2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-7433-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRITO56286.2022.9964904

DT-DL Based Hybrid Approach for Early Detection of Diabetes Using PIMA Dataset

Huma Naz
Department of Computer Science, School of Computer Science
University of Petroleum and Energy Studies
Dehradun, India
huma.naz@ddn.upes.ac.in

Dr. Rahul Nijhawan
Department of Computer Science, School of Computer Science
University of Petroleum and Energy Studies
Dehradun, India
rahul.nijhawan@ddn.upes.ac.in

Dr. Neelu Jyothi Ahuja
Department of Computer Science, School of Computer Science
University of Petroleum and Energy Studies
Dehradun, India
neelu@ddn.upes.ac.in

***Abstract*— According to the International Diabetes Foundation (IDF) 2019 report, 463 million people are living with diabetes, which is likely to increase to 700 million by 2045. Diabetes Mellitus is a metabolic disease with a persistent increase in pervasiveness. Therefore, it is one of the most severe challenges in developed and developing countries. Detection of diabetes is critical in the early phase so that the progression of the disease can be stopped at a definite complication. Therefore, it is mandatory to build an automatic tool that can predict diabetes at an early stage. The machine learning approach has been proven promising for detecting diabetes accurately and is hence used in many diabetes detection processes. This study presents the practical hybrid approach by combining two machine learning algorithms (Decision tree, Deep learning) that may provide prediction and early detection of diabetes securely and effectively. This study's result confirms that the proposed framework can be applied for early diabetes detection with ID3 and J48 algorithms to analyze the performance. The accuracy achieved by these algorithms was recorded as 94.6% and 88.2%, respectively. Additionally, results are compared with the proposed DT-DL hybrid classification approach, which shows that our proposed approach outperforms other traditional approaches with an accuracy of 96.62%. The outcome of this study shows that the hybrid approach of DT-DL provides the most promising results with the best-extracted features.**

Keywords- Diabetes prediction, Machine Learning, Hybrid DT-DL approach, Decision Tree, Deep learning, PIMA Indian dataset

I. Introduction

Diabetes is considered one of the major health problems by WHO and other leading health agencies, and it affects many people worldwide. It can be considered the most common and non-communicable disease in the twenty-first century (diabetes prediction). According to the WHO, 18 million people die each year due to cardiovascular disease, and diabetes is one of the major influencing factors of CHD (coronary heart disease). WHO has also advised adding a progressive strategy for Socioeconomic change in society which shows that diabetes can be prevented mainly by introducing modern and western lifestyles. According to the 8th atlas statistics of the international diabetes federation (IDF), diabetes is at risk globally [1]. As per the 2017 report of IDF, a patient is predicted to die from diabetes impediments every seventh second, and it's increasing at a very high pace [2]. The pervasiveness of diabetes is expected to rise to 9.9% by 2045.

Diabetes is a metabolic disease in which glucose does not metabolize in the body. The human body consists of insulin-named hormones responsible for transforming the starches, sugar and other food aspects into required energy for the body [3]. Diabetes can cause long-span impairment and dysfunction of the pancreas's beta cells, which can further harm the eyes, nerves, kidneys, heart, and blood vessels of the human body and invites numerous other disorders directly or indirectly [4]. Attention must be paid to diabetes because it may lead to macro-vascular and micro-vascular complications, which can drain the quality of life. Along with this, it adversely impacts the developed and developing countries' economies and throughput.

As per the official statistics, the affected people counted by this disease were nearly 110 million in 2017. Diabetes Mellitus (DM) can be characterized into two types, first is type 1 diabetes which arises due to the pancreas' failure to produce insulin and requires insulin injections for treatment[4]. The second type of DM is type 2 diabetes which can be treated by medication and typically accounts for 90% of the cases worldwide. The prevalence of diabetes in women is estimated to be 9.0% and 9.6% in men. It is expected to increase to a prevalence level of 19.9% in people 65-79 years of age. Figure 1 demonstrates the rate of diagnosed, undiagnosed and total diabetic persons according to the national statics report.

