International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST229394

A Systematic Review on Machine Learning Algorithms for Diagnosis of Diabetes in Health Care Systems

Podila Mounika[1], Ch .Swetha[2], Dr. Mahesh Kotha[3], D Anusha[4]
[1]Assistant Professor, G Narayanamma Institute of Technology and Science, Hyderabad, India
[2]Assistant Professor, AI Department, Vidya Jyothi Institute of Technology, Hyderabad, India
[3]Assistant Professor, CSE- (AI&ML) Department, CMR Technical Campus, Hyderabad, India
[4]Assistant Professor, AI Department, Vidya Jyothi Institute of Technology, Hyderabad, India

Article Info
Volume 9, Issue 3
Page Number : 422-433

Publication Issue
May-June-2022

Article History
Accepted : 01 June 2022
Published : 07 June 2022

ABSTRACT

Diabetes is an ongoing metabolic problem that influences an expected 463 million individuals around the world. Meaning to work on the therapy of individuals with diabetes, computerized wellbeing has been generally taken on as of late and produced a gigantic measure of information that could be utilized for additional administration of this persistent sickness. Exploiting this moves toward that utilization computerized reasoning and explicitly profound learning, an arising sort of AI, have been broadly taken on with promising outcomes. In this paper, we present a complete survey of the utilizations of profound advancing inside the field of diabetes. We led a deliberate writing search and recognized three primary regions that utilization this methodology: finding of diabetes, glucose the executives, and determination of diabetes-related inconveniences. The hunt brought about the determination of 40 unique exploration articles, of which we have summed up the critical data about the utilized learning models, improvement process, principal results, and pattern strategies for execution assessment. Among the examined writing, it is to be noticed that different profound learning strategies and systems have accomplished state of-the-craftsmanship execution in numerous diabetes-related assignments by outflanking regular AI draws near. In the mean time, we distinguish a few restrictions in the ongoing writing, like an absence of information accessibility and model interpretability. The fast advancements in profound learning and the expansion in accessible information offer the likelihood to address these difficulties soon and permit the far and wide arrangement of this innovation in clinical settings.

Keywords : Diabetes, deep learning, machine learning, diabetic complications, artificial intelligence, prediction.

I. INTRODUCTION

Diabetes is a gathering of deep rooted metabolic problems caused by inadequate insulin emission or disabled insulin activity. The Worldwide Diabetes Federation appraises that there are 463 million individuals (95% certainty span: 369-601 million) living with diabetes in 2019, a big part of whom, notwithstanding, remain undiscovered, because of the complicated pathogenesis of diabetes. The worldwide pervasiveness of diabetes is projected to altogether expansion in the approaching 10 years. Thusly, forestalling and treating diabetes has been a significant weight for public economies, medical care frameworks, and individual clinical uses, particularly for low-and centre pay nations most of individuals with diabetes requiring exogenous insulin utilize the purported basal-bolus insulin treatment, which comprises on estimating glucose levels with a glucose levels meter and conveying multiple daily injections (MDI) with an insulin pen or with an insulin siphon (consistent subcutaneous insulin mixture (CSII)) [6]. For individuals living with diabetes, it is indispensable to keep up with blood glucose (BG) levels in a typical reach. In any case, hyperglycaemia or hypoglycaemia can cause short and long haul difficulties in micro vascular and macro vascular, including neuropathy, nephropathy, retinopathy, stroke, cardiovascular illness, and fringe vascular sickness. By and by, BG control is trying for individuals with diabetes, since there are a lot of day to day factors that impact BG levels, for example, dinner ingestion, work out, liquor, sickness, and stress. Hence self management, e.g., ideal BG estimation, chemical conveyance, furthermore, adherence to suggested way of life are very significant, however, every one of them require multidisciplinary information in clinical practice, particularly for kids and teenagers. In addition, because of the great entomb and intra-populace fluctuation in the glucose energy interaction and pharmacokinetics, it is troublesome to find an ideal restorative procedure for all individuals [9]. In late many years, ceaseless glucose observing (CGM) frameworks [10] and shut circle chemical conveyance frameworks [13], otherwise called the counterfeit pancreas (AP), have been broadly investigated, targeting creating programmed glucose guideline and freeing the weight from glucose the executives. An AP framework utilizes CGM, a shut circle control calculation, and an insulin siphon to convey insulin by CSII. It has been demonstrated to actually decrease glycaemic control what's more, is prescribed to some T1D accomplices [15].

