2021 7th International Conference on Advanced Computing & Communication Systems (ICACCS)

Supervised and Unsupervised Machine Learning based Review on Diabetes Care

Tannu Chauhan
BCA.Delhi Technical Campus
Greater Noida, India
tannuchauhan0801@gmail.com

Surbhi Rawat
BCA.Delhi Technical Campus
Greater Noida, India
surbhi.rawat109@gmail.com

Samrath Malik
BCA.Delhi Technical Campus
Greater Noida, India
samarth.malik21@gmail.com

Pushpa Singh
CSE. Delhi Technical Campus
Greater Noida, India
pushpa.gla@gmail.com

2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) | 978-1-6654-0521-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICACCS51430.2021.9442021

Abstract*— Sedentary lifestyle, poor diet and work pressure lead the diabetes disease which may cause several fatal health issues like heart attack, strokes, kidney failure, nerve damage etc. Diabetes can be effectively managed when caught early with high accuracy. Machine Learning (ML) approaches are very effective to early detection and prediction of diabetes. The goal of this paper is to offer the inclusive examination of the diagnosis of diabetes by supervised and unsupervised ML algorithms. This survey includes papers on the diagnosis of diabetes from 2018-2020. Decision tree based algorithm such as C4.5, AdaBoost, XGBoost, etc., have predicted the diabetes with high accuracy. Unsupervised learning techniques such as PCA and K-Mean are also useful in the attribute selection and outlier detection from the large dataset. This study reveals that K-Mean and SVM have also diagnosed and evaluated diabetes by high accuracy as an amalgamation of supervised and unsupervised machine learning techniques.

***Keywords*—** *Machine Learning, Supervised, Unsupervised, Diabetes, Decision Tree, Prediction*

I. INTRODUCTION

Diabetes is one of the most dangerous chronic diseases that could lead to others serious complicating diseases. Diabetes diseases are also called as diabetes mellitus, which describes a set of metabolic disease. Diabetes leads to many other kinds of diseases and that are- heart attack, blindness, kidney diseases and so on. Diabetes is also called as Diabetes Mellitus is a chronic disease and is considered as one of the deadliest diseases.

Diabetes disease can be categorized as Type 1 or Type 2. If the pancreas does not create adequate amount of insulin in body, is called as Type 1. In Type 2, the body either cannot effectively use the insulin that it produces or an inadequate amount of insulin is released into the bloodstream [1]. Type 1 disease generally occurs in children and adolescents, but it can occur in older people. Type 2 diabetes is normally milder compare to type 1 and 90 % people have type 2 diabetes. Type 1 diabetes can be cured by inserting insulin into the fatty tissue under the skin of patient. However, Type 2 diabetes can be cured by having a healthy diet, weight and exercising.

Many of diseases can be prevented if diabetes can be diagnosed in the early stages. Early diagnosis and prediction of disease is possible due to recent technological development of IoT, Artificial Intelligence (AI) and Block chain in the current healthcare system [2]. AI presented a paradigm shift in diabetes care from conservative management approaches to construct the targeted data-driven precision care [3]. IoT offers connected environment to the smart healthcare system [4]. ML and deep learning are AI based techniques. ML has a potential of improving efficiency and decrease the cost of treatment in the healthcare system. Various texts are available for diagnosis and prediction of diabetes based on data mining and ML. Data mining and ML methods are equally important to their specific objective. Data mining techniques are useful to extract rules and pattern from the vast amount of diabetes data set, while ML is significant to learn and automate the machine along with pattern recognition. Several ML techniques are used to form digital support in diabetes care. These include support vector machine (SVM), Decision Tree (DT), random forest (RF), classification and regression trees, Logistic Regression (LR) k-nearest neighbor (KNN), neural network, K-Mean, Principle Component Analysis (PCA) based algorithm for better diabetes care. Various texts have been available for automatic diabetes detection, prediction and management via ML and AI [5].

