Detailed Review on Data Mining Based Machine learning Techniques for Prediction of Diabetes Disease

Priyanka Tiwari [1], Mr. Varun Singh[2]

[1]*M.Tech Scholar,* [2]*Assistant Professor*

Department of Computer Science, Rewa Institute of Technology, Rewa, India

Abstract— In this paper, data mining methods used for diabetes data studies and disease forecasting were integrated.In this paper, the diabetic prediction problemit was well studied. The disease predictions were explored using different tools of DM. The use of diabetes prediction medical data collection was analyzed.For researchers, it would be a data mining major advantage for diabetes as it may reveal hidden details from a large amount of data related to diabetes.Different techniques of DMhelp research into diabetes and in the end improving the quality of diabetes patient care.This paper uses methods for DM focused on diabetic data set to perform a comprehensive survey of disease prediction.This paper presents a survey of different technologies for DM and the classification of neural networks used to predict the risk of diabetes disease based on risk factors.A person's risk level is defined using techniques such as K-Nearest Neighbor Algorithm, Decision Trees, Genetic Algorithm, Naive Bayes, etc. and the reliability is high when using more of the above technique's attributes and combinations.

Keywords—Data Mining, Diabetes Disease, Machine learning techniques, Prediction techniques.

I. INTRODUCTION

Data mining (DM)has played a significant role in the design of medical Expert programs for the identification and classification of advanced patterns.Medical expert systems is an active field of research in which machine learning (ML) data analysts and experts are constantly determined to compose themmore specific. Save the enhanced diagnostic systemsmedical practitioners ' time with better performance.Such devices also support physicians and surgeons in their daily routine [1]. Popular techniques of data mining used in nearly all sectors are described as:Support vector machine (SVM),Decision Tree, Knearest neighborhood (KNN),Bagging algorithm, Artificial Neural Network (ANN), etc.DM is an important step in the exploration of information in a database (KDD), iterative method of information cleaning, data addition, data collection, Recognition of patterns and knowledge identification in DM [2].

DM refers to knowledge extraction from large data amounts. This helps us to explore and examine the broad patterns in large datasets using statistical and artificial intelligence. The methodology of DM is used to forecast possible future developments or to uncover secret patterns in information behavior. Experts are widely used techniques like Artificial Neural Networks, Clustering, Decision Trees, Association rule algorithms, Classification, etc.[3].DMWas widely used in interaction, credit analysis, stock market forecasting,education, marketing, finance, health and medicine, risk prediction, data acquisition, scientific discovery, identification of fraud, etc.However, DM has animportant presence in all medical field Diagnoses for different diseases, such as kidney failure, diabetes, lung cancer, breast cancer, kidney stone, skin cancer, liver disease, diabetes, etc.DM applications include data analysis for better health Policymaking, hospital error prevention, early detection, and prevention deceptive coverage claims for different diseases, more value for money, cost savings and save more lives by increasing the death rate [4].

II. DIABETES DISEASE

The infection has taken place,however, based on the eating habits of humans. Many variables exist. That affectsthe occurrence of the disease.As the lifestyle of people has shifted, their dietary preferences have dramatically changed, with chunk diets, high protein foods, cigarettes,lack of exercise, and so on.It is possible to list a number of factors for the causes of diabetes diseases. In order to identify the existence of the disease, it is therefore important to track the changing conditions of blood sugar. Prediction of disease is the process of predicting or the probability of illness being diagnosed on the basis of signs and values.or the probability of illness is identified on the basis of signs and values glucose,Precision is also dependent on other factors characteristics.It depends onlifestyle,physical work, calorie use, and age.

There is a stronger correlation of diabetes with nephropathy, heart disease, including neuropathy, micro vascular, macro-vascular complications,and retinopathy.This leads to organ and tissue damage

thathappens in one-third of the population with diabetes.This allows the physician to classify prediabetes patients and perform a detailed study of their glucose sensitivity Insulin resistance to complications of the vascular system.The history of different diabetic patients to perform disease prediction. Such records are collected and named as a set of diabetic data from different medical organizations.The information in you can use the medical data collection to train the algorithm and a prediction process can be done by supplying the input set of values. On the other hand, the diabetic prediction has a variety of approaches.The process of collecting relevant information from a huge array of databases is DM. Such techniques ofDM can be adapted to the diabetes problem. likewise, there is numerous other scientific approaches to the prediction issue [5].