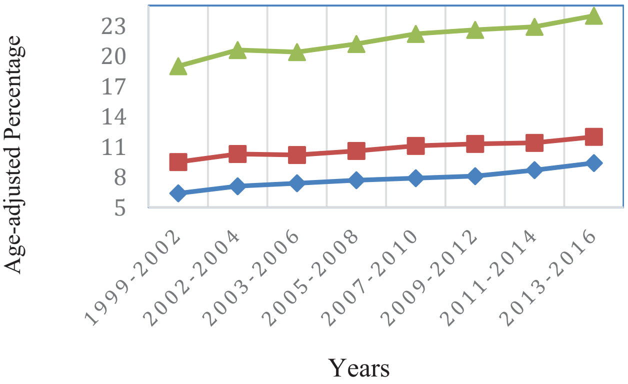

Fig. 1. Prevalence of total, diagnosed, undiagnosed diabetes from 1999-2016 [2]

The main aim of this research work is to propose a predictive tool using a hybrid approach for early diabetes detection. Over the years, the machine learning technique has proven to be very useful for extracting hidden patterns [5]. Hence, this research paper presents a novel hybrid approach combining two machine learning algorithms (Decision tree and Deep learning) with an accuracy of 93.24% for diabetes detection onset. Where Deep Learning

978-1-6654-7433-7/22/$31.00 ©2022 IEEE

(DL) learns the representation from the more complex relationship of data and provides more efficient results, and a Decision tree (DT) is an algorithm which is easy to implement as a predictive model to produce more accurate results for statistics data [6].

The rest of the paper is ordered in the following manner: Section 2 presents the related work done on diabetes detection using an integrated approach. Section 3 of the paper puts forth the description of the dataset and methodology design. Section 4 presents the experimental outline used to examine the model's performance. Finally, the paper concludes with section 5.

II. Literature Review

Over the year, much research has been done on diabetes detection using machine learning algorithms. Except for machine learning's numerous excellent features and fantastic generalization capabilities, machine learning algorithms are widely used as a prediction model for disease prediction and detection. Based on recent diabetes prediction studies, it's been observed that the PIMA Indian dataset (PID) is a commonly used dataset primarily collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). However, data mining and machine learning technologies have been used for a decade. These algorithms combine statistical analysis to extract hidden features and predict the outcome more precisely.

According to Patilet al. [7], a hybrid approach can be proven helpful for efficient decision-making and disease prediction. Therefore, the author suggested a hybrid model that combines the C4.5 and K-means clustering algorithm to build an ensemble classifier model that can approve the functional dataset class label and provide a classification accuracy of 92.38%. In the previous year of diabetes research, numerous scholars have shown the potential of prediction classification for medical officials and practitioners. Therefore, Kandhasamy et al. [8] compared four different as well as standard classifiers named K-Nearest Neighbors (KNN), J48, Support Vector Machine (SVM) and Random Forest (RF) for diabetes prediction. The achieved outcomes from all four classifiers presented that the J48 classification algorithm outperforms with a 73.48 accuracy rate before the data preprocessing, and Random Forest, along with KNN(k=1), gives better outcomes after data preprocessing.

Ahmed et al. [9] compared the outcomes of the multilayer perceptron model in a neural network with J48 and ID3 Decision tree classifiers. The outcome reveals that the J48 classifier performs better after the pruning process and gives an accuracy rate of 89.3%. Then Cedenoet. al [10] presented a novel approach for diabetes detection using artificial met plasticity on a multilayer perceptron neural network with an accuracy rate of 89.93%. All the presented studies were carried out on PID as investigational data. According to recent studies by numerous researchers' data preprocessing is a significant aspect to consider for better diabetes prediction outcomes.