II. MACHINE LEARNING OVERVIEW

Machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the extract information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise. Many industries apply machine learning to extract relevant data. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves without being explicitly programmed. Many mathematicians and programmers apply several approaches to find the solution of this problem which are having huge data sets.

Supervised learning: It consists of a given set of input variables (training data) which are pre labelled and target data [5]. Using the input variables it generates a mapping function to map inputs to required outputs. Parameter adjustment procedure continues until the system acquired a suitable accuracy extent regarding the teaching data.

Unsupervised learning: In this algorithm we only have training data rather a outcome data. That input data is not previously labelled. It is used in classifiers by recognizing existing patterns or cluster in the input datasets [4].

Table 1. ML algorithms for various model building approaches

Learning type	Model building	Examples
Supervised	Algorithms or models learn from labelled data (task-driven approach)	Classification, regression
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, associations, dimensionality reduction
Semi-supervised	Models are built using combined data (labelled + unlabeled)	Classification, clustering
Reinforcement	Models are based on reward or penalty (environment-driven approach)	Classification, control

Reinforcement learning: Applying this algorithm machine is trained to map action to a specific decision hence the reward or feedback Signals are generated. The machine trained itself to find the most rewarding actions by reward and punishment using past experience

There are massive numbers of algorithms used by machine learning are designed to erect models of machine learning and implemented in it [4]. All algorithms can be grouped by their learning methodology, as follows:

Regression algorithms: In Regression algorithms predictions are made by the model with modelling the relationship between variables using a measure of error[25]. Continuously varying value is predicted by the Regression technique. The variable can be a price, a temperature.

Instance based learning algorithms: In the algorithms which based on Instance, decision problem is a issue with illustration of training data build up a database and compare test data then form a prediction. Instance-based learning method is famous as lazy learner.

Algorithms using Decision Tree: Algorithms using Decision trees are used mainly in classification problem. They splits attributes in two or more groups by sorting them using their values. Each tree have nodes and branches [4]. Attributes of the groups are represented by each node and each value represented by branch [5].

Baysian algorithms: Machine Learning is multidisciplinary field of Computer Science like Statistics and algorithm. Statistics manages and quantifies the uncertainty and are represented by bayesian algorithms based on probability theory and Bayes' Theorem.

Data Clustering algorithms: This algorithm split items into different types of batches. It groups the item set into clusters in which each subset share some similarity. It is unsupervised learning method and its methods are categorized as hierarchical or network clustering and partitioned clustering.

Learning algorithms using Association Rule: Learning algorithms using Association rule are generally utilized by the organization commercially when multidimensional datasets are huge in size. They are used as extraction methods that can explore observed relationships between variables and data.

Algorithms using Artificial Neural Network: Artificial neural networks models are based on the biological neuron structure and uses supervised learning. It consists of artificial neurons which have weighted

interconnections among units. They are also well known by parallel distributed processing networks.

Deep Learning algorithms: Deep Learning methods upgraded the artificial neural networks They are more complex neural networks are large in size.

Algorithms using Dimensionality Reduction: Dimensionality reduction method is widely used in case of large number of dimensions, large volume of space concerned. Then that problem requires a statistical significance. Dimensionality reduction methods used for minimizing the number of dimensions outlined the item and removes unrelated and un essential data which lessen the computational cost. Some of these methods are used in classifying and regression.

Ensemble Algorithms: They are based on unsupervised Learning. It groups the teaching data into many types of classes of data. Self-supporting models for learning are built for those groups. To make correct hypothesis all learning models are combined.