III. DATA MINING DIABETES DISEASEPREDICTION

Scientists in bioinformatics are commonly utilizing DM techniques. Bioinformatics is the science in which data from biological sequences and molecules is processed, organized, interpreted, extracted and used.In recent years, the methods of scientific discovery and data mining have been commonly used to derive Vast biological database trends. Biological information volume is increasing rapidly. Such data sets are evaluated involves inputting structure and generalizations from the data to make sense of the information.In the diagnosis of many diseases, the DM and bioinformatics interaction plays a fundamental role [3].

This paper analyzes and contrasts various data mining andMethods of machine learning in diabetes. Identification and prediction is part of the diagnosis and prognosis of the disease. Recent and popular DM clinical data techniques include Bayesian, SVM and decision tree,artificial neural network, random forest algorithms, etc. This provides the issues and observations of different factors about these techniques [6].

MATERIALS AND METHODS

It is possible to classify the Diabetic neuropathy methods and nephropathy prediction based on the methods and measures used.Originally, for the prediction, various DM algorithms can be used.Diabetes mellitus type 2 was predicted using k means clustering algorithm and Regression of logistics.A method was validated using the Pima Indian Diabetes dataset for its accuracy. The Waikato method was used to evaluate efficiency.Similarly, for diabetic prediction, the KNN and Naive Bayes algorithms are used in met combination [18] In Mapping attributes are used to assign data to objects from categorical and mathematical values of attributes and rules of the association are used for predicting disease.In any prediction of disease, the DM algorithms are of great importance. In, methods such as regression, ANN, GMM, and SVM, are evaluated for disease prediction for their results.In any prediction of disease, the DM algorithms are of great value. In methods such as regression,ANN, GMM,and SVM, are evaluated for disease prediction for their results.The method uses the CART algorithm to understand the normal habits of people including sleep, water, walking, MI. Likewise, The diabetic prediction was performed with DM techniques. The Type two diabetic Mellitus was the Enhanced k predicted means algorithm. Thevaluation is carried out with the data set of PIMA. A hybrid diabetic prediction model is described in as an extension.System k means diabetic prediction clustering and decision tree algorithm.An approach was tested using the data set of the UCI.

A. Fuzzy Based Prediction

The neural diabetic prediction has been presented using fuzzy logic in diabetic prediction.Within two steps, the prediction is performed. Gaussian kernel function was used to distribute the data in the first level.Second, data points were trained in the neural network and fuzzy logic was used to predict diabetes.Likewise, with a case-based approach, the Meth diabetic prediction has been approached with NN and Fuzzy logic.The method uses fuzzy logic,CBR, and NN, to predict the disease first.Therefore, the prediction is confirmed using a law-based approach. A fuzzy classifier has been introduced Methfor the association rule for diabetic prediction in which the system Generates rules of the association from a huge collection of data. Using data set rules are exploited and developed.rules created are used to predict diabetes risk. A performance of classificationhas been enhanced.

B. SVM Based Prediction

Support vector machine is an algorithm that can be applied to different classification issues.for machine learning.Diabetic neuropathy and nephropathy can be effectively classified on the basis of the SVM classifier. In this section, these methods are discussed. SVM with a diabetic prediction of Naive Bayes is presented in. The combined model uses the data collected from 400 plus patients to perform diabetic prediction. AnSVM is presented as a decision support system for diabetic prediction with a random forest ensemble approach.The formula lists the genes by the strength and the same was used to classify the genetic material that causes disease.

C. Genetic Algorithm Based Prediction

Genetic algorithm way to select a feature. Any classifier's performance depends heavily On the consideration of the feature. This section contains a list of the GA-based

prediction method for diabetes.The selection of features plays an important role in predicting disease. A genetic algorithm-based technique for the prediction of diabetic retinopathy has been proposed for development.The binary map of the ship was developed using the techniques of segmentation. SVM is used to isolate and identify morphological features.The genetic algorithm was used here for the choice of apps.A multifactorial genetic model for diabetic peripheral nephropathy prediction has been presented. However, in feature selection, the Method uses the standard variants to predict disease.

D. ANN Based Prediction

In order to perform effective classificationto identify hidden values, the neural networks are more efficient.Methods for diabetic prediction based on ANN are discussed here. A hybrid method of ANN and regression model in approached the issue of diabetic prediction.The approach was tested with different data sets and its output was calculated in different parameters. This method reduces the error rate[5].

