorithm and Data clustering algorithms. The dataset is collected via questionnaire distribution with participants. All the dataset was organized and analysed using a computer program SPSS 20. This model was successfully implemented with input as symptoms that may appear in an individual during the early stages of diabetes and also based on the physical condition of the individual. This research study avoids the patients from undertaking certain blood tests, checking the diastolic and systolic blood pressure etc. Thereby creating a user friendly interface and environment for the patient's without any requirement of a doctor or hospital staff [14].

Anand et al. discourse on setting up a connection between diabetes hazard prone to be created from a man's every day way of life exercises, for example, his/her dietary patterns, resting propensities, physical movement alongside different pointers like BMI (Body Mass Index), abdomen outline and so forth. Diabetes Mellitus or Diabetes has been depicted as more terrible than Cancer and HIV (Human Immunodeficiency Virus). It creates when there are high glucose levels over a drawn out period. As of late, it has been cited as a hazard factor for creating Alzheimer, and a main source for visual deficiency and kidney disappointment. Aversion of the ailment is an interesting issue for investigate in the medicinal services group. Numerous procedures have been found to discover the reasons for diabetes and cure it. At first, a Chi-Squared Test of Freedom was performed trailed by utilization of the Truck (Classification and Regression Trees) machine learning calculation on the information lastly utilizing Cross-Validation, the inclination in the outcomes was expelled [15].

Chitkara et al. done research to recognize the vocal qualities of patients having type 2 diabetes mellitus with sound people. There has been much ongoing exploration utilizing voice parameters of obsessive people. The investigation has been completed on 177 voice tests in age gathering of 40-60 years including solid and diabetic guys and females. The acoustic investigation was finished utilizing MDVP. The voice parameters like jitter, sparkle, smoothed sufficiency irritation remainder, clamor to symphonious proportion, relative normal annoyance, adequacy bother remainder indicate huge distinction in their qualities for type 2 diabetes mellitus patients when contrasted with non-diabetic people. Results are viewed as an imperative advance towards non-obtrusive conclusion of sort 2 DM infection [16].

Ghuse et al. these days there is tremendous measure of information put away in certifiable databases and this sum keeps on developing quick. The significant utilization of oddity or anomaly location is misrepresentation identification. Social insurance extortion prompts generous misfortunes of cash every year in numerous nations. Successful misrepresentation discovery is vital for decreasing the cost of Health mind framework. Extortion and manhandle on restorative cases turned into a noteworthy worry for medical coverage organizations a decades ago. Extortion includes purposeful misdirection or deception expected to bring about an unapproved advantage. It is stunning on the grounds that the occurrence of medical coverage extortion continues expanding each year. Information mining which is partitioned into two learning methods viz., managed and unsupervised is utilized to recognize false claims. Essentially arbitrary timberland calculation and coordinations relapse calculation systems are utilized for extortion identification in medical coverage. Information mining consequently sifting through colossal measures of information to discover known/obscure examples bring out significant new recognitions and make expectations [17].

Table 1: Comparative Study of Various Techniques

Author	Paper	About Paper	Dataset	Tools	Advantages	Accuracy	Gaps	Methods
Yuvaraj et al.	Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster	Proposes the novel usage of machine learning calculations in hadoop based groups for diabetes	Pima indian diabetes dataset	Hadoop	Enhance the accuracy of disease prediction	86.71%	Unwanted features affects the accuracy And Time consuming	Naïve Bayes, Decision tree, Random forest
Anooj et al.	Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules	proposed clinical decision support system for the risk prediction of heart patients	Cleveland data, Hungarian data, Switzerland data	MATLAB(7.10)	removing the missing values and other noisy information.	57.851%	Time consuming& really depends on medical experts	Automated approach for the generation of weighted fuzzy rules and developing a fuzzy rule-based decision support system
Emrana Kabir et al.	An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques	proposed system calculates and compares the accuracy of C4.5 and KNN	Pima indian diabetes dataset	WEKA	More active and accurate decision	90.43%	only works with numeric feature vector	Decision tree, KNN
Subramanian et al.	Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm	proposed CANFIS model combined the neural network adaptive capabilities and the fuzzy logic qualitative approach	Heart disease dataset	Neuro solution software	Affordable costs		Unwanted biases, errors which affects the quality of service	CANFIS, Genetic optimization
ShravanKumar et al.	Expert System Design to Predict Heart and Diabetes Diseases	To design an expert system that predicts the heart disease and diabetes disease with reduced number of attribute using data mining technique	Pima indian diabetes dataset and Heart disease dataset	Patterns	Reduced no. of parameters in very less time	83.63%, 85.96%	Extreme complications	Decision tree algorithm
Shouman et al.	Using data mining techniques in heart disease diagnosis and treatment	Determines the gap resulting from the previous theories of research	Cleveland heart disease dataset	Data mining tools	Provide reliable performance and to handle the error, complexity	84.14%, 89.01%	Received less attention	Naïve bayes, Decision tree, Neural network

