IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING IN HEALTHCARE FOR PREDICTING DIABETES AND HYPOXEMIA: A SURVEY

[1]B. Naga Lakshmi, [2]M. Robinson Joel

[1]Research Scholar, [2] Associate professor

[1] Department of Computer Applications, [2]Department of Computer Science

[1] Ponnaiyah Ramajayam Institute of Science and Technology(Deemed to be University), Thanjavur,India

[2] Ponnaiyah Ramajayam Institute of Science and Technology(Deemed to be University), Thanjavur,India

Abstract: One of the most important goals is to develop a medical diagnosis system for disease prediction. Machine learning techniques and technologies have been effectively applied in a variety of areas, including medical diagnostics. Machine learning algorithms may be extremely useful in developing a health system to address health-related issues, such as assisting doctors in diagnosing diseases at an early stage. Diabetes is a chronic disease marked by high blood sugar levels. It has the potential to induce a variety of serious illnesses, including stroke, kidney failure, and heart attacks. When our blood oxygen level falls below the normal range, we may experience hypoxemia symptoms. It could result in a slew of complications. Heart disease, heart attack, stroke, neuropathy, nephropathy, retinopathy and vision loss, hearing loss, and so on are all symptoms of heart disease, heart attack, or stroke. This research focuses on a review of machine learning techniques for predicting diabetes and hypoxemia disease.

Index Terms **- Machine Learning algorithms, Diabetes, Hypoxemia, Healthcare, Disease.**

I. INTRODUCTION

Monitoring raised glucose levels in blood for expanded duration of time cause to a metabolic disorder called diabetes. Diabetes is characterized by insulin inadequacy. This can be either due to the body's weakness to produce enough insulin (Type 1 Diabetes) or the inefficiency to use the insulin produced for the metabolic process (Type 2 Diabetes). Such defective metabolism results increase of glucose in the body thus leads to raised Blood Glucose Levels (BGL). The untreated Diabetes chronic disease leads to associated problems of the heart (cardiovascular diseases), eyes (retinopathy pigmentation), legs (resulting in amputation), kidneys, stroke and such other complications. Recently diabetes has no cure, yet regular monitoring and maintaining blood glucose levels assure control over adverse effects and complications associated with this disease.Maintaining the precise balance of oxygen saturated blood should be important to our health. When our blood oxygen level goes outside the typical range, we may begin experiencing the symptoms of Hypoxemia like Shortness of breath, Coughing or wheezing, Headache, Rapid heartbeat. A blue coloration to the skin, lips and fingernails. The untreated Diabetes chronic disease leads to associated problems of the Heart disease, heart attack, stroke, Neuropathy, Nephropathy, Retinopathy and vision loss, Hearing loss etc.

That is why it is so important to know the warning symptoms to look for and to monitor a health care provider regularly for routine wellness screenings. Computer Aided Diagnosis System is a rapidly growing dynamic area of research in medical industry. The recent researchers in machine learning machine learning guarantee the improved accuracy of perception and diagnosis of disease. The Computer Aided Diagnosis System is able to think by developing intelligence by learning. There are many types of Machine Learning Techniques are available to classify the datasets. Supervised, Unsupervised, Semi-Supervised, Reinforcement, Evolutionary learning, and deep learning algorithms are used to classify the datasets. This research focuses on developing a diabetes prediction model using machine learning algorithms.

The following is a breakdown of the paper's structure:

Section II provides an overview of previous research on diabetes prediction and machine learning techniques.

The approach and machine learning algorithms are presented in Section III.

The Conclusion and References are found in Section IV.

II. RELATED WORKS

Madhusmita Rout et al. [1] Proposed a classification model used with machine learning techniques to enhance the prediction accuracy of predict diabetes. Classification and clustering methods are applied to build the predictive model. For this study, the PIMA Indian dataset is collected from the UCI Machine Learning Repository which contains a record of 769 patients with nine attributes. SVM, KNN, K-means, Random Forest, Regression,Decision Tree, Outlier Detection, Rule mining algorithms were used.

Md. Kamrul Hasan et al.[2] proposed framework for the diabetes Prediction. This framework has used the proposed ensemble model from the PID dataset. KNN, DT, RF, AB, NB, XB machine learning algorithms were used. The algorithm predict the similar level accuracy.

Sisodia et al. [3] discuss predicting the diabetes disease using three classifiers name as such as Naïve Bayes (NB), Support vector machine (SVM), and Decision tree (DT). An experiment has performed on the Pima Indian Diabetes Database (PIDD). The performance metric has measured in the term of Precision, Accuracy, Recall, and F-Score. The Results obtained show Naïve Bayes outperformed among the three algorithms with 76.30 % Accuracy.

Alkaragole et al. [4] have used a ten years record of a dataset from 130 US clinics which consist of 1000 patients records and 9 attributes from the UCI standard dataset to propose a hybrid framework. They have compared and analyzed different data mining techniques using the apache server. SVM and Decision Tree algorithms have used to build the proposed model which achieved an accuracy of 94%.

