A Comparative Analysis on Diagnosis of Diabetes Mellitus Using Different Approaches – A Survey

Ramya S[1], Kalaivani D [2]

[1]*Student, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, India.*

[2]*Associate Professor, Department of Computer Technology, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore, India.*

Corresponding Author: ramyasiva31@gmail.com

Abstract: Diabetes Mellitus is commonly known as diabetes. It is one of the most chronic diseases as the World Health Organization (WHO) report shows that the number of diabetes patients has risen from 108 million to 422 million in 2014. Early diagnosis of diabetes is important because it can cause different diseases that include kidney failure, stroke, blindness, heart attacks, and lower limb amputation. Different diabetes diagnosis models are found in literature, but there is still a need to perform a survey to analyze which model is best. This paper performs a literature review for diabetes diagnosis approaches using Artificial Intelligence (neural networks, machine learning, deep learning, hybrid methods, and/or stacked-integrated use of different machine learning algorithms). More than thirty-five papers have been shortlisted that focus on diabetes diagnosis approaches. Different datasets are available online for the diagnosis of diabetes. Pima Indian Diabetes Dataset (PIDD) is the most commonly used for diabetes prediction. In contrast with other datasets, it has key factors which play an important role in diabetes diagnosis. This survey also throws light on the weaknesses of the existing approaches that make them less appropriate for a diabetes diagnosis. In artificial intelligence techniques, deep learning is widespread and in medical research, heart rate is getting more attention. Deep learning combined with other algorithms can give better results in diabetes diagnosis and heart rate should be used for other cardiac disease diagnoses.

Key Words: —***Diabetes Mellitus, World Health Organization, Pima Indian Diabetes Dataset, Deep learning.***

I. INTRODUCTION

Among medical diagnosis, a diabetes diagnosis is one of the major challenges. The World Health Organization (WHO) report shows that the number of diabetes patients has risen from 108 million to 422 million in 2014. An estimate shows that by 2045, this number may reach 629 million. In 2016, the estimated 1.6 million deaths were reported due to diabetes. Early diagnosis of diabetes is significant in lowering the chances of different diseases like kidney failure, stroke, blindness, heart attacks, and lower limb amputation. Many machine learning techniques have been used in the medical diagnosis system. They have proven to be accurate in diagnosis, successful in treatments, and cost-efficient.

Manuscript revised June 26, 2022; accepted June 27, 2022. Date of publication June 30, 2022.

This paper available online at www.ijprse.com

ISSN (Online): 2582-7898; SJIF: 5.59

Diabetes Mellitus is a metabolic disorder in which the body is unable to use insulin or to store and use glucose for energy and does not make insulin [1]. Different classification techniques are used to deal with different medical problems. There are multiple types of diabetes, such as Type1, Type 2, and gestational diabetes. In type 1, the pancreas fails to produce sufficient insulin for the body. Whereas in Type 2, the body is unable to use insulin properly. It is the most common type of diabetes. The third type of diabetes is Gestational Diabetes. It occurs in pregnant women having high glucose levels in the blood [4].

Deep learning is a subset of machine learning in Artificial Intelligence (AI) that can self-learn from the data. It is also capable of unsupervised learning. It can learn a large amount of unstructured and unlabeled data that even a human brain can take years to understand. Deep learning uses multiple layers to extract features from raw data. Deep learning models are based on artificial neural networks, and Convolutional Neural Network (CNN) is one of them. Architecture of simple neural network is shown in Fig. 1.

