Diabetes Prediction Using Machine Learning Analytics

Conference Paper · May 2022

DOI: 10.1109/COM-IT-CON54601.2022.9850922

5 authors, including:

Reshmi S.
National Institute of Technology, Silchar
3 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Saroj Kr. Biswas
National Institute of Technology, Silchar
172 PUBLICATIONS **1,803** CITATIONS
SEE PROFILE

Arpita Nath Boruah
Assam down town University
24 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Dalton meitei Thounaojam
National Institute of Technology, Silchar
48 PUBLICATIONS **717** CITATIONS
SEE PROFILE

Diabetes Prediction Using Machine Learning Analytics

S. Reshmi
Computer Science and Engineering
NIT Silchar
Silchar, India
s_pg_21@cse.ac.in

Saroj Kr. Biswas
Computer Science and Engineering
NIT Silchar
Silchar, India
bissarojkum@yahoo.com

Arpita Nath Boruah
Computer Science and Engineering
NIT Silchar
Silchar, India
arpita.boruah@hotmail.com

Dalton Meitei Thounaojam
Computer Science and Engineering
NIT Silchar
Silchar, India
dalton.meitei@gmail.com

Biswajit Purkayastha
Computer Science and Engineering
NIT Silchar
Silchar, India
biswajit@nits.ac.in

Abstract— **Diabetes, an incurable disease which occurs because of high blood sugar levels over a prolonged time period, requires early prediction to significantly reduce its severity. Now-a-days Machine Learning (ML) community has been working on diabetes prediction and many researches have been done since decades for its prediction. Keeping in view the severity of this disease, this paper introduces a model, named Diabetes Expert System using Machine Learning Analytics (DESMLA), exploring the diabetes data to predict the disease more effectively. The diabetes dataset is imbalance in nature. And therefore, DESMLA model uses 5 most prominent oversampling techniques namely SMOTE, Borderline SMOTE, ADASYN, KMeans SMOTE, Gaussian SMOTE to get rid from this class imbalance problem of diabetes dataset. DESMLA model uses Decision Tree (DT) and Random Forest (RF) as classifiers along with all the data preprocessing steps for diabetes prediction. The experimentation results shows that DESMLA model with KMeans SMOTE and Gaussian SMOTE performs better.**

Keywords—Class Imbalance Problem, Data Mining, Decision Tree, Machine Learning, Random Forest

I. Introduction

Among the top 5 countries globally, India stands second with 69.20 million people with diabetes and another 36.50 million pre-diabetes [1], high-risk diabetes, and cardiovascular disease. Diabetes mellitus [2] also known as diabetes, a ubiquitous disease and has no permanent treatment. The pancreas [3] produces insulin which has a significant role in regulating blood glucose levels. There are three significant diabetes mellitus: Type 1[4], Type 2 [5], and gestational diabetes [6] [7]. Diabetes symptoms differ upon how much the blood sugar is exalted. Type 1 diabetes occurs due to lack of insulin. Symptoms of Type 1 diabetes are mostly severe, which include increased thirst, frequent urination, starvation, weight loss. The person suffering from Type 1 diabetes requires to inject insulin on daily basis. Insulin resistance causes Type 2 diabetes and occasionally combined with an absolute shortage insulin. Following a healthy lifestyle such as a nutritious diet, proper exercise, can help prevent diabetes. Without a previous diagnosis of diabetes, when a pregnant develop a high blood glucose level then it leads to Type 2 diabetes mellitus.

People can make an earlier decision about diabetes mellitus by ML by using their daily physical examination data. The challenges faced for the ML method are how to select the valuable features and the correct classifier to get highly accurate results. Recently, for diabetes prediction, various ML algorithms have been used, like DT [8-11], RF [12-13], Naïve Bayes (NB), Support Vector Machine (SVM) etc. DT is one of the popular ML methods because of its simplicity and transparency. However, RF has a greater classification power in compared to DT as it generates a large number of DTs for prediction reducing the overfitting problem. Therefore a model, Diabetes Expert System using Machine Learning Analytics (DESMLA) is proposed to explore diabetes data to predict the diabetes more effectively. With the high demand of use of ML techniques in the medical fields, an enormous amount of data is collected. The characteristics of the data plays a vital part in performance of ML techniques. Hence the characteristics of the data need to be examined before using any ML techniques. Thus, in the proposed DESMLA, Machine Learning Analytics (MLA) is used to detect diabetes using DT and RF more efficiently. In the proposed DESMLA model, the five most prominent oversampling techniques namely SMOTE [14], Borderline SMOTE [15], ADASYN [16], KMeans SMOTE [17], Gaussian SMOTE [18] are used to get rid from class imbalance problem of diabetes dataset after which feature selection is applied using Pearson's Correlation Coefficient (PCC) and then by using DT and RF diabetes is predicted. Finally from the experimentation it can be concluded that DESMLA with KMeans SMOTE and Gaussian SMOTE works better than others.