III. RELATED WORK

Diabetes is a very dangerous disease and does not able to cure. If this disease affects once, it will maintain in your life time. At the same time, your blood having too much of glucose can cause health issues. Like kidney disease, heart disease, stroke, eye problems, dental disease, foot problems, nerve damage. So you can take step to oversee your diabetes and avert these complications. The most general type of diabetes is type 1 and type 2. In this type of diabetes create problems like the body does not able to produce or use insulin. But there are also other kinds of diabetes, like gestational diabetes, which crop up during pregnancy. Gestational diabetes causes high blood sugar that can affect your pregnancy and you baby's health. Several machine learning and data mining methods are used to diagnoses diabetes and administering diabetes. This study focuses on new developments in machine learning which have made significant impacts in the detection and diagnosis of diabetes. In this work, the machine learning algorithms are used to classifying diabetes patients. The classification accuracy is achieved by the classifying diabetes patients.

Diabetes is a very dangerous disease and does not able to cure. If this disease affect once, it will maintain in your life time. At the same time, your blood having too much of glucose can cause health issues. Like kidney disease, heart disease, stroke ,eye problems, dental disease, foot problems ,nerve damage.so you can take step to oversee your diabetes and avert these complications.

The familiar types of diabetes:

- *Type 1 diabetes*
- *Type 2 diabetes*
- *Gestational diabetes*

Type1 Diabetes

Body does not able to produce insulin. Its affect children and young adults. Also it can affect at any age. Peoples affected by this type of diabetes to take insulin every day.

Type2 Diabetes

Body does not able to produce or use insulin. This type of diabetes mostly affected on middle aged and up in years.

Gestational Diabetes

Women's are mostly affected by this type of diabetes. This type of diabetes develops during
pregnancy. Gestational diabetes causes high blood sugar that can affect your pregnancy and you baby's health.

Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to

generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression technique. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes.[1] Machine learning is considered to be one of the most important artificial intelligence features supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming.

Table 2. Summary of selected articles from the literature on diabetes diagnosis.

Cases	Models	Data Sources	Development Process	Main Outcomes	Baselines
Classification of diabetes‡	Denosing AE	Mount Sinai Data Warehouse⋆ (ICD-9)	Normalization; pre-process to obtain raw features; the data of training, validation and testing: 704,587, 5000, 76,214 patients	AUC: 0.907	Original descriptors, PCA (0.861)
Prediction of diabetes‡	Modified LSTM, attention pooling layer	An EHR dataset from a regional hospital (7191 patients, ICD-10)	The split for training, validation and testing: 2/3, 1/6 and 1/6 from 53,208 admissions	Precision of diagnosis, intervention, unplanned readmission: 66.2%, 78.7%, 79.0%	SVM, RF, plain RNN, LSTM (65.7%, 78.2%, 75.9%)
Detection of diabetes†‡	RBM and RNN	PID dataset from UCI repository⋆	Feature selection by RFs; min-max normalization; the ratio for training and testing data: 80%, 20%	Sensitivity and precision: 90.66%, 75%	N/A
Prediction of diabetes†‡	Modified 1-D CNN and FC layers	25 breath samples collected by MOS sensors with 1000-sec intervals	The data for training and testing: 15 samples, 10 samples; leave-one out cross-validation	AUC of T1D, T2D, healthy subjects: 0.9659, 0.9625, 0.9644	SVD, SVM, PCA
Detection of diabetes	5-layer CNN, LSTM, and SVM	ECG data sampled at 500 Hz with digital bandpass filtering and thresholding collected from 40 people	Heart rate variability (HRV) data from 71 ECG datasets (each contains 1000 samples); 5 fold cross-validation	Validation accuracy: 95.7%	Previous work using HRV
Detection of diabetes‡	DMLP with dropout	PID dataset from UCI repository⋆	The ratio of training and validation data: 90% and 10%	Accuracy: 88.41%	Previous work on the same dataset
Prediction of diabetes‡	DMLP	A population dataset (4814 participants,the majority are overweight)	Data cleaning (imputing missing values with the median); the ratio of training and testing data: 80% and 20% from 656 T2D subjects	AUC without and with HbA1c: 0.703, 0.840	SVM (0.679,0.825)
Prediction of the onset T2D‡	DMLP and a linear model	Practice Fusion dataset (9948 patients, ICD-9)⋆	Feature extraction by grouping 1312 features; the ratio of training and validation data: 70%, testing data: 30%; 10-fold cross-validation	Sensitivity: 31.17%, AUC: 84.13%	RF (29.12%, 16.07%)
Detection of diabetes‡	2 layer AE and a softmax layer	PID dataset from UCI repository⋆	Training the layer one by one with previous output; fine-tuning by supervised learning	Sensitivity: 87.92%, specificity: 83.41%, accuracy: 86.26%	Previous work on the same dataset
Prediction of diabetes‡	DBN	PID dataset from UCI repository⋆	Min-max normalization; feature selection by PCA; pre-training for RBMs; supervised fine-tuning	Sensitivity: 100%, F1 score: 0.808	DT, LR, RF, SVM, NB (75.9%, 0.760)
Detection of undiagnosed diabetes‡	2 hidden layer DMLP with dropout	An EHR dataset from a national survey (31,098 subjects, 4 years)	Combining 2013-2016 datasets;selecting features by LR; the data of training and testing: 11456 and 4444 subjects	AUC: 80.11%	LR, KNN, SVM, AdaBoost, Gaussian NB, RF (79.05%)