IV. MACHINE-LEARNING METHOD TO DIABETES DISEASEPREDICTION TECHNIQUE

Machine learning is the field of science that deals with how Based on experience, machines learn. Researchersthe term "machine learning" is similar to the term "artificial intelligence" because learning is the key entity in the broadest sense of the word ability.Feature called smart. UseDM and ML and techniques in DM research is a vital approach to using large volumes of knowledge-based diabetes-related data from available resources.A particular disease's severe social impact makes DM one amongthe top priorities of medical science research, which inevitably leads to huge amounts of data. That is why,there is no doubt that DM approaches In terms of analysis, Management and other related aspects are of great concern to clinical administration, machine learning, and data mining.As a result, as part of this report, efforts have been Study current literature on machine learning and approaches to data mining in diabetes research [7].

A. Association Rule Learning

Training in association rules is a methodology that finds similar patterns in different databases and finds clear rules in the database. (I.e. regularly made observations). Understanding the law of connection helps in the process of decision making.Therefore, In the study of business baskets, internet usage mining, medical diagnosis where X and Y are sets of different items, rule learning is widely valuable.Association guidelines use a minimum aid level stated by the user and a minimum trust value defined by the user.

B. Classification

Classification is a way to find rulesseparate information in various group. The method of classification makes a set of similar observations from the large file.A classification implementing Algorithm is defined as a classifier, a mathematical method that implements an algorithm for classification.The major task of this methodology is to recognize and arrange similar observations in a selection from the large dataset. Classifications methods help find different trends for the large dataset.

C. Cluster Analysis

Clustering or research a practice that finds similar characteristics to group objects in the same group. It's a data grouping method.In many places, such as bioinformatics, machine learning, identification of patterns, etc., cluster analysis is widely used.Clustering is a segmentation like technique, grouping parallel observations. Using the rule of association, various observation patterns are created and then grouping techniques according to that pattern.Asimilar type of data is stored in the same groups in this approach and these groups are known as clusters. Artificial neural network and nearest neighbor searches are various techniques used in cluster analysis. Clustering helps to group items in sequence analysis.

D. Decision Tree

The decision making tree is a prediction, tree-like model. Each tree branch Refers to a condition and leaves, if the condition is met,denotes the result.With a question or a condition, each branch has been reported with two or more answers.Each answer can lead to a different question or situation. aDecision tree classifies data by state without a lack of information and assists in the decision making process.The decision-making tree can be used as a decision-making tool.

E. Neural network

The neural network can beused to identify and model different patterns. The neural network receives a set of inputs and is used to predict another output. The numerical output of the neural network.For fraud customer response prediction, detection,and much more, neural networks, image understanding, are widely useful. In outer analysis or clustering, neural networks also help.

F. Regression Analysis

Analysis of regression makes similar observations from the observations from different patterns.Analyzes the relationship Determines how the value of a dependent variable varies with the two variables independent variable's value change. It tests the independent variable's status. In prediction, this technique has been used.

Analysis of regression uses an independent variable function called the regression function.The analysis of regression is a type of systematic analysis of data.

G. Sequential Pattern Mining

The sequence is a set of prearranged transactions. The sequential mining pattern as the name implies finding patterns that are different from the datasets.The methodology identifies trends that happen over and over again. Sequential mining patterns help to identify current trends or related occurrences on a regular basis two styles of sequential mining patterns that are mining strings and mining objects. The algorithm recognizes a set of frequently occurring observations using different rules in the mining of items.Sequential pattern mining is commonly used in commercial applications. [8] is commonly used in sequential pattern mining.Literature Survey.

P. Sonar and K. JayaMalini [2019] In such cases, after consultation, To obtain their reports, the patient has to go to the clinic center.Because they need to invest their time and money at all times. But with the growth of machine learning approaches, we have the versatility to seek a solution to the current problem; we have advanced information processing system maltreatment capable of predicting whether or not the patient has a polygenic disease.In fact, sickness forecasting initially ends up supplying the patients before it becomes critical. The extraction of information has the ability to delete hidden details from a vast amount of information related to diabetes.The mean of this research is to increase a system that could more accurately predict a patient's diabetic risk level. The development of the model is based on methods of categorization, such as algorithms for Decision Tree, ANN, Naive Bayes and SVM [9].

M. F. Faruque et.al. [2019]This research work uses machine learning methods to explore Different risk factors associated with this disease. Techniques for learning machines provide effective learning outcomes by creating predictive models from clinical medical data sets obtained from diabetic patients to obtain information. It may be useful to predict diabetic patients to derive information from such results. We use four common machine learning algorithms in this study, namely Naive Bayes (NB), C4.5 Decision Tree (DT), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM),to predict diabetic Mellitus in adult population information.Our experimental results show that, relative to other machine learning methods, Higher accuracy was achieved in the C4.5 decision tree[10].