In this paper, we will review the several ML techniques for diabetes detection and prediction. There are mainly two categories of learning i.e. supervised and unsupervised learning that made foremost impacts in the detection, prediction and treatment of diabetes. This literature survey

978-1-6654-0521-8/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: Malta College of ArtsScience&Tech. Downloaded on September 30,2021 at 12:57:19 UTC from IEEE Xplore. Restrictions apply.

firstly focused on key words associated to the supervised and secondly on unsupervised ML techniques mainly from 2018 to 2020.

The remainder of the paper is arranged as follows. Section 2 and section 3 represent supervised and unsupervised ML techniques respectively to the analysis, diagnosis, classification and prediction of diabetes disease. Section 4 deliberated the findings of the review as a part of result and discussion. Lastly, section 5 concludes the paper.

II. SUPERVISED LEARNING TECHNIQUES ON DIABETES DISEASE

Supervised learning algorithms take direct feedback for the prediction. Supervised learning can be categorized in classification and regression methods. KNN, DT, SVM, LR, Artificial Neural Network (ANN), Naïve Bayes (NB) etc., are some popular algorithm of supervised learning. The basic objective of classification techniques is to detect and predict of the possibility of diabetes in patients with maximum accuracy.

National Institute of Diabetes and Digestive Kidney Disease dataset and many techniques like Data transformation, Association rule mining is also used in [6]. In this study, ANN, RF, K-means clustering techniques are used to predict diabetes with maximum accuracy. Artificial neural network outperformed with highest accuracy of 75.7%. The classification algorithms were used to predict diabetes with maximum accuracy by applying various ML algorithms such as SVM, NB Classifier, and DT. Experiments were implemented on PIDD (Pima Indian Diabetes Data Set) database [7] and Naïve Bayes has gain highest accuracy i.e. 76.30%. Reference [8] proposed ensemble method that offers 90.36% accuracy on PIDD dataset. An early detection method for diabetes detection was suggested by using optimal feature selection. DT, RF, and NB were applied and highest accuracy had achieved as 82.3% by the NB classifier [9].

Author used PIDD database and their proposed diabetics' prediction system was based on two stages data preparation and classification. Tigga & Garg [10] applied ML methods for the detection of diabetes. The aim of their research work was to predict the risk of diabetes among individual with maximum accuracy depending on their family background and lifestyle. Numerous ML algorithms were used like: LR Method, KNN Classifier, SVM, NB Classification Method, DT Classification, RF Classification. Data was collected through a questionnaire on which experiments were performed and same experiments were performed on the PIDD database. RF Classification Method has highest accuracy i.e. 94.10% as compared to other algorithms. The Linear Discriminant Analysis (LDA), RF, KNN, NB, J48 and SVM were used for classification. Author in [11] proposed predictive models using Logistic Regression and Gradient Boosting Machine (GBM) methods on Canadian patients aged between 18-90 years to recognize patients with high risk of evolving diabetes.

Heart rate variability (HRV) signal had taken from ECG could be efficiently applied to the non-invasive detection of diabetes. A HRV signal based on deep learning architecture was proposed for the classification of diabetes. Author applied long short-term memory (LSTM), convolutional neural network (CNN) and its variation to retrieve temporal dynamic features and these features have been applied with SVM for the classification [12]. Various ML algorithms were implemented on Weka tool and found that the sequential minimal optimization (SMO) algorithm was provided better performance compared to other supervised learning algorithm [13]. It has observed that paradigm is shifted to ML to deep learning [5]. The following papers have been studied related to supervised learning and represented in table 1.

TABLE I
SUMMARY OF MAJOR FINDINGS OF DIABETES PREDICTION USING SUPERVISED LEARNING

S. No.	Ref. No.	Findings	Best Algorithm
1	[14]	85% algorithms were related to supervising learning and 15% were related to unsupervised algorithms. It was witnessed that SVM was most usually used classification algorithm of the diabetes.	SVM
2.	[15]	Various supervised ML techniques were compared to reveal that which algorithm is appropriate for the prediction of diabetes.	SVM
3.	[7]	Three classification algorithms such as DT, SVM and NB were applied to identify diabetes at an early stage.	NB
4.	[9]	Author applied DT, RF, and NB.	NB
5.	[16]	Estimated the future diabetes risk by using Gradient Boosting, LR and NB.	Gradient Boosting.
6.	[17]	Comprised some external features that were accountable for diabetes along with regular features like Glucose, BMI, Age, Insulin, etc. Classification accuracy is enhanced with new dataset.	AdaBoost with 98.8% accuracy

Authorized licensed use limited to: Malta College of ArtsScience&Tech. Downloaded on September 30,2021 at 12:57:19 UTC from IEEE Xplore. Restrictions apply.