The selection of parameters and preprocessing methods is vital to obtain more accurate results. Therefore, Vijayan V. [11] studied different preprocessing techniques and suggested that principal component analysis (PCA) and discretization methods are more convenient for diabetes detection. The author reviewed that preprocessing improves the outcomes of the Decision Tree (DT), and Naive Bayes (NB) classifier, while the SVM outcomes were diminished. Based on the literature cumulative study, it has been observed that presented outcomes are not worthy in the field of classification and disease prediction. Besides these mentioned studies, Polat & Gunes [12] have proposed a novel approach for early diabetes prediction using fuzzy logic, PCA and a Neuro-fuzzy inference system for detecting lymph disease with a prominent Accuracy of 88.83%.

According to Pang et al. [13], SVM is a widely used and demanding classifier nowadays and works efficiently for linear and nonlinear datasets. Therefore, the author proposed the Relief F-SVM-based approach for breast tumour diagnosis and prediction with a prominent accuracy rate of 90.0%. The specificity and sensitivity rates for the proposed prediction model were 98.7% and 73.8%. Sangaiah & Kumar [14] proposed a novel hybrid approach for breast cancer disease detection using an entropy-based genetic algorithm and Relief classifier. The outcomes achieved by this classifier were commendable, with reduced dimensionality and a higher accuracy rate. Shunye Wang [15] projected an improved prediction model using the k-means clustering algorithm using the Huffman tree structure. Several researchers have worked on detecting diabetes with the PIMA dataset to date. Thus Table 1 displays some of the excellent work done on the Pima dataset and our proposed method. Our proposed work achieved a rate of 96.62% on PID, which is also shown in Table 1.

TABLE I. COMPARISON OF EXISTING APPROACH WITH PROPOSED METHOD

Authors	Methods	Accuracy obtained (in %)
Ref No. [16]	Firefly and Cuckoo Search Algorithms	81%
Ref No. [7]	K-means Clustering, C4.5 Algorithm	92.38%
Ref No. [8]	J48 DT, KNN, Random Forest, and SVM	73.48%
Ref No. [9]	Multilayer Perceptron model, J48, ID3	89.3%
Ref No. [4]	Feedforward NN.	82%
Ref No. [17]	NB	79.56%
Ref No. [18]	SVM	78%
Ref No. [19]	LDA – MWSVM	89.74%
Ref No. [20]	Neural Network with Genetic Algorithm	87.46%
Ref No. [21]	K-means and DT	90.03%
Ref No. [22]	PCA, K-Means Algorithm	72%
(Proposed Work)	*DT-DL Based Hybrid approach, ID3, J48((Highest accuracy achieved using DT-DL)*	*96.62%*

III. METHODOLOGY

A. *Dataset Description*

Various experiments have been done using the PIDD. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney (NIDDK). PID was taken from the UCI ML repository for this work [23]. The reason for selecting this dataset is that most people in modern times live an identical lifestyle that includes a high reliance on processed food and declining physical activities. PIMA is a group of Native Americans who lived in an area now known as Central Arizona. Due to their genetic predisposition, they could survive on low carbohydrates for many years. However, in the recent past, PIMA group suddenly shifted from their traditional diet towards processed food, followed by decreased physical activity. Consequently, they were detected with high levels of Type II diabetes, so since 1965, their health data have been used in many diabetes studies.

PIDD includes a certain number of medical predictors and one variable target. The predictor variables are the number of pregnancies, BMI, Blood Pressure, skin thickness, insulin level, age, Glucose, and Diabetes Pedigree Function shown in table 2. All the participants in the PIDD study are females up to the age of 21. The dataset has 768 instances divided into 268 non-diabetic and 500 diabetic instances. The target variable identifies whether a person is non-diabetic (represented by 0) or diabetic (represented by 1). The description of different parameters of each attribute in the dataset, including max value, min value, standard deviation, missing value, mean and median, are given in table 3.