This manuscript is distributed into five sections. In section 2 a survey on prediction of diabetes is performed. The Machine Learning methodologies namely DT and RF are illustrated in section 3 followed by section 4 which discusses the results, and finally section 5 draws the conclusion.

II. Literature Survey

Several researchers used ML methods to predict diabetes. Some of them are mentioned in this section.

Alam et al. [19] applied ANN techniques and recorded an accuracy of 76.82%. Canadian Primary Care Sentinel

Surveillance Network and classifier Bootstrap aggregating, Adaptive Boosting, and DT were used by Perveen et al. [20] and they found that Adaboost can predict diseases giving better accuracy. Sisodia et al. [21] showed the comparison of SVM, NB, and DT using PIDD and finally concluded that NB as the better classifier with 76.86% accuracy. After reducing the dimensionality of PIDD, Sivaranjani et al. [22] used SVM and RF to detect diabetes. Tigga et al. [23] used logistic regression on PIDD and found the count of pregnancies, level of glucose and BMI as utmost important. In Diwani et al. [24] Naive Bayes and DT were trained using 10 fold cross-validations. Experimentation showed that NB gave better performance of 76.30% accuracy. Zou et al. [25] done the experimentation on PIDD using RF, DT, ANN as classifier and Minimum Redundancy Maximum Relevance (mRMR) and PCA methods as feature reduction procedure. From the experimentation it was observed that RF with the mRMR feature reduction method gave best performance with 77.21% accuracy.

Kandhasamy et al. [26] compared J48, SVM, RF and K-Nearest Neighbors (KNN). Experimentation were done in two procedure, one by preprocessing and other without preprocessing using 5 fold cross validation. Yuvaraj et al. [27] used RF, DT, and the Naïve Bayes for predicting diabetes. After using Information Gain method to extract the relevant features, they used the classifier for prediction and found that the RF gave the highest accuracy. Boruah et al. [28] proposed a way to find risk factor of Parkinson's disease by using DT. The rules generated from DT are processed to find the important factor, which is/are the main cause of the disease. An enhanced model was put forwarded by new Tafa et al. [29] for predicting diabetes using SVM and NB for the data set acquired as of three distinct locations in Kosovo which consist of 402 patients out of which a total of 80 were diagnosed with diabetes of Type 2 form. The dataset comprises eight attributes. The proposed approach had enhanced to 97.6% which is much better than SVM and Naïve Bayes. Khanam et al. [30] used 7 ML algorithms on PIDD to detect diabetes and concluded that Logistic Regression and SVM worked better in prediction. Boruah et al. [31] put forwarded a methodology to predict Parkinson's disease. In the proposed approach, the dataset was firstly treated for class imbalance problem using Borderline SMOTE, Safe-Level SMOTE and SMOTE, and then using DT Parkinson disease is detected. From the experimentation Borderline SMOTE with DT gives the best accuracy and thus is further processed to find the risk factor of Parkinson's disease.

III. The Proposed Method DESMLA

The proposed model Diabetes Expert System using Machine Learning Analytics (DESMLA) consisting data preprocessing and classification as 2 of its steps. In data preprocessing, the dataset is first preprocessed then the model is trained using the DT and RF. The work flow associated with the proposed DESMLA is as shown in the figure 1.

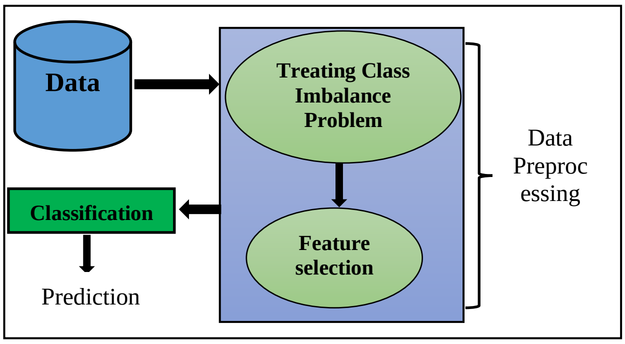

Fig.1. Workflow diagram of DESMLA

A. Data Preprocessing

In this step the data is analyzed and preprocessed to balance the dataset and to select the feature set. This step is subdivided into 2 sub steps: class balancing and feature selection.

1) Treating Class Imbalance Problem

If one of the class is extremely high compared to the other classes present in the dependent variable then it is termed as the class imbalance problem in ML. Which means there is a bias towards the majority class present in the dependent variable. Fraud detection, medical diagnosis, e-mail classification are areas where such data can be found. Hence, to have a proper prediction diabetes, class imbalance must be rectified. There is an assumption of even data distribution within classes in ML algorithms. The extensive issue in the class imbalance problem is that the algorithm will not learn the patterns in the minority class as it does not have enough data leading to high misclassification errors for the minority class.

Inorder rectify class imbalance problem, the proposed DESMLA uses SMOTE techniques namely, borderline SMOTE, ADASYN SMOTE, Means SMOTE, Gaussian SMOTE:

i. **SMOTE:** SMOTE stands for Synthetic Minority Oversampling Technique. The synthetic points are created for