Author	Paper	About Paper	Dataset	Tools	Advantages	Accuracy	Gaps	Methods
Thangarasu et al.	Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques	Predicts the diabetic disease from clinical database by using Neural Network algorithm s world	Cluster data set	Data mining tools	Reducing costs, reliability		Noise data sometimes leads to misleading results, complexity	Neural network algorithm, Fuzzy logic relations, Hybrid Genetic Algorithm and Data clustering algorithms
Anand et al.	Prediction of diabetes based on personal lifestyle indicators	Discourse on setting up a connection between diabetes hazard prone	Dehradun's clinics and university dataset	R studio	Variables are independent	75%	Response collected was less, avoid the over-fitting of data	CART method
Chitkara et al.	Voice based Detection of type 2 Diabetes Mellitus	Done research to recognize the vocal qualities of patients having type 2 diabetes mellitus	177 voice samples dataset	MDVP (multi-dimensional voice program)	Reduction in their values		Increase in spectral noise is due to the increase in value of NHR which can be due to variations in amplitude and frequency.	Time domain method, parameters like Jitter, Shimmer and NHR etc
MacDougall et al.	Integrating Health Information Technology into Clinical Guidelines	Brings the use of research based evidence into practice so as to develop clinical guidelines into practice	Institute of medicine	HER	increase accessibility, providing data directly into systems		Decrease time spent reviewing and asking about a patient's health history.	Updated PaJMa
Hang et al.	The Research and Implement of Diabetes Diagnosis Expert System	Acquainted how with found the diabetes analyze master framework		Expert system framework	reduce the rate of misdiagnose effectively		For not correct analysis i.e.useless	positive reasoning, reverse reasoning and two-way reasoning.
Sundar et al.	Development of a Data Clustering Algorithm for Predicting Heart	Finds out the accuracy of the result by using the K-means Clustering Algorithmic	Cleveland Heart Disease database	WEKA	simplicity and speed which allows it to run on large dataset	66.00%.	missing data, inconsistent data, and duplicate data	Decision Tree, Naïve Baye's , Neural Network and K-means
Srinivas et al.	Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks	Uses data mining application techniques in the Healthcare	Data collected from UCI repository	WEKA	reduce medical errors, improve the quality of service		lack of effective analysis tools to discover hidden relationships	Decision Tree algorithms,Neural Network ,neuro-Fuzzy

V. CONCLUSION

In this work, we have discussed about previous data processing techniques to retrieve information and current development in the research of medical sciences. Further we have elaborated terminologies and techniques of learning in data mining and machine learning. Literature survey is discussed to study about previous researches in this field of technology.

REFERENCES

[1] Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan, "*An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques*", International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 396-400, February 16-18, 2017.

[2] P. K. Anooj, "*Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules*," J. of King Saud Uni. Comput. and Inform. Sci., ELSEVIER, Vol. 24, pp. 27-40, 2012.

[3] Purushottam, K.Saxena and R. Sharma, "*Efficient Heart Disease Prediction System*," Proced. Comput. Sci., ELSEVIER, Vol. 85, pp. 962 – 969, 2016.