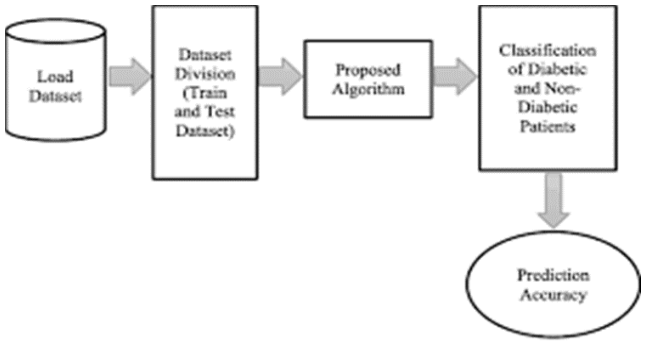

Fig.1. General flowchart for diabetes diagnosis

Fuzzy logic is a method of reasoning that is modelled upon human cognitive and analytical abilities. It involves both possibilities of YES or NO. A computer gives the output as TRUE or FALSE that in human language is equivalent to YES or NO. Different techniques are used by researchers for diabetes diagnosis such as Backpropagation neural network (BPNN) [3]. Similarly, researchers in Ref. [4] show the performance of the Small World FANN model in diabetes diagnosis. Moreover, an artificial neural network-based approach is presented in Refs. [16]. Many researchers used Pima Indian Diabetes Dataset (PIDD) for a diabetes diagnosis. Pima Indian Diabetes Dataset consists of eight parameters. Those parameters include the number of times pregnancy has occurred, BMI, plasma glucose, diastolic blood pressure, systolic blood pressure, skinfold thickness, diabetic pedigree function, and Class 0 or 1 (0 means non-diabetic while 1 means diabetic patient). The literature review shows that PIDD might be the best dataset for diabetes diagnosis as it has a large number of values making it a standardized dataset. Other small datasets are also discussed in the literature for example data collected from patients directly, data collected through surveys, heart signals (ECG signals), CGM Signals, images dataset, Eye dataset, Skin dataset, and Ayurvedic dataset.

II. LITERATURE REVIEW

This work presents a survey for diabetes diagnosis with some new contributions outlined below:

- Literature Review is performed to analyse the existing latest approaches for diabetes diagnosis with some suggestions for future research.
- Several related schemes from the last decade have been searched as per research questions and carefully studied to identify strengths and weaknesses.
- The quality evaluation has been performed to verify articles linked with research questions

The rest of the document is divided as follows: Section 2 throws light on the Literature Review of studied articles. Section 3 presents an analysis of the survey considering different evaluation measures, and Section 4 includes a comprehensive conclusion. 2. Literature review the literature review helps us in identifying specific areas or research questions, or gaps in the literature that already exists. 2.1. Research questions the main objective of this research is to find a question for our research. Question: “Is there any algorithm that has better accuracy using large dataset/Pima Indian Diabetes Dataset?” 2.2. Databases Digital Libraries used are: (i) Science Direct (www.sciencedirect.com/) (ii) IEEE (www.ieeexplore.ieee.org/) (iii) Springer (www.springerlink.com/) (iv) Others (https://scholar.google.com.pk/).

III. METHODOLOGY

Collection of study: This collection of articles for study are collected based on: (i) Research articles on diabetes diagnosis (ii) Research articles with available PDF (iii) Research articles vary from the last decade (iv) Articles based on surveys if required Early Diabetes diagnosis is important for human health to saves them from the fatal effects of diabetes. In the past few years, different techniques have been introduced using a variety of models and approaches to diagnose diabetes. Those techniques include neural network-based approaches, deep learning approaches, and machine learning approaches, decision making approaches, k-NN approach, retinal images-based approaches and face image-based diagnosis techniques. 2.4. Neural network approaches Researchers in Ref. [3] proposed Back Propagation Neural Network (BPNN). Graphical User Interface (GUI) was built in MATLAB. Pima Indian Diabetes Dataset is used by researchers to test their proposed methodology. Once loading of the dataset is completed, parsing was performed. After reading values one by one, they were stored to train ANN using Back Propagation Neural Network. In feature extraction phase values were classified with similar features and also arranging of groups was done in the column. Normalization was the next step of the proposed technique. Data values were represented within 0 and 1. Normalization removes data redundancy and guarantees data dependencies. The training was the last step of the proposed technique. Up to 9 iterations were performed to train the proposed system. The minimum

error was found in the 3rd iteration. Best results were obtained at lower epoch values. Results were created using the regression plot and validation plot. Feed Forward ANN (FFANN) becomes prominent in today's world because of its computational speed and efficiency. Researchers in Ref. [4] presented the performance of the Small World FANN model in diabetes diagnosis. For the investigation, researchers considered four-layered FFANN. There were eight inputs in the network that includes one output neuron. They used two hidden layers in FFANN. Two different network topologies were used for FFANN. SW-FANN Activation function used by researchers in the proposed methodology is bipolar-sigmoid function. Backpropagation learning algorithm with training was used for the training process of SW-FFANN.