TABLE II. DESCRIPTION OF PID ATTRIBUTES

S No	Predicators	Description of Predicators	Unit
1.	Preg	Number of times a female participant is pregnant	-
2.	Plasma glucose	Glucose concentration in 2 hours in an oral glucose tolerance test	Mg/dl
3.	Diastolic blood pressure	Diastolic blood pressure (upper blood pressure)	mmHg
4.	Triceps skinfold thickness	Skin thickness of participants in mm The collagen content concludes it	Mm
5.	Insulin	Participant's 2-Hour serum insulin	Mm U/Ml
6.	Body Mass Index	Weight of a participant in Kg/ (height in m) ^2)	Kg/m2
7.	Diabetes pedigree function	Appealing attributes used for Diabetes diagnosis	-
8.	Age	Age of participants	-
9.	Outcome	Diabetes onset with diabetic and non-diabetic patients	-

TABLE III. DETAILED DESCRIPTION OF PID ATTRIBUTES

S No.	Predicators	Missing Values	Mean	Std Dev	Range	Data Type
1.	Pregnancy	0	3.845	3.370	0-17	INTEGER
2.	Glucose	0	120.89	31.973	0-199	INTEGER
3.	Diastolic Blood pressure	0	316.56	1096.927	0-122	INTEGER
4.	Skinfold Thickness	0	51.697	88.690	0-99	INTEGER
5.	Insulin	0	819.49	3873.732	0-846	INTEGER
6.	Body Mass Index	0	60.769	92.015	0-67.1	REAL
7.	Diabetes pedigree function	0	0.472	0.472	0.078-2.42	REAL
8.	Age	0	33.241	11.760	21-81	INTEGER
9.	Outcome	(Diabetic instances – 268) (Non-diabetic instances – 500) (Total instances – 768)				POLYNOMIAL

B. Data-Preprocessing

Pre-processing is the method applied to the dataset before commencing its processing. Typically, preprocessing modifies the raw data, which can enhance the classification ability of processing. It also includes detecting and removing outliers where outliers are the inconsistent type of data. It performs attribute extraction, normalization, integration, aggregation, and discretization [24]. The dataset consists of 8 attributes with 768 total instances, considered to be analyzed using our proposed approach.

However, it is being analyzed that the dataset contains inconsistent values, which are processed by applying diverse available preprocessing methods to recover the data quality. Therefore, a data-cleaning operation was performed on the dataset to remove the outliers and integrate the various attributes. In this paper, we used traditional approaches to remove the outliers and integrate the attributes, as shown in Figure 2. After removing noise and inconsistent data, the SMOTE algorithm was applied to PIDD to eliminate the class imbalance problem [25].

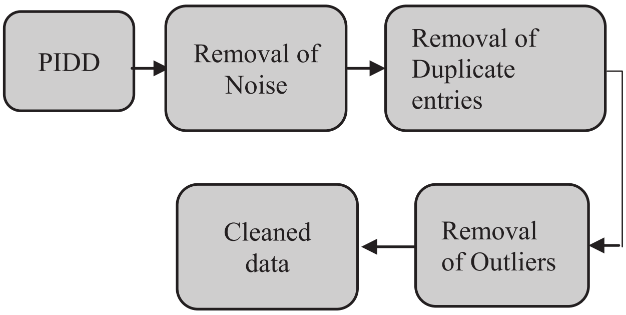

Fig. 2. Steps for traditional preprocessing of data

After passing through the traditional pre-possessing steps, we had the processed data, whose Mean and standard deviation (SD) were lesser and different from the unprocessed data, as shown in Table 4. Lesser Mean and SD implied that data was clustered rather than spread out. In our approach, the first step of preprocessing was the removal of noise, and this step is performed by eliminating the irrelevant data which did not secrete its attribute. The second step was the removal of duplicate values to eliminate identical data to reduce the overhead. The last step was removing the outliers, which was done using diverse available DM techniques like DBSCAN and Z-score.