They done a Comparative Result between Proposed hybrid algorithm and other algorithms are shown as follows

Classifier	Sensitivity	Specificity	Accuracy
Decision Tree	83%	83%	86%
Naïve Bayes	86%	82%	90%
SVM	86%	85%	91%
Proposed Ensemble SVM+ Decision Tree (iteration=100)	91%	91%	94%

Table 1: Classifier Analysis

Saru et al. [5] have Compared different classifiers and accuracy for better prediction using Pima Indian dataset by building a predictive model. WEKA software has used to implement the Decision Tree, Naive Bayes, and KNN algorithms. The bootstrapping technique was used to enhance the accuracy rates. The proposed ensemble method obtained an accuracy level of 94.44%.

Yahyaoui et al. [6] proposed a medical Decision Support System (DSS) for predict and detect diabetic patients based on SVM and the Random Forest (RF) and compared conventional machine learning with a Fully Convolutional Neural Network (CNN). They have perceived from their method that RF was more efficient for predicting diabetes which yielded accuracy to be 83.67%.

Ayman Mir et al. [7] performed an analysis to predict diabetes disease using Machine Learning techniques on big data of healthcare. They have used several algorithms such as Naive Bayes, Support Vector Machine (SVM), Random Forest and Simple CART. The accuracy for Nave Bayies was 77%, SVM was 79.13%, RF was 76.5%, and Simple CART was 76.5%

Yichuan Wang et. al [8] had proposed a data analytics structure for the healthcare sector, with the help of this data analytics structure , the five big data analytics entities like Pattern's Analysis, unstructured Data Analysis, Decision Support, Predictive and traceability were identified.

Aeshah Saad Alanazi et al [9] proposed the model which combines two algorithms of machine learning algorithms namely Support Vector Machine and Random Forest to predict the diabetes. They have used a real dataset collected from Security Force Primary Health Care. The proposed model has achieved 98% of accuracy, ROC 99%. The result shows that the Random Forest algorithm has better accuracy score when compared with Support Vector Machine.

Md. Tanvir Islam et al. [10] have used three popular Machine Learning algorithms called AdaBoost, Bagging and Random Forest. They have collected real time information of both diabetic and non-diabetic people. The dataset has consisted 464 instances with 22 unique risk factors. They trained and test the algorithms. They used AdaBoost, Bagging, Bagging algorithms . Among the three algorithms, AdaBoost gave 97.84% accuracy, Bagging gave 98.28% accuracy and Random Forest gave 99.35% accuracy with respect to predict diabetes disease precisely.

Not yet prediction system has developed for Hypoxemia.

The table2 exhibits, for predicting diabetics a list of various algorithms with accuracy.

s.no	Algorithm	Accuracy	Year
1	SVM, KNN, K-means, Random Forest, Regression, Decision Tree	80%	2020
2	KNN,DT,RF,AB,NB,XB	81%	2020
3	Naïve Bayes (NB), Support vector machine (SVM), and Decision tree (DT)	76.30%	2018
4	SVM and Decision Tree	79%	2019
5	Decision Tree, Naive Bayes, and KNN	76%	2019
6	SVM , Random Forest (RF) ,CNN	83.60%	2019
7	Naive Bayes, SVM, Random Forest ,Simple CART	77% 79.13% 76.5%	2018
8	Pattern's Analysis, unstructured Data Anal	81%	2018
9	Support Vector Machine and Random Forest	80%	2020

Table 2: Classifier Algorithm Comparison

III. METHODOLOGY

As we mentioned in the introduction section, the machine learning techniques are used to predict diabetes. Both the classification and clustering methods are applied to build the predictive model. SVM, KNN, K-means, Random Forest, Regression, Decision Tree, Outlier Detection, Rule mining algorithms are used for prediction.

3.1 Data Pre-processing

The noisy or missing value which is responsible for the performance degrade of a dataset, we use data pre-processing method for cleaning a dataset, normalized values are obtained. Data pre-processing involves dimensional reduction, data cleaning, normalization, data transformation, feature extraction for of dataset, etc... The final result is structured from unstructured dataset.

3.2 Data Tool

Many tools like Anaconda, WEKA, Rstudio, MatLab, Orange, Knime, Apache, etc... are already available to carry out this experiment. Anaconda distribution is one of the open-source tool and available for different OS platforms. Data visualization should be very important as it needs to clear for presenting the gaps of previous results with the experimental results. so a clear analysis should be done for further researches.

3.3 Algorithms and Techniques

A supervised learning approach can performed on any type of data values. Classification can classify the new data. This technique helps to identifying the class labels where new data can be fit. The different labels and groups can be formed based on similarity from the data set by the clustering method.

Artificial neural network (ANN)- Complex and non-linear relationships can be easily carried out using ANN. ANN is a large connection of neurons consist the input, hidden and output layers.

Examples - MLP, CNN, RNN, FNN, RBFNN.

3.3.1 Multilayer perceptron (MLP) - It's every node uses activation function and also it uses a back propagation method for training and minimizes errors.

3.3.2 Convolutional Neural Networks (CNN) - It easily detects the important features and filters the input. K-means- It helps find and form clusters that help data to be grouped based on the mean calculation of each cluster. Each data point can be labeled so the statistics is useful to explore the various patterns for identification.