IV. Results and Analysis

The data set used for this research was PIDD taken from the UCI repository. The rewiring process was applied to the best regular topology for SW-network construction. DGlobal and DLocal parameters were calculated for each rewiring step. The artificial neural network-based approach was presented in Ref. [16]. The artificial neural network has three main layers: input, hidden, and output layers. The input layer gets raw data. Hidden layer's function is determined using inputs and weights assigned to them. The data was entered into a JNN tool that determines the values of attributes. Afterwards, training, testing, and validation of data were performed. The proposed system provided output in binary numbers. 0 as a diabetic patient and 1 as a healthy person. An average error rate of the proposed system was 0.010. The number of epochs performed on the dataset was 158,000. Samples for training data were 767, and samples to validate the system were 237. Researchers in Ref. [12] used skin impedance and heart rate variability for the detection of diabetes. Artificial neural networks were used for classification. Skin impedance data were collected from 11 patients having diabetes that include six females and five males with an average age of 40 ± 8 years. Also, data of eight normal persons were collected that includes five females and three males with an average age of 24 ± 3 years. To measure signal power at different frequencies, Welch Method was used. ECG data was collected from 20 healthy persons including fourteen males and six females with an average age of 22 ± 7 years. Also, data of 20 diabetic patients were collected including eight females and twelve males with an average age of 40 ± 8 years. Pre-processing was performed on raw ECG signal removing baseline drift in a signal using median filtering. The noise of high frequencies was also removed using butter worth a lowpass filter. Then smoothing of the ECG signal was performed using the Savitzky-Golay filter. Table 1 briefly explains different neural network approaches for the diagnosis of diabetes. All approaches show better results, but ANN [12] outperforms all other neural network approaches. 2.5. Machine learning approaches ANFIS was proposed in Refs. [1] that was based on Sugeno FIS. The proposed methodology was the hybridization of an artificial neural network and Fuzzy inference system having a learning ability. Features were adapted from the Artificial Neural Network. ANFIS comprised of two parts antecedents, and conclusion. It consists of five layers having its own functionality. X and Y were values of input against nodes, while fuzzy sets were represented as Ai and Bi. The triangular membership function was used in the proposed techniques. The output of the first layer becomes the input of the second layer. Normalization of data was performed in the third layer. Datasets used to perform experiments were taken from the locals of Bhubaneswar, Odisha, India. Levenberg-Marquardt backpropagation algorithm was used to train the ANN system. Researchers in Ref. [11] used Pima Indian Diabetic Dataset to classify diabetic patients and diabetes diagnosis using different machine learning techniques. To classify diabetic patients and normal persons, some sets of characteristics were used that are selected according to WHO criteria. Researchers use those sets of characteristics as features vectors. Feature vectors were composed of all eight features from a selected dataset. Three stages of the evaluation were performed by researchers [11]. The first one showed a comparison of the state of diabetic and non-diabetic peoples. The second evaluation stage used hypothesis testing to check if the feature vector showed different distributions for diabetic and non-diabetic patients. In the last stage classification, the analysis was performed to assure whether all features can discriminate between diabetic patients and non-diabetic patients. Machine learning classification algorithms like J48, JRip, MultilayerPerceptron, RandomForest, HoeffdingTree, and BayesNet were used. Weka tool was used for performing classification analysis. The null hypothesis got rejected by all eight features, statistically showing that all these features can distinguish between diabetic and non-diabetic patients. Five different techniques of machine learning were used in Ref. [15] for diabetes diagnosis and pre-processing of data. Those techniques include DNN, Logistic Regression, Decision Tree, SVM, and Naïve Bayes. Those techniques were used on Pima Indian Diabetic Dataset to calculate the accuracy of cross-validation. Five pre-processing steps were performed on the