TABLE IV. DETAILED DESCRIPTION OF PIDD ATTRIBUTES AFTER PREPROCESSING

S No	Predicators	Mean	Std Dev	Data Type
1.	Pregnancy	0.845	0.362	INTEGER
2.	Glucose	120.89	30.624	INTEGER
3.	Diastolic Blood pressure	70.775	12.327	INTEGER
4.	Skin Thickness	29.241	10.553	INTEGER
5.	Serum Insulin	153.63	111.361	INTEGER
6.	BMI	32.762	6.497	REAL
7.	Diabetes pedigree Function	0.527	0.338	REAL
8.	Age	30.646	10.050	INTEGER

C. *Proposed Approach for Diabetes Detection*

The primary purpose of this proposed research study is to extract the most prominent features with the help of the proposed hybrid approach that is needed to detect diabetes at an initial stage. The projected architecture contains three main parts; the first part is the data preprocessing part with sampling methods explained in the above section. The next part is the hybrid approach of DT and DL for the classification and prediction of diabetes. The final part of the methods shows the outcome achieved by using a hybrid approach. Fig 3 shows the flow chart of the proposed architecture.

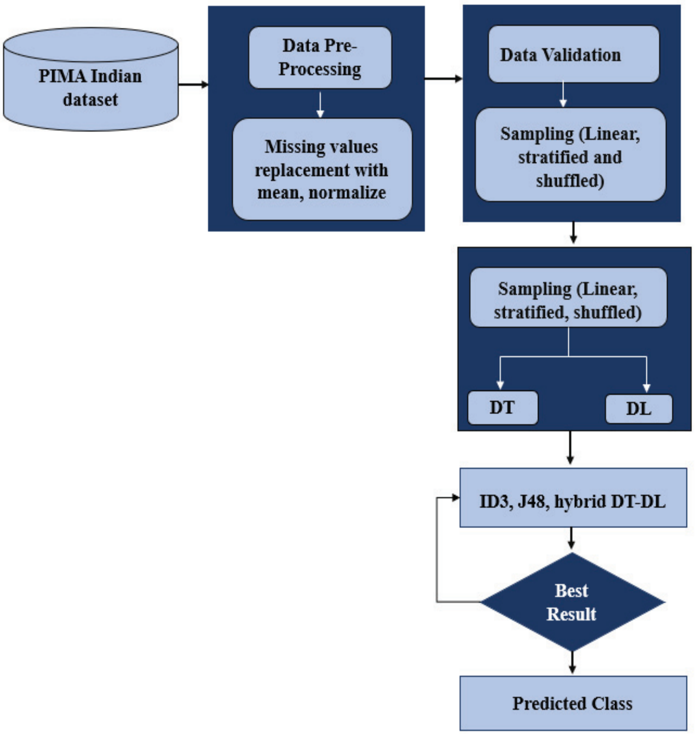
PIMA Indian dataset
Data Pre-Processing
Missing values replacement with mean, normalize
Data Validation
Sampling (Linear, stratified and shuffled)
Sampling (Linear, stratified, shuffled)
DT
DL
ID3, J48, hybrid DT-DL
Best Result
Predicted Class

Fig. 3. A proposed architecture for Diabetes Detection

The proposed methodology combines the Deep learning and Decision tree algorithm used in several previous research studies to evaluate the risk factors in diseases like diabetes. It is being observed that the performance of machine learning algorithms tends to decrease with crucial artificial intelligence problems like speech recognition and text classification. Therefore, the shortcoming of these Machine Learning (ML) algorithms is that they improve the demands of the DL methods, giving better disease prediction results [26]. DL has various applications in healthcare, and DT is a very powerful machine-learning model that works well with relational datasets. The clarity of representation in ML nodes makes the decision tree a distinct classifier. In our proposed architecture, the hybrid approach of DL and DT is operated on the PID, and it takes eight inputs of the PID and one output as the predicted outcome that shows whether a person has diabetes or not. The individual characteristics of each algorithm are represented here.