H. Abbaset.al. [2019] We review the San Antonio Heart Study data in this paper and use Using the computer to predict progression of type two diabetes in the future. Develop a prediction model, we use the aid vector machines than 10 Characteristics that are well known as strong predictors of future diabetes in the literature.We use 10-fold cross-validation for template learning and hold-out array to test it because of the unbalanced complexity of the dataset in terms of class labels.The results of this study indicate an 84.1 % validity reliability with an average 81.1% recall rate There are more than 100 iterations. The results of this study will help identify the high-risk population with type 2 diabetes in the future [11].

A. Aldallal and A. A. A. Al-Moosa [2018]. This work aims to develop application software to predict the incidence and recurrence of non-communicable diseases (NCDs) to be used by doctors and other medical practitioners.In this task, the predictive data mining method was applied. Bahrain Defense Force Hospital records of patients were used to test the proposed software application.This procedure was performed and checked at the specified hospital by the actual practitioner. The results showed that the prediction model can effectively, accurately and most importantly, immediately predict the diseases of NCDs.This software will help a doctor make the right decisions about the threats to patient health [12].

S. N. Singh and K. Kathuria [2018] This paper aim to find solutions to Decision trees to improve accuracy for the effective data processing technique. Diabetes is one of the chronic and fatal diseases caused by changes in insulin production and use in the body leading to high levels of blood sugar and some long-term complications. According to the 2017 IDF report, diabetes has affected 425 million adults worldwide. The effective diagnosis of diabetes can be made by the discovery of knowledge of available medical records.The conventional approaches focus only on data mining techniques but are lacking in proper data preparation and feature selection of attributes [13].

D. Duttaet.al. [2018] We should find out in this paper what are the critical elements for the cause of diabetes. Factor and feature selection have become the focal point of a lot of research in use regions for which tens or a large number of factors are available. Similarly, we will concentrate on the most essential features to predict a person's future chances of developing diabetes. Diabetes is an uprising disorder, especially due to the type of food we have these days and the contradictory eating scheme and schedule we follow.Diabetes is caused primarily by obesity or high levels of insulin, and so on. [14].

D. Shettyet.al. [2107] DM methodology's aim Is to think of a set of data and transform it into a logical framework for further use**.**Our review focuses on this aspect of the training development of clinical conclusion by gathering diabetes information and building an emotionally supportive network of smart therapeutic options to benefit

a physician**.A** primary objective of this study is to assemble the Intelligent Diabetes Disease Prediction Program,which predicts disease with diabetesthe database of the person with diabetes. In this method, we suggest using algorithms such as Bayesian and KNN (K-Nearest Neighbor) to apply and test the list of diabetes patients by taking different diabetes prediction attributes [15].

B. Alić et.al.[2017]This paper provides a description of machine learning approaches used by classification of diabetes and cardiovascular diseases (CVD), Artificial Neural Networks (ANNs) and Bayesian Networks (BNs).The study of comparison has been posted on selected articles between 2008 and 2017. Multilayer feed-forward neural network with Liebenberg Marquardt learning algorithm is the most frequently used type of ANN in selected papers.On the other hand, the Naive Bayesian network, which retrospectively showed the highest precision values for diabetes and CVD classification, is the most commonly used type of BN.99.51% and 97.92%.In addition, the Mean performance measurements of observed networks showed better results with ANN,indicating the higher chance of achieving more accurate results in diabetes or CVD classification is when applied to ANN[16].

B. D. Kanchan and M. M. Kishor [2016] A number of DM algorithm methods used to predict diabetic disease are addressed in this research paper.DM is a widely used methodin bioinformatics research by health organizations to classify diseases like diabetes and cancer.PCA study was conducted in this paper, which determines a minimum number of necessary attributes to improve the accuracy of different machine learning algorithms that are supervised.This work is aimed at researching supervised algorithms for machine learning to predict diseases of diabetes.DM has a number of key techniques such as pre-processing, categorization.Diabetic is a condition that affects health. that can be prevented in both urbanized and emerging countries such as India.The data categorization Consists of 1865 different attributes that are generated by gathering patient information database. Two types of blood tests, urine tests [17] are examples in the data set.