dataset. After each step, the accuracy of all algorithms was calculated and compared. Those data pre-processors include imputation, scaling, normalization separately, imputation and scaling, imputation and normalization. Imputation was the process to calculate the missing values of a dataset. After performing data pre-processing steps, the comparison of the results showed that Naïve Bayes and Decision Tree performed the same on the original dataset and the scaled dataset in terms of accuracy. All other classifiers also showed good results in terms of accuracy on a scaled data set. Different machine learning models: k-NN, Naïve Bayes, Decision Tree, Random Forest, SVM, and logistic regression were used in Ref. [13] to identify type 2 diabetes using electronic health records. From the total number of 23,281 diabetes-related patients, 300 samples were selected. All samples were un-labelled. Two clinical experts were called to label the dataset. From 300 samples, 161 were typed 2 diabetic patients, 60 were non-diabetic patients and 79 samples were unconfirmed. 78.3% of samples were incomplete those 79 samples were dropped. The feature construction model was used to convert that electronic health records (raw data) into statistical features so that it can be used as input for classification models. Related features were summarized using summation to form new features. From 36 features, eight features were extracted using feature summarization. These features were used as input for classification models like k-NN, Naïve Bayes, Decision Tree, Random Forest, SVM, and logistic regression. Also, the ability to diagnose type 2 diabetes was tested using the same classification models. Weka tool was used to apply those classification models on a dataset. Proposed classification model performance based on parameters such as accuracy, precision, specificity, sensitivity, and AUC. The graph-based approach was proposed in Ref. [9] to classify the retinal image. Retinal vessels are of 2 types' veins and arteries. The most important phase is the extraction of retinal vessels to detect vascular changes. The retinal image of a patient was used to calculate the artery vein ratio. Diabetes recognition was done using an artery-to-vein ratio. The implementation of the proposed system was done in different stages. The first one was the preprocessing. Extraction of the green channel from scanned retinal images was done in this stage. This stage improved unprocessed image quality by removing noise and eliminating irrelevant information. The Green channel image was calculated using equation Eqn (1) [9]. g = G R + G + B (1) Enhancement was used to clear an image. Edge detection was the next stage of the proposed technique. Edge detection techniques were applied to the retinal image to extract blood vessels. Researchers in Ref. [9] used the canny edge detection technique. Kirsch template was used to identify the presence of edge and finally to extract blood vessels from the retinal image. Graph-based methods were applied for retinal vessel classification. Graphs were represented using links and nodes. Object detection was done from different images after extracting unique features. To detect damaged parts, the MSER algorithm was used. Classified images and extracted features were considered as input. Row and column-wise values of the image were compared. Any part having maximum value was considered diabetes. The proposed methodology showed 88% accuracy. Iris images and machine learning techniques were used in Ref. [19] to diagnose type 2 diabetes. For this purpose, 338 subjects were considered, 180 out of them were diabetic, and 158 were non-diabetic patients. Subjects were selected on three factors that include: gender ratio, standard deviation, and diabetes duration age (vary from 1 to 25 years), and average age. Iris images were attained using I-SCAN-2. Gray infra-red images of size (left and right iris) 640 × 480 were acquired. Using the iris image, suitable features were extracted from regions of interest. Inner and outer boundaries of iris were used in segmentation. Rubber-sheet normalization was used to plot extracted iris into a fixed rectangle. Region of interest was cropped from iris according to the tail, head, and body pancreas organ. A threshold was then applied to generate the edge map. Centre point and radius of pupil were considered as main parameters. For each feature, the scoring criteria was calculated. Different machine learning algorithms were used by researchers for classification. Those algorithms include SVM, Naïve Bayes, Random Forest, NN, Adaptive boosting model, and generalized linear model. A decision support system was proposed in [35] that used the AdaBoost algorithm with Decision Stump as a base classifier for classification. The proposed methodology was implemented in four different phases. Local and global dataset collection was performed. The global dataset was used for training and testing, a local dataset was used. The dataset used for this research was collected from various places in Kerala, India. Pima Indian Diabetes Dataset was considered as a global dataset, while the dataset collected from Kerala was considered as a local dataset. Missing values in the local dataset were fulfilled by replacing them with the mean value. In the second phase, AdaBoost algorithm was applied to a global dataset to train the proposed system. Different base classifiers (SVM, NB, Decision Stump, and DT) were also used along with the AdaBoost algorithms. In the third phase, the validation of the proposed system was achieved using local dataset. Finally, the accuracy of AdaBoost algorithm with base classifiers was calculated. AdaBoost algorithm with

Decision Stump as a base classifier showed the best accuracy of 80.729% for diabetes prediction. Also, it showed less error rate. Table 2 briefly explains different machine learning approaches for a diabetes diagnosis. ANFIS [1] found to be more accurate than others, but with a smaller dataset.

Many researchers used Pima Indian Diabetes Dataset to test and train their proposed techniques. Here is the graphical representation of their results in terms of accuracy using the same Pima Indian Diabetes Dataset (PIDD). Fig. 3 shows the best accuracies achieved using different approaches for diabetes prediction, among them Deep Neural Network [24] achieved the highest accuracy. It has four hidden layers, and different combinations of no. of neurons are applied to hidden layers to achieve the best result. Along with that, SVM [38] also achieved the best accuracy as there is a slight difference in their accuracies. 4. Conclusion Among different challenges in the medical diagnosis system, diabetes detection is one of the major technical challenges. Early diagnosis of diabetes is important as delayed detection may lead to different diseases that include kidney failure, stroke, blindness, heart attacks, and lower limb amputation. Machine learning techniques have been introduced in the medical diagnosis system as they have proven to be accurate in diagnosis, successful in treatments, and more cost-efficient. Deep learning is a subset of machine learning in AI, which has the capability of self-learning from data. It is also capable of unsupervised learning. It can learn large amounts of unstructured and unlabelled data that for the human brain may take years to understand. This research is done on existing techniques to perform a survey of the diagnosis of diabetes.