D. *Deep Learning & its Architecture*

The multilayer perceptron model is used in this research work for evaluating the risk factor in diabetes. In a multilayer perceptron, each layer processes the information received from the preceding layer and passes it to the next layer. The neurons in the input layer are directly connected to the input feature values in the dataset from which learning has to be performed. The number of neurons in the input layer corresponds to the number of input features in the dataset. The output of each hidden layer neuron and output layer neuron is the summation of the product of the preceding layer neurons with the connected weights [27]. The activation function defines the neuron functions. The activation functions are used in the hidden and output layer. The hidden layer implements the nonlinear activation function, whereas the output layer implements the linear activation function. The ANN generates a response, determining the degree of adjustment required in the network parameters. The weight adjustment is performed by calculating the mean square error in the network. This adjustment is known as learning or training the network [28].

DL network optimization performs a significant role in better outcomes, and DL performs the optimization on its network by tuning the parameters with their behaviour. Several parameters are available for DL optimization, including adaptive learning rate, root means square error, momentum, Dropout, and L1 and l2 regularization, as shown in Table 5.

TABLE V. KEY PARAMETERS OPTIMIZED IN DL NETWORK IMPLEMENTATION

Layers	Units	Type	Momentum	Mean Weight	Weight RMS	Mean Bias	Bias RMS
1	8	Input	-	-	-	-	-
2	50	Maxout	0.002166	0.000000	-0.002163	0.194934	0.489399
3	50	Maxout	0.002624	0.000000	-0.004571	0.140682	0.995029
4	2	Softmax	0.001182	0.000000	0.022341	0.433838	0.000000

E. *Decision Tree*

As the name implies, the DT classifier works like a structure of the tree, its internal nodes test each attribute of the dataset, and the final node represents the outcome of the DT [29]. DT is the most basic and efficient classification method for prediction. The attribute with the highest weight is considered to be the tree's root, and all the internal nodes are working to apply a decision on the considered root node. DT is a simple and easily interpretable classification problem, although noise and outliers in the training data can cause overfitting in the dataset. When DTs are built using training data, many branches over expanded due to noise. Therefore it is essential to offer preprocessed data to DT for better results.

IV. EXPERIMENTAL RESULTS & ANALYSIS

In our proposed method, the results are achieved by applying a hybrid classification approach that combines the

DL and DT algorithms. The outcomes are compared with the traditional decision tree algorithms (J48 and ID3 algorithms) to display maximum accuracy in diabetes prediction. J48 is an algorithm which is also known as C4.5, developed by Ross Quinlan. It is an extension of an earlier proposed ID3 algorithm [30]. The J48 algorithm is often referred to as a statistical classifier, and ID3 (Iterative Dichotomiser 3) is a precursor of the C4.5 algorithm.

Moreover, this algorithm generates the decision tree by fixed sets of nodes. Compared with these three classifiers, our proposed hybrid approach provides promising accuracy (96.62%) that can further be applied to develop a prominent tool for diabetes detection onset. Further, it can help the healthcare practitioner and be the second estimation for improving decisions depending on extracted features.

A. Performance Metrics

Accuracy($\mathbf{A_C}$): Accuracy of a measurement is the ratio of correctly classified observations to the total number of observations [31] or accuracy can be the closeness of a measured value to a standard or known value. Accuracy can be measured using equation 1. Accuracy is a good performance measure when target variable class data is nearly balanced.

$$A_C = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \qquad (1)$$

Here, true positives are abbreviated as TP, true negatives abbreviated as TN, false-positive abbreviated as FP and false-negative abbreviated as FN.

Sensitivity ($\boldsymbol{S_V}$): S_V is a proportion of true positives correctly classified [32]. Sensitivity is a performance measure that evaluates the true positives from each class label and can be measured using equation 2.

$$S_V = \mathrm{RC} = TPR = \left(\frac{TP}{TP + FN}\right) \qquad (2)$$

Specificity ($\mathbf{S_p}$): S_p is a proportion of true negatives correctly classified [33]. Specificity is an evaluation metric that measures the true negatives from each class label and can be measured using equation 3.

$$S_p = \left(\frac{TN}{TN + FP}\right) \qquad (3)$$

Table 6 shows the Sensitivity, Specificity, Accuracy, Precision, Recall, and F-Measure. As shown in table 6, the proposed DT-DL scheme outperforms all the performance measures and gives the best results for diabetes onset with an accuracy rate of 96.62 %. Fig 4 displays a graphical comparison between the Evaluation criteria of the diabetes detection technique.