L. H. Anjaneya and M. S. Holi [2016]The risks of diabetes in children and adults have increased since the last decade. Different approaches to early detection and prevention of diabetes have been suggested.For identification, some approaches use EMG signals of diabetes,Because of the acceleration ofthroughout signal processing, EMG signals cannot effectively identify the signal.To solve this, we propose a new approach by considering EMG signals ' time domain and frequency domain features and performing the classification that Where using a neural network. Use the MATLAB software to execute this process 97.05% accuracy of the proposed approach is shown in the simulation study. [18].

V. CONCLUSION

In this paper, different methods of disease prediction were also reviewed and discussed the question of diabetes prediction.The methods are evaluated in different parameters of prediction for their results.The results obtained In a graph was plotted. In the future, an issue of prediction methods will be established and a new approachisconsidered to improve the prediction problem.The aim of DM methodology is to learn about information from a set of data and make it a logical framework to be used further. The design of medical conclusion learning through the collection of diabetes data and the creation of an emotionally supportive network of smart therapeutic choice to help the doctors. The primary goal is to build the Intelligent Diabetes Disease Prediction Program, which offers diabetes disease prediction using the registry of patients with diabetes. We analyze different techniques of machine learning by taking different attributes of diabetes for diabetes disease prediction.

REFERENCES

[1] Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., & Nawaz, R. (2017). *An expert system for diabetes prediction using autotuned multi-layer perceptron. 2017 Intelligent Systems Conference (IntelliSys).* DOI:10.1109/intellisys.2017.8324209

[2] Liu, W., Man, Z., Hua, L., Chen, A., Wang, Y., Qian, K., & Zhang, Y. (2014). *Data mining methods of lung cancer diagnosis by saliva tests using surface-enhanced Raman spectroscopy. 2014 7th International Conference on Biomedical Engineering and Informatics.* DOI:10.1109/bmei.2014.7002849.

[3] Shivakumar, B. L., &Alby, S. (2014). *A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes. 2014 International Conference on Intelligent Computing Applications.* DOI:10.1109/icica.2014.44.pp.1-7.

[4] Chauhan, D., &Jaiswal, V. (2016). *An efficient data mining classification approach for detecting lung cancer disease. 2016 International Conference on Communication and Electronics Systems (ICCES).*

[5] Baiju, B. V., &Aravindhar, D. J. (2019). *Disease Influence Measure Based Diabetic Prediction with Medical Data Set Using Data Mining. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT),* Pp.1-6. DOI:10.1109/iciict1.2019.8741452

[6] B. Senthil Kumar 1, Dr. R. Gunavathi2," A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 12, December 2016,pp.1-5.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., &Chouvarda, I. (2017). *Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116.* DOI:10.1016/j.csbj.2016.12.005.pp.1-46.

[8] Satyam Shukla+, DharmendraLal Gupta # and BakshiRohit Prasad*.” Comparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques”, International Journal of Database Theory and Application Vol.9, No.9 (2016), pp.107-118.

[9] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 367-371. DOI: 10.1109/ICCMC.2019.8819841.

[10] M. F. Faruque, Asaduzzaman and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, 2019, pp. 1-4. DOI: 10.1109/ECACE.2019.8679365.

[11] H. Abbas, L. Alic, M. Rios, M. Abdul-Ghani and K. Qaraqe, "Predicting Diabetes in Healthy Population through Machine Learning," *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, Cordoba, Spain, 2019, pp. 567-570. DOI: 10.1109/CBMS.2019.00117.

[12] Aldallal and A. A. A. Al-Moosa, "Using Data Mining Techniques to Predict Diabetes and Heart Diseases," *2018 4th International Conference on Frontiers of Signal Processing (ICFSP)*, Poitiers, 2018, pp. 150-154. DOI: 10.1109/ICFSP.2018.8552051.

[13] S. N. Singh and K. Kathuria, "Diabetes Diagnosis using different Data Preprocessing Techniques," *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2018, pp. 1-4.doi: 10.1109/CCAA.2018.8777332.

[14] D. Dutta, D. Paul, and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (ICON)*, Vancouver, BC, 2018, pp. 924-928. DOI: 10.1109/IEMCON.2018.8614871

[15] D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, 2017, pp. 1-5. DOI: 10.1109/ICIIECS.2017.8276012.

[16] B. Alić, L. Gurbeta, and A. Badnjević, "Machine learning techniques for classification of diabetes and cardiovascular diseases," *2017 6th Mediterranean Conference on Embedded Computing (MECO)*, Bar, 2017, pp. 1-4. DOI: 10.1109/MECO.2017.7977152.

[17] B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using the principle of component analysis," *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, Jalgaon, 2016, pp. 5-10. doi: 10.1109/ICGTSPICC.2016.7955260.