V. Conclusion

This study includes papers from the last decade. Diabetes is one of the fatal diseases, and its early and accurate detection is important to save humans from other fatal effects. Many researchers have proposed methodologies that showed better results in terms of accuracy. Among all approaches which use the same dataset (PIDD) for their model training and testing, Deep Neural Network performed better. On the contrary, DNN has shortcomings like it requires more computational time and frequent adjustment of parameters [15]. It is a well-known fact that deep learning performs more accurately on image datasets. Therefore, image datasets should also be considered for a diabetes diagnosis. Hence, a model must be introduced in the future that should be able to overcome these issues. These different methods (PIDD, ECG, E-Nose and facial images, etc.) for diabetes diagnosis have the clinical advantage as there is no need for a blood sample for diabetes diagnosis as using these methods, diabetes can be diagnosed without any pain. As deep learning is getting more attention nowadays, deep learning should be combined with different algorithms to achieve better accuracy as deep learning can learn large amounts of unstructured and unlabelled data that even the human brain might take years to understand. Also, as there are only a few diabetes datasets available on the internet, so more public data should be available to do research. More research should be performed using Heart Rate as it takes less bandwidth, and its computational complexity is also less. It can also be used in the cloud or mobile devices. HR signals should also be used to detect other cardiac diseases.

References

[1]. Swain Aparimita, Mohanty Sachi Nandan, Das Ananta Chandra. Comparative risk analysis on prediction of diabetes mellitus using machine learning approach. IEEE 2016:3312–7.

[2]. Nirmala Priya Shirley Muller and M. Diagnosis of gestational diabetes mellitus using radial basis function. IEEE 2016;4.

[3]. Sneha Joshi Miss, Borse Megha. Detection and prediction of diabetes mellitus using back-propagation neural network. IEEE 2016;4.

[4]. Erkaymaz Okan, Ozer Mahmut. Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes. Sci. Direct 2016:8.

[5]. Shu Ting, Zhang Bob, Tang YY. Using k-NN with weights to detect diabetes mellitus based on genetic algorithm feature selection. IEEE 2016:6.

[6]. Ali Mohebbi, Aradottir Tinna B, Johansen Alexander R, Bengtsson Henrik, Fraccaro Marco, Mørup Morten. A deep learning approach to adherence detection for type 2 diabetics. IEEE 2017:4.

[7]. Hariyanto Riyanarto Sarno, Rehman Wijaya Dedy. Detection of diabetes from gas analysis of human breath using E-nose. IEEE 2017:6.

[8]. Musale Reena, Paithane AN. Design and develop an algorithm for a diabetic detection using ECG signal. IEEE 2017:6

[9]. Mangrulkar RS. Retinal image classification technique for diabetes identification. In: 2017 Int. Conf. Comput. Methodol. Commun. ICCMC Erode IEEE; 2017. p. 961–6.

[10]. Chen Wenqian, Chen Shuyu, Zhang Hancui, Wu Tianshu. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In: 2017 8th IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS Beijing IEEE; 2017. p. 386–90.

[11]. Mercaldo Francesco, Nardone Vittoria, Santone Antonella. Diabetes mellitus affected patients' classification and

diagnosis through machine learning techniques. Sci. Direct 2017; 112:2519–28.

[12]. Das Tarak, Guha Sayanti, Ghosh Arijit, Basak Piyali. Early detection of diabetes based on Skin impedance Spectrogram and heart rate variability noninvasively. In: 2017 1st Int. Conf. Electron. Mater. Eng. Nano-technol. IEMENTech Kolkata IEEE; 2017. p. 1–5.

[13]. Zheng Tao, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Sci. Direct Jan. 2017; 97:120–7

[14]. Jahangir Maham, Afzal Hammad, Ahmed Mehreen, Khurshid Khawar, Nawaz Raheel. An expert system for diabetes prediction using auto-tuned multilayer perceptron. In: IEEE, vol. 2017 intelligent systems Conference (IntelliSys). London: IEEE; 2017. p. 722–8.

[15]. Wei Sidong, Zhao Xuejiao, Miao Chunyan. A comprehensive exploration to the machine learning techniques for diabetes identification. In: 2018 IEEE 4th world forum internet things WF-IoT IEEE; 2018. p. 291–5.

[16]. El_Jerjawi Nesreen Samer, Abu-Naser Samy S. Diabetes prediction using artificial neural network, vol. 478. Springer; Oct. 2018. p. 12.

[17]. Wu Han, Yang Shengqi, Huang Zhangqin, He Jian, Wang Xiaoyi. Type 2 diabetes mellitus prediction model based on data mining. Sci. Direct 2018; 10:100–7.