TABLE VI. COMBINED PERFORMANCE MEASURES OF THE PROPOSED APPROACH

MEASURES	METHODS		
	DL - DT	ID3	J48
Sensitivity (%)	96.62	91.12	82.12
Specificity (%)	94.02	86.41	80.23
Accuracy (%)	95.45	94.46	88.51
Precision (%)	94.72	94.35	89.42
Recall (%)	97.86	86.35	80.13
F-Measure (%)	94.03	85.16	76.15

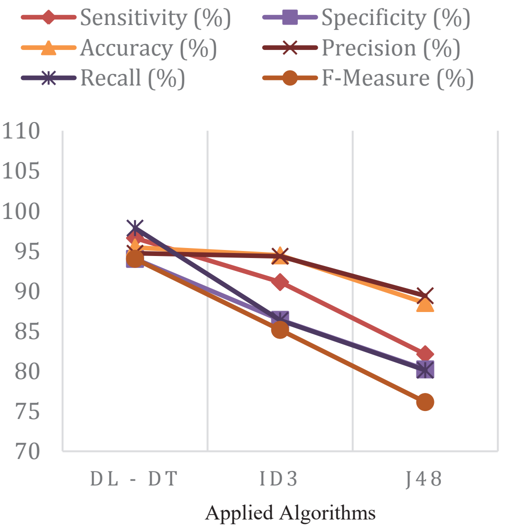

Fig. 4. Comparison Performance for Various Diabetes Prediction Schemes

V. CONCLUSION & FUTURE SCOPE

This study presents the DT-DL-based hybrid approach for detecting diabetes using the PIMA diabetes dataset. Diabetes has been listed as a significant cause of death in developing and developed countries and invites a lot of other diseases like brain stroke, kidney failure, etc. therefore; early diabetes detection becomes a significant domain of research. In this manuscript, we proposed a combination of DT and DL algorithms to predict diabetes at an initial phase. The outcome achieved on the PIMA dataset is higher than other proposed methodologies on the same dataset, as shown in Table 6. The proposed hybrid approach outperforms traditional machine learning algorithms with an accuracy rate of 96.62%. In the future, we intend to increase the dataset size to make our study more robust and develop a flexible predictive tool in the form of an app that can use the proposed DT-DL algorithm to help healthcare specialists in early diabetes detection.

REFERENCES

[1] Cuschieri, S., 2019. Type 2 diabetes–An unresolved disease across centuries contributing to a public health emergency. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 13(1), pp.450-453.

[2] "IDF Diabetes Atlas, 9th edition", January 23,2019. Available at: https://www.diabetesatlas.org/en/resources/

[3] "Global Report on Diabetes", April 21, 2016. Available at: https://www.who.int/publications/i/item/9789241565257

[4] Zhang, Y., Lin, Z., Kang, Y., Ning, R. and Meng, Y., 2018. A feed-forward neural network model for the accurate prediction of diabetes mellitus. International Journal of Scientific and Technology Research, 7(8), pp.151-155.

[5] Zhang, L.M., 2015, October. Genetic deep neural networks using different activation functions for financial data mining. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 2849-2851). IEEE.

[6] Naz, H. and Ahuja, S., 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19(1), pp.391-403.

[7] Patil, B.M., Joshi, R.C. and Toshniwal, D., 2010. Hybrid prediction model for type-2 diabetic patients. Expert systems with applications, 37(12), pp.8102-8108.

[8] Kandhasamy, J.P. and Balamurali, S.J.P.C.S., 2015. Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, pp.45-51.

[9] Ahmad, A., Mustapha, A., Zahadi, E.D., Masah, N. and Yahaya, N.Y., 2011, July. Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus. In International conference on digital information processing and communications (pp. 537-545). Springer, Berlin, Heidelberg.