[4] Parthiban Latha and Subramanian R., "*Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm*" International Journal of Biological and Life Sciences 3:3 2008.

[5] Uppin ShravanKumar and M A Anusuya, "*Expert System design to predict Heart and Diabetes Diseases*", International Journal of Scientific Engineering and Technology Vol: 03, 2014.

[6] Shouman Mai, Tumer Tim, Stocker Rob, "*Using Data Mining Techniques in Heart Disease Diagnosis and Treatment*", International Conference on Electronics, Communications and Computers, 2012, IEEE, Northcott Drive, Canberra.

[7] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, "*Machine Learning and Data Mining Methods in Diabetes Research*",Comput Struct Biotechnol J. Vol. 15, pp. 104–116, 2017.

[8] Sheena Angra, Sachin Ahuja, "*Machine Learning and its Applications: A Review*", IEEE, International Conference On Big Data Analytics and computational Intelligence, pp. 57-60, 2017.

[9] N. Yuvaraj, K. R. SriPreethaa, "*Diabetes prediction in healthcare systems usingmachine learning algorithms on Hadoop cluster*", Springer, Cluster Computing, 2017.

[10] Mac Dougall Candice, Percival Jennifer and Mc Gregor Carolyu, "*Integrating Health Information Technology into Clinical Guidelines*", Annual International Conference of the IEEE, EMBS Minneapolis, Minnesota, USA, September 2-6, 2009.

[11] Srinivas K, Kavihta Rani B. and Dr. Govrdhan A., "*Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks*", International Journal on Computer Science and engineering Vol. 02, No. 02, pp. 250-255, 2010.

[12] Sundar V Bata and Tevi T, Saravanan N, "*Development of a Data Clustering Algorithm for Predicting Heart*", International Journal of Computer Applications(0975-888) Volume 48, No. 7, June 2012, Coimbatore, India.

[13] M Nirmala Devi, Balamurugan.S Appavu alias, U.V Swathi, "*An amalgam KNN to predict Diabetes Mellitus*", IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, Madurai, Tamil Nadu, India, 2013.

[14] Thangarasu Gunasekar and Assoc. Prof. Dr. Dominic P.D.D, "*Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques*", IEEE 978-1-4799-0059-6, Tronoh Perak, Malaysia, 2014.

[15] Ayush Anand, Divya Shakti, "*Prediction of Diabetes Based on Personal Lifestyle Indicators*", IEEE, International Conference on Next Generation Computing Technologies, pp. 673-676, 2015.

[16] Divya Chitkara, Dr. R.K. Sharma, "*Voice based Detection of type 2 Diabetes Mellitus*", IEEE, International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics, 2016.

[17] Namrata Ghuse, Pranali Pawar, Amol Potgantwar, "*An Improved Approch For Fraud Detection In Health Insurance Using Data Mining Techniques*", Int. J. Sc. Res. in Network Security and Communication, Vol. 5, Iss. 5, pp. 27-32, June 2017.

Authors Profile

Monika pursed Bachelor of technology from Punjab Technical University,India in year 2016. She is Currently pursuing M.Tech in Department of computer science and Engineering from Punjab Technical University main campus, kapurthala, india. Her main research work focuses on machine Learning Algorithms, Data Mining, Big Data Analytics.

Pooja Sharma pursed Master degree from Guru Nanak Dev University, Amritsar, India, She has Gold Medal by holding 1st position in Master Degree from Guru Nanak Dev University, Amritsar, India and Ph.D. from Punjabi University, Patiala, India, 2013. She is currently working as Assistant Professor in Department of Computer Science and engineering, IKGPTU Main Campus, Kapurthala (since June, 2017).She has published more than 15 research papers in reputed international journals including (SCI & Springers) and conferences including IEEE and it's also available online. Her main research work focuses Digital Image Processing, Computer Vision, Pattern Recognition, Image Retrieval, Image Reconstruction, Face Recognition. She has 7 years of teaching experience and 3 years of Research Experience.