3.3.3 Genetic Algorithm - It is used to find the quality results by a better optimization for complex and large data sets. It uses three operators such as mutations, crossover, and selection together for obtaining an enhanced solution for a given problem.

3.3.4 Decision Tree - It can be applied to numerical and categorical data. It has pros like easier visualization and understands. But sometimes due to a variation in data, it will generate the decision trees which could be very complex and unstable and leads to over-fitting. It should be overcome with bagging and boosting algorithms.

3.3.5 Logistic regression (LR) - The outcome of LR is specific to certain events. Most of the medical cases, it is used to determine the severe state of a patient using logistic regression models. LR can be binomial, ordinal or multi-nominal. LR can only work for the binary variables i.e. the result will be '0' or '1'.

3.3.6 Support vector machine (SVM) - It works better for high dimensional spaces but the estimation about the probability is not direct. So further fold-cross validation may be calculated. The outcome should understand by the hyper-plane that should differentiate the given classes very clearly.

3.3.7 K-Nearest Neighbors (KNN) - It's a non- parametric and lazy learning algorithm because it does not learn during the training phase. Easy implementation on a large dataset. The variable k value depends upon the data i.e. if the variable k value is large, the noise effect can be reduced. But the computation time should be high because of each query instance will be calculated from all the trained samples present in a dataset.

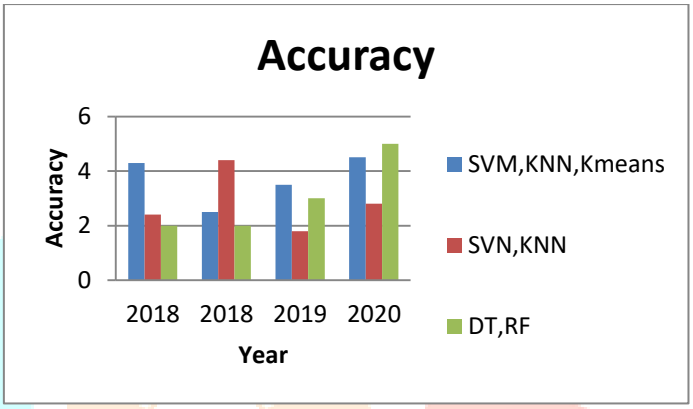

Fig 1: Comparison graph

IV. CONCLUSIONS

The use of machine learning is also increased rapidly as they are useful for handling huge amounts of data in healthcare. The aim of this work was to use machine learning to make easy practitioner and machine learning researchers in healthcare. We have done a Literature Survey on best algorithms from all the list of papers. We have focused on the algorithms which are used and their accuracy level. This study will reduce the exhaust of the practitioner and researcher to choose the best algorithm while predicting the diseases like Diabetes . There is no system developed to predict Hypoxemia. The various algorithms of machine learning such as Support vector machine (SVM), Decision tree (DT),naive bayes, AdaBoost, Bagging, Bagging algorithms, random forest, Convolutional Neural Network (CNN), KNN algorithm and logistic regression algorithm are reviewed to predict diabetes

References

[1]. Madhusmita Rout and Amandeep Kaur , Prediction of Diabetes Risk based on Machine Learning Techniques in 2020 International Conference on Intelligent Engineering and Management (ICIEM).

[2] Md.Kamrul Hasan, Md.Ashraful Alam, Dola Das, Eklas Hossain (Senior Member, Ieee), Mahmudul Hasan Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, IEEE Access 2020.

[3] Sisodia, D., Sisodia, D.S., "Prediction of Diabet es using Classification Algorithms," International Conference on Computational Intelligence and Data Science (ICCIDS 2018),ELSEVIER. Procedia Computer Science, ISSN 1877-0509,vol132.

[4] Alkaragole, M. L. Z., & Kurnaz, A. P. S. (2019). Comparison Of Data Mining Techniques For Predicting Diabetes Or Prediabetes By Risk Factors.

[5] Saru, S., & Subashree, S. (2019). Analysis and Prediction of Diabetes Using Machine Learning. International Journal of Emerging Technology and Innovative Engineering.

[6]. A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK),Ankara,Turkey,2019.

[7]. A. Mir and S. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare", in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018.

[8]. Wang, Y., and Kung, L. A., Terry Ant hony Byrd, "Understanding its capabilities and potenial benefits for healthcare organizations". Journal of Technological Forecasting and Social 126:3–13, 2018.

[9]. Aeshah Saad Alanazi and Mohd A. Mezher , "Using Machine Learning Algorithms For Prediction Of Diabetes Mellitus", International Conference on Computing and Information Technology, University of Tabuk, Kingdom of Saudi Arabia. Volume: 02, Issue: ICCIT- 1441, Page No.: 55 - 57, 9th & 10th Sep. 2020.

[10]. Md. Tanvir Islam, M. Raihan, Nasrin Aktar, Md. Shahabub Alam4, Romana Rahman Ema5 and Tajul Islam ,"Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches" ,11th ICCCNT 2020.