[10] Marcano-Cedeno, A., Torres, J. and Andina, D., 2011, May. A prediction model to diabetes using artificial metaplasticity. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 418-425). Springer, Berlin, Heidelberg.

[11] Vijayan, V.V. and Anjali, C., 2015, April. Decision support systems for predicting diabetes mellitus—A Review. In 2015 Global conference on communication technologies (GCCT) (pp. 98-103). IEEE.

[12] Polat, K. and Guneş, S., 2007. Automatic determination of diseases related to lymph system from lymphography data using principles component analysis (PCA), fuzzy weighting preprocessing and ANFIS. Expert Systems with Applications, 33(3), pp.636-641.

[13] Peng, H. and Fan, Y., 2017. Feature selection by optimizing a lower bound of conditional mutual information. Information Sciences, 418, pp.652-667.

[14] Sangaiah, I. and Kumar, A.V.A., 2019. Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction. *Cluster Computing*, *22*(3), pp.6899-6906.

[15] Yao, H., Duan, Q., Li, D. and Wang, J., 2013. An improved K-means clustering algorithm for fish image segmentation. Mathematical and Computer Modelling, 58(3-4), pp.790-798.

[16] Haritha, R., Babu, D.S. and Sammulal, P., 2018. A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms. International Journal of Applied Engineering Research, 13(2), pp.896-907.

[17] Iyer, A., Jeyalatha, S. and Sumbaly, R., 2015. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

[18] Kumari, VA and Chitra, R., 2013. Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), pp.1797-1801.

[19]

[20] Calisir, D. and Dogantekin, E., 2011. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7), pp.8311-8315.

[21] Dadgar, S.M.H. and Kaardaan, M., A Hybrid Method of Feature Selection and Neural Network with Genetic Algorithm to Predict Diabetes.

[22] Chen, W., Chen, S., Zhang, H. and Wu, T., 2017, November. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In 2017 8th IEEE International conference on software engineering and service science (ICSESS) (pp. 386-390). IEEE.

[23] Patil, R.N. and Tamane, S., 2017. A novel scheme for predicting type 2 diabetes in women: using kmeans with PCA as dimensionality reduction. Int J Comput Eng Appl XI (VIII), pp.76-87.

[24] "Machine Learning: Pima Indians Diabetes",April 14,2018 . Available at: https://www.andreagrandi.it/2018/04/14/machine-learning-pima-indians-diabetes/.

[25] Jalili, M. and Niroomand, M., 2016. Type 2 diabetes mellitus. Tintinalli's Emergency Medicine, 7.

[26] Wu, Han, et al. "Type 2 diabetes mellitus prediction model based on data mining." Informatics in Medicine Unlocked 10 (2018): 100-107.

[27] Swapna, G., Vinayakumar, R. and Soman, K.P., 2018. Diabetes detection using deep learning algorithms. ICT express, 4(4), pp.243-246.

[28] Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., 2018. Informatics in medicine unlocked.

[29] Davazdahemami, B. and Delen, D., 2019. The confounding role of common diabetes medications in developing acute renal failure: A data mining approach with emphasis on drug-drug interactions. Expert Systems with Applications, 123, pp.168-177.

[30] Pei, D., Zhang, C., Quan, Y. and Guo, Q., 2019. Identification of potential type II diabetes in a Chinese population with a sensitive decision tree approach. Journal of diabetes research, 2019.

[31] Mantovani, R.G., Horváth, T., Cerri, R., Junior, S.B., Vanschoren, J. and de Carvalho, A.C.P.D.L.F., 2018. An empirical study on hyperparameter tuning of decision trees. arXiv preprint arXiv:1812.02207.

[32] Salzberg, S.L., 1994. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.

[33] Centers for Disease Control and Prevention, 2020. National diabetes statistics report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services, pp.12-15.

[34] Chen, W., Chen, S., Zhang, H. and Wu, T., 2017, November. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In 2017 8th IEEE International conference on software engineering and service science (ICSESS) (pp. 386-390). IEEE.