Table 2. Comparative analysis of diabetes prediction using machine learning methods

S.No	Method Name	Dataset	Size	Classifier used	Feature selection method	Speed	Classification accuracy
1	Kamrul Hasan	PID	768	NN, DT, RF, MLP,AB, XB, and NB	CA, ICA, and CRB	Slow	78.9%
2	Quan Zou	uzhou and PIDD	68994, 768	N J4	PCA and mrM	Slow	80.84%
3	Nishith Kumar	PIDD	768	GPC, LDA, QDA, and NB	Kernels	Fast	91.97%
4	Maniruzzuman	NHANES	9858	NB, DT, RF, and AB	LR	Slow	92.75%
5	V. Jackins	PIDD	768	B and RF	CRB	Fast	74.46%
6	N. Sneha	PIDD	2500	SVM, RF,NB, DT, and KNN	CRB	Slow	82.3%
7	S. Mohapatr	PIDD	768	MLP	None	Fast	77.5%
8	D. Sisodia	PIDD	768	NB, SVM,and DT	None	Fast	76.3%
9	Orabi	Egyptian National Research Centr	Not mentioned	DT	Not mentioned	Slow	84%
10	O. M. Alad	PIDD	768	NN	None	Fast	Prediction

IV. LITERATURE SURVEY

Rob Law (1998) [7] applies neural networks to forecasts occupancy rates for the rooms of Hong Kong hotels and finds that neural networks outperforms naïve extrapolation model and also superior to multiple regression. This research studied the feasibility incorporating the neural network to predict the rate of occupancy of rooms in Hong Kong hotel industry.

Authors Hua et al. (2006) [8] described support vector machines approach to predict occurrences of non zero demand or load time demand of spare parts which used in petrochemical enterprise in china for inventory management. They used a integrated procedure for establishing a correlation of explanatory variables and autocorrelation of time series of demand with demand of spare parts. On performing the comparison the performance of SVM method with this LRSVM model, Croston's model , exponential smoothing model, IFM method and Markov bootstrapping procedure., it performs best across others.

Authors Vahidov et al. (2008) [9] compares the methods of predicting demand in the last of a supply chain, the naive forecasting and linear regression and trend moving average with advanced machine learning methods such as neural networks and support vector machines, recurrent neural networks finds that recurrent neural networks and support vector machines show the best performance.

Wang (2007) [10] describes the machine learning method with genetic algorithm (GA)-SVR with real value GAs, . The experimental findings investigates this , SVR outshines the ARIMA models and BPNN

regarding the base the normalized mean square error and mean absolute percentage error .