7.	[18]	Author applied four popular ML algorithms SVM, NB, KNN, C 4.5 DT on adult population data to predict diabetic mellitus.	C4.5 DT
8.	[11]	Predictive models using LR and techniques on Canadian patients aged between 18-90 years to recognize patients with high risk of evolving diabetes.	LR & GBM
9.	[10]	Author collected the 952 responses online and offline survey having 18 questions associated to health, lifestyle and family background for the prediction of Type 2 diabetes.	RF with highest accuracy of 94.10%
10.	[19]	Early detection of type 2 diabetes based on multivariate regression methods like Glmnet, RF, XGBoost, LightGBM.	XGBoost with Accuracy 88.1%
11.	[13]	Diabetic Patient was monitored using NB, SMO, J48, ZeroR, OneR, simple logistic and RF.	SMO
12.	[12]	Suggested LSTM, CNN and its mixtures for extracting complex temporal dynamic features of the input HRV data. These features were applied with SVM for classification.	Moving ML to Deep Learning
13.	[5]	Review various papers for feature selections, different datasets, classification algorithms such as KNN, SVM, Random Forest, etc., Deep learning approach.	Moving ML to Deep Learning

III. UNSUPERVISED LEARNING TECHNIQUES ON DIABETES DISEASE

Unsupervised learning algorithms do not take any feedback for the prediction. This learning finds the hidden patterns in data. Two simple concepts i.e. principal component analysis (PCA) and cluster analysis are used in unsupervised ML. PCAs eliminate extremely associated features by using covariance matrix, eigenvalues and eigenvectors. K-Mean, Self Organized Model (SOM), PCA and LDA etc. are some well-known unsupervised learning algorithm. The K-means clustering algorithm is implemented as pre-processing steps and outlier detection.

The K-means clustering algorithm first employed to discover and delete outliers in diabetes data set and then classification algorithm SVM was applied [26]. K-means algorithm was used to categorize the patients into Healthy and Diabetic clusters. In this work, healthcare dataset of pregnant women from healthcare is taken to build a predictive diabetics model. K-Means algorithm is found to be 78% accurate [21].

Further, due to having large dimension of diabetes dataset, it is significant to identify principle components or attributes that are participating in the detection and prediction of diabetes. Limited text are available that applied unsupervised learning for the prediction of diabetes as shown in table 2.

TABLE II
SUMMARY OF MAJOR FINDINGS OF DIABETES PREDICTION USING UNSUPERVISED LEARNING

S. No.	Ref. No.	Findings	Best Algorithm
1.	[21]	85% Clustering methods and 15% non-clustering methods were used to build a predictive model for diabetes	k-means with 78% accuracy
2.	[22]	K-Means and Hierarchical Clustering methods for diabetes prediction	K-Means

Generally, supervised learning is used for the classification and prediction of the diabetes and unsupervised learning is used as a pre-processing. Pre-processing is a way that prepare the raw data and building it appropriate for a ML model.

IV. RESULT & DISCUSSION

Diabetes Mellitus is one of the serious diseases. Age, obesity, sedentary life style, hereditary diabetes, living style, poor diet, high blood pressure, etc. are the main reason of diabetes. From the table 1, it has been observed that decision tree or variation of decision tree such as XGBoost, AdaBoost and RF are most widely used classification algorithm of the supervised learning. Trend is shifting from ML to deep learning. ANN is recently very popular ML methods, which perform well in many aspects [5] [12].