Authors Chen et al. (2011) [11] presents a method forecast the tourism demands that is SVR built using chaotic genetic algorithm (CGA), like SVRCGA, which overcome premature local optimum problem. This paper reveal that suggested SVRCGA model outclass other methodologies reviewed in the research paper.

Turksen et al. (2012) [12], presents next-day stock price prediction model which is based on a four layer fuzzy multi agent system (FMAS) structure. This artificial intelligence model used the coordination of intelligent agents for this task. Authors investigates that FMAS is a suitable tool for stock price prediction problems as it outperforms all previous methods.

Shahrabi et al. (2013) [13] proposed a method for estimating tourism demand which is a new combined intelligent model i.e. Modular Genetic-Fuzzy Forecasting System using a genetic fuzzy expert systems and finds that accuracy of predicting power of MGFFS is better than approaches like Classical Time Series models , so it is suitable estimating tool in tourism demand prediction problems.

Chen Hung et al. (2014) [14] proposes forecasting model for tourists arrival of Taiwan and Hong Kong named as LLSSVR or logarithm least-squares support vector regression technologies. In combinations with fuzzy c-means (FCM) and Genetic algorithms (GA) were optimally used and indicates that method explains a better performance to other methods in terms of prediction.

Guang-Bin Huang et al. (2015) [15] explores the basic features of ELMs such as kernels , random features and random neurons, compares the performance of ELMs and shows it tend to outshine classification, support vector machine and regression applications.

Wang et al. (2016) [16] proposed a novel forecasting method CMCSGM based Markov-chain grey model which used algorithm of Cuckoo search optimization to make better the performance of the Markov chain grey model. The resultant study indicates that the given model is systematic and fine than the traditional MCGM models.

Barzegar et al. (2017) [17] demonstrates model predict multi-step ahead electrical conductivity i.e. indicator of water quality which is needed for estimating the mineralization, purification and salinity of water based on wavelet extreme learning machine hybrid or WAELM models and extreme learning machine which exploiting the boosting ensemble method. The findings showed that upgrading multi WA ELM and multi WAANFIS ensemble models outshines the individual WAELM and WA ANFIS constructions.

Authors Fouilloy et al. (2018) [18] suggested a statistical method employing machine learning model and to analyze and applied it to solar irradiation prediction working hourly. This methodology used the high, low and medium meteorological variability like Ajacio, Odeillo , Tilos . They compared model with auto regressive moving average and multi-layer preceptor .

Makridakis et al. (2018) [19] presents Machine Learning methods to statistical time series forecasting and compared the correctness of those methods with the correctness of conventional statistical methods and found that the first one is better and outtop using the both measures of accuracy. They provide the reason for the accuracy of learning models is less that of statistical models and suggested some other achievable ways .

Zhang et al. (2018) [20] suggests a design of multi kernel ELM or MKELM method for segregation of motor imagery electroencephalogram or EEG and

investigate performance of kernel ELM and impacts of two different functions of kernel such as polynomial and Gaussian kernel Compares MKELM method gives greater segregation accuracy than other algorithms indicates betterment of the suggested MKELM based .

Table 3 A comprehensive study of the machine learning methods done by some researchers

Algorithm	Method used/innovation	Application and future work	Results and limitations
J48, AdaBoost, And bagging on base classifier [2]	The model was performed on Canadian Primary Care Sentinel Surveillance Network dataset with several features to train on. The author used ensemble methods AdaBoost on base classifier J48 DT.	The author claimed that these ensemble algorithms can be used on other disease datasets to increase accuracy.	The AdaBoost algorithm with the J48 as the base classifier showed the maximum accuracy followed by bagging and then the J48 classifier. The AROC was used as the parameter.
NB with clustering [3]	Dataset used was the PIMA Indians Diabetes Dataset with eight attributes. The model is NB performed on prior clustering. This model is compared with only the NB model. Five hundred and thirty-one instances of data were divided into 5 clusters. The fourth cluster was the only one used for testing, which consisted of 148 instances.	By collecting a large amount of data for training, the accuracy can be increased by many-fold, helping people by developing a system that gives them a correct prediction without having to consult a doctor.	The parameters used for evaluation are accuracy, sensitivity, and specificity. The model with clustering showed a 10% increased accuracy, rise in sensitivity by 53.11% but the imitation caused here was the fall of specificity by 10.99% and also a reduced amount of dataset.
DTs, LR, and NB with bagging And boosting [4]	Initial datasets were collected from primary care units, which (through further changes) consisted of 11 features and a data of 30122 people. The three algorithms are used along with bagging and boosting methods, which are to decrease overfitting and increase accuracy.	The final model obtained with highest accuracy was deployed on a commercial web application.	The following data shows the accuracy with bagging and boosting. DT 85.090, LR 82.308, NB 81.010, Bagging with DT (BG+DT) 85.333, bagging with LR (BG+LR) 82.318, bagging with NB (BG+NB) 80.960, boosting with DT (BT+DT) 84.098, boosting with LR (BT+LR) 82.312, and boosting with NB (BT+NB) 81.019. RF 85.558 shows the maximum accuracy. The ROC was used for final validation.
	The author collected a raw	The author proposed	The RF classifier was the algorithm that

LR, KNN, SVM, LDA, NB, DT, and RF[12]	dataset from Noakhali medical hospital containing 9843 samples with 14 attributes. Eighty percent of the data was taken for training and the rest for testing.	that we can enhance the accuracy of early treatment to lessen the suffering of patients.	performed the best in classifying data and LR showed the worst performance. Although machine learning classifiers are widely used, they still lack in terms of accuracy against deep learning models.
LR and DTs [12]	The dataset was prepared using a questionnaire carried out for 1487 individuals in which 735 were diabetic and the remaining 752 negatives. A Pearson chi-square test was carried out on all the characteristics. The models' performance was evaluated on three parameters: accuracy, sensitivity, and specificity.	Recently, many researchers have been implementing various algorithms and networks to compare them and find out the most feasible one. DTs and LR are among the ones that are most used.	LR achieved a ACC of s76.54%,sensitivity of 79.4%, and specificity of 73.54% on the testing data while the DT gained an accuracy of 76.97%, sensitivity of 78.11%, and specificity of 75.78%. Overall, the DT model performed better than the LR model. The model poses a limitation of the dataset. It is collected only from one area of China, if it had been collected from different regions, the model implementation could be more practical.
SVM and LR [13]	Practice fusion de-identified dataset was used for the study taken from Kaggle containing data of approximately 10000 patients.	LR is a model that is widely used in public health and clinical practice for disease detection and to calculate risks.	On using a smaller subset of features, the LR model performed slightly better than the SVM model.
DTs and NB[15]	The dataset taken for consideration was the PIMA Indian diabetes database. On applying feature selection, the author obtained five features. 10-fold cross validation was used for data preparation after which the J48 algorithm – DTs and NB is applied.	The author proposed to gather information for the dataset from different people to make a more representative model. The work can be further enhanced to include automation.	Using a percentage split of 70:30, the J48 DT algorithm correctly classified 177 instances (76.95%) whereas the NB got an accuracy of 79.56%. The accuracy obtained performed better on the percentage split, which shows
RF and XG Boost [16]	The author used the PIMA diabetes dataset. Using Jupyter Notebook as an IDE, the author trains the model using 8 attributes of the total 9 provided in the dataset.	The author suggests the use of more algorithms in this branch of machine learning like hybrid model for better accuracies.	The accuracy gained on the RF classifier came out to be 71.9%. The hybrid model proposed through XG boost gained an accuracy of 74.1%.