Unsupervised learning methods such as PCA, LDA are mainly applied for the dimensionality reduction. Unnecessary attributes in diabetes data set are misleading the accuracy of classifier. Hence, we can have combination of supervised and unsupervised learning for the better prediction and detection of diabetes. Zhu et al. in [23] integrated PCA, K-Mean and logistic regression. PCA was applied for dimensionality reduction, k-means for clustering, and LR for classification.

Authorized licensed use limited to: Malta College of ArtsScience&Tech. Downloaded on September 30,2021 at 12:57:19 UTC from IEEE Xplore. Restrictions apply.

TABLE III
SUMMARY OF MAJOR FINDINGS OF DIABETES PREDICTION USING SUPERVISED & UNSUPERVISED LEARNING

S. No.	Ref. No.	Findings	Best Algorithm
1.	[6]	Apriori method has been used to establish a strong relationship of diabetes BMI and blood glucose level. ANN, RF and K-means clustering methods were applied for the prediction of diabetes	ANN with accuracy of 75.7%
2.	[20]	K-Mean was applied for outlier detection and then SVM was applied for the classification	K-Mean & SVM
3.	[23]	PCA,K-Means and LR algorithm. PCA boosted the K-Means.	PCA, K-Mean, And logistic regression
4.	[24]	Author compared SVM and k-means & SVM on PID data set f	K-Mean & SVM with 99.64 % accuracy.
5.	[25]	A survey report on SVM technology using a different kernel and K-mean in order to diagnose the SVM.	K-Mean & SVM

Figure 1 represents the available literature based on supervised, unsupervised and both on diabetic prediction and care. From the table 1, 2 and 3, it has been observed that supervised learning has utilized 65%, unsupervised learning has applied 10% and both have applied 25%.

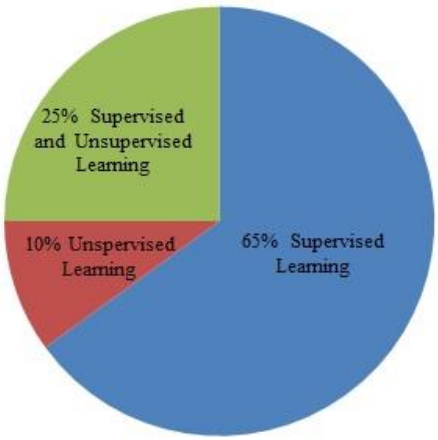

Fig. 1: Paper distribution for diabetes prediction & Care

This work can be extended with Blockchain and IoT in order to transmit the prediction in a transparent and secure manner to doctor, patient, diagnostic lab, etc.

V. CONCLUSION

In this paper we surveyed studied the current state of-the-art in order to predict and detect diabetes disease. Diabetes is a chronic disease and must be diagnosed earlier before it could reach a dangerous state. Various Supervised learning algorithms applied such as Naïve Bayes, SVM, Decision tree etc. As a conclusion, decision tree based classifiers have the potential to detect the diabetes in early stage. It is clear that the model improves accuracy and precision of diabetes prediction when combined with an unsupervised learning method such as PCA and K-Mean. K-Mean and SVM have also diagnosed and evaluated diabetes based on accuracy. Deep Learning based algorithm such as ANN, CNN etc. are also performing well in order to obtain better results in diabetes care system.

References

[1] V. Agrawal, P. Singh, and S. Sneha, "Hyperglycemia Prediction Using ML: A Probabilistic Approach," In International Conference on Advances in Computing and Data Sciences, vol. 1046, pp. 304-312, April 2019, Springer, Singapore.
[2] P. Singh and N. Singh, "Blockchain with IoT and AI: A Review of Agriculture and Healthcare," International Journal of Applied Evolutionary Computation (IJAEC), vol. 11, pp. 13-27, 2020.
[3] S. Ellahham, "Artificial Intelligence in Diabetes Care ", The American Journal of Medicine, 2020.
[4] P. Singh, and R. Agrawal, "A customer centric best connected channel model for heterogeneous and IoT networks. Journal of Organizational and End User Computing (JOEUC), vol. 30, pp. 32-50, 2018.
[5] J. Chaki, S. T. Ganesh, S. K. Cidham and S. A. Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," Journal of King Saud University-Computer and Information Sciences, 2020.
[6] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig and Z. Abbas, "A model for early prediction of diabetes," Informatics in Medicine Unlocked, 16, 100204, 2019.
[7] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia computer science, vol. 132, pp. 1578-1585, 2018.
[8] M. Alehegn, R. Joshi and P. Mulay, "Analysis and prediction of diabetes mellitus using machine learning algorithm," International Journal of Pure and Applied Mathematics, vol. 118, pp. 871-878, 2018.
[9] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," Journal of Big data, vol. 6, pp. 13,2019.
[10] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," Procedia Computer Science, vol. 167, pp. 706-716, 2020.
[11] H. Lai, H. Huang, K. Keshavjee, A. Guergachi and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques,"BMC endocrine disorders, vol. 19, pp. 1-9, 2019.
[12] G. Swapna, R. Vinayakumar and K. P. Soman, "Diabetes detection using deep learning algorithms," ICT Express, vol. 4, pp. 243-246,2018.
[13] A. Rghioui, J. Lloret, S. Sendra and A. Oumnad, "A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms In Healthcare," Multidisciplinary Digital Publishing Institute, vol. 8, pp. 348, September 2020.
[14] I. Kavakiotis, O. Tsave, , A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda, "Machine learning and data mining methods in diabetes

Authorized licensed use limited to: Malta College of ArtsScience&Tech. Downloaded on September 30,2021 at 12:57:19 UTC from IEEE Xplore. Restrictions apply.

research," Computational and structural biotechnology journal, vol. 15, pp. 104-116, 2017.
[15] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," In 2018 24th International Conference on Automation and Computing (ICAC), pp. 1-6. IEEE, September 2018.
[16] R. Birjais, A. K. Mourya, R. Chauhan, & H. Kaur, "Prediction and diagnosis of future diabetes risk using Machine Learning Approach". SN Applied Sciences, vol. 1, 1112, 2019.
[17] A. Mujumdar and V. Vaidehi, "Diabetes prediction using Machine Learning Algorithms". Procedia Computer Science, vol. 165, pp. 292-299, 2019.
[18] M. F. Faruque and I. H. Sarker, "Performance analysis of Machine Learning Techniques to predict diabetes mellitus". In International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp. 1-4, February 2019.
[19] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh and G. Stiglic, "Early detection of type 2 diabetes mellitus using Machine Learning-based prediction models," Scientific reports, vol. 10, pp. 1-12, 2020.
[20] M. Alirezaei, S. T. A. Niaki and S. A. A. Niaki, "A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using Support Vector Machines," Expert Systems with Applications, vol. 127, pp. 47-57, 2019.
[21] D. C. Sujatha, D. M. Kumar and M. C. Peter, "Building predictive model for diabetics data using K Means Algorithm," International Journal of Management, IT and Engineering, vol. 8, pp. 58-65, 2018.
[22] M. Raihan, M. T. Islam, F. Farzana, M. G. M. Raju and H. S. Mondal, "An Empirical Study to Predict Diabetes Mellitus using K-Means and Hierarchical Clustering Techniques," In 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) , pp. 1-6, IEEE, July 2019.
[23] C. Zhu, C. U. Idemudia and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Informatics in Medicine Unlocked, 17, 100179, 2019.
[24] S. Afzali and O. Yildiz, "An effective sample preparation method for diabetes prediction," Int. Arab J. Inf. Technol, vol. 15, pp. 968-973, 2018.
[25] N. I. Alghurair, "A Survey Study Support Vector Machines and K-MEAN Algorithms for Diabetes Dataset". Academic Journal of Research and Scientific Publishing, vol.2,pp. 14-25, 2020.

Authorized licensed use limited to: Malta College of ArtsScience&Tech. Downloaded on September 30,2021 at 12:57:19 UTC from IEEE Xplore. Restrictions apply.