DT (J48) and NB[17]	The author used the PIMA Indian diabetes dataset with 8 attributes, which was reduced to 5 based on the feature selection. The pre-processing was performed used the WEKA using 10-fold validation. The model was created using the 70% dataset and the rest was used fortesting.	In future, it is planned to gather the information from different locales over the world and make a more precise and general prescient model for diabetes conclusion.	The J48 algorithm was 76.95% accurate with other parameters like kappa statistic, MAE, RMSE, relative absolute error, and root relative absolute error. The NB algorithm was accurate up to 79.56%. Since this model is not optimally configured, a developed model would require more training data for creation and testing.
LR, DT, RF, SVM [19]	The author used the PIMA Indian women dataset concerned with women's health with 8 attributes. Different models were trained for this dataset under different hyperparameters.	The author proposed to create advanced models on RF because of its highest accuracy and ability to overcome overfitting.	Different models were compared on basis of accuracy. RF gained the highest accuracy with 77.06% followed by SVM.

V. OBSERVATIONS

After conducting a survey of various articles on diabetic prediction models, we strongly recommend our study because of the following reasons:

a) We have included recent articles.
b) We have presented a comparative statement of major diabetic prediction models which will help other researchers to understand and evaluate the models.
c) Advantages and disadvantages have been presented.
d) Various Strategies to predict diabetes have been discussed in the paper.

All the methods proposed so far for the prediction of diabetes are focusing more towards feature selection strategy and few machine learning methods such as random forest, naive Bayes, support vector machine, and decision trees, whereas only a few features are to be selected for prediction purpose. While studying all of these articles, the challenges that we faced are as follows:

a) The major challenge in prediction purpose was the absence of a larger dataset since the publicly available dataset contains only nine attributes, one being the class attribute. Time and effort are being spent on those features that have no potential to be selected for prediction purposes.
b) Most of the authors have dropped missing values from the standard dataset, which can affect the results as the size of the dataset decreases.
c) General machine learning algorithms are applied to the dataset; only one author has made use of AdaBoost and gradient boost technique. None of the authors has made use of the recurrent neural network or deep learning technology, which can help in increasing the efficiency. So, a method needs to be developed which can deliver more accurate results, has to be fast in terms of processing, and is more effective for the prediction purpose.

VI. CONCLUSION

In this paper, we present a comprehensive review of the current trend in machine learning technologies for diabetes research. We performed a systematic search, selected a collection of articles, and summarized the key information focusing on three areas: diagnosis of diabetes, glucose management, and diagnosis of diabetes related complications. In these areas, various DNN architectures and learning techniques have been applied and obtained superior experimental performance that previous conventional machine learning approached. On the other hand, several challenges have been identified from the literature including data availability, feature processing, and model interpretability. In the future, there is great potential to meet these challenges by transferring the latest advances in deep learning technologies into massive multi-modal data of diabetes management. We expect that deep learning technologies will be widespread in clinical settings and largely improve the treatment of people living with diabetes.

VII. REFERENCES

[1]. Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Manag Process 5(1): 1–14.

[2]. Taiyu Zhu,"Deep Learning for Diabetes: A Systematic Review",2168-2194 (c) 2020 IEEE.

[3]. Barik S, Mohanty S, Mohanty S, Singh D (2021) Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In: Mishra D, Buyya R, Mohapatra P, Patnaik S (eds) Intelligent and cloud computing. Smart innovation, systems and technologies, vol 153. Springer, Singapore, pp 399–409.

[4]. Ravindra Changala, "Development of Predictive Model for Medical Domains to Predict Chronic Diseases (Diabetes) Using Machine Learning Algorithms and Classification Techniques", ARPN Journal of Engineering and Applied Sciences, VOL. 14, NO. 6, MARCH 2019, ISSN 1819-6608

[5]. Ephzibah EP (2011) A hybrid genetic-fuzzy expert system for effective heart disease diagnosis. In: Wyld DC, Wozniak M, Chaki N, Meghanathan N, Nagamalai D (eds) Advances in computing and information technology. first international conference, ACITY 2011, July 2011. Communications in computer and information science, vol 198. Springer, Berlin, Heidelberg, pp 115-121.

[6]. M.Rajeswari,"A Review of Diabetic Prediction Using Machine Learning Techniques", International Journal of Engineering and Techniques - Volume 5 Issue 4,July 2019.

[7]. Ravindra Changala, "Statistical Models in Data Mining: A Bayesian Classification" in International Journal of Recent Tredns in Engineering & Research (IJRTER), volume 3, issue 1, pp.290-293. in 2017.

[8]. Zheng T, Xie W, Xu LL, He XY, Zhang Y, You MR et al (2017) A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 97:120–127.

[9]. Md. Kamrul Hasan, Md. Ashraful Alam, D. Das, E. Hussain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Acess, vol. 8, pp. 76516–76531, 2020.

[10]. Umair Muneer,"Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", Hindawi,Journal of Healthcare Engineering Volume 2021, Article ID 9930985.

[11]. V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," Fe Journal of Supercomputing, vol. 77, no. 5, pp. 5198–5219, 2021.

[12]. Ravindra Changala, "Evaluation and Analysis of Discovered Patterns Using Pattern Classification Methods In Text Mining" in ARPN Journal of Engineering and Applied Sciences,Volume 13,

Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018.

[13]. Aishwarya M,"Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science 165 (2019) 292–299.

[14]. N. Sneha and T. Gangil, “Analysis of Diabetes Mellitus for Early Prediction Using Optimal Feature Selection,” Journal of Big Data, vol. 6, 2019.

[15]. Ravindra Changala, “Integrating Different Machine Learning Techniques for Assessment and Forecasting of Data” in Springer series, August-2015.

[16]. Perveen S, Shahbaz M, Guergachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci 82:115–121.

[17]. Khan NS, Muaz MH, Kabir A, Islam MN (2019) A machine learning-based intelligent system for predicting diabetes. Int J Big Data Anal Healthc 4(2):1.

[18]. Nai-Arun N, Moungmai R (2015) Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci 69:132–142.

[19]. Roshi Saxena,"A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey", Hindawi, Journal of Healthcare Engineering Volume 2022, Article ID 8100697.

[20]. Mahesh Kotha, "A Survey on Predicting Uncertainty of Cloud Service Provider Towards Data Integrity and Economic" 2019 IJSRST | Volume 6 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X.

[21]. Kocher T, Holtfreter B, Petersmann A, Eickholz P, Hoffmann T, Kaner D et al (2019) Effect of periodontal treatment on HbA1c among patients with prediabetes. J Dent Res 98(2):171–179.

[22]. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q (2013) Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 29(2):93–99.

[23]. Sheikhi G, Altınçay H (2016) The cost of type II diabetes mellitus: a machine learning perspective. In: Kyriacou E, Christofides S, Pattichis CS (eds) XIV mediterranean conference on medical and biological engineering and computing 2016. IFMBE proceedings, vol 57. Springer, Cham, pp 818-821.

[24]. Mahesh Kotha, "Predict Trustworthiness of Cloud Services Using Linear Regression Model", International Journal of Advanced Science and Technology Vol. 29, No. 4, (2020), pp. 5737 – 5745.

[25]. Raja Krishnamoorthi,"A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Hindawi, Journal of Healthcare Engineering Volume 2022, Article ID 1684017.

[26]. Ravindra Changala, “Automated Health Care Management System Using Big Data Technology” ,Journal of Network Communications and Emerging Technologies (JNCET), Volume 6, Issue 4, April (2016), 2016, pp.37-40,ISSN: 2395-5317, ©EverScience Publications.

[27]. Vinoda Reddy, “Recurrent Feature Grouping and Classification for action model prediction in CBMR”, International Journal of Data Management and Knowledge Process, Vol.7, No.5/6, November 2017, pp. 63 74.http://dx.doi.org/10.5121/ijdkp.2017.7605.

Cite this article as :

Podila Mounika, Ch .Swetha, Dr. Mahesh Kotha, D Anusha, "A Systematic Review on Machine Learning Algorithms for Diagnosis of Diabetes in Health Care Systems", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 3, pp. 422-433, May-June 2022. Available at doi : https://doi.org/10.32628/IJSRST229394

Journal URL : https://ijsrst.com/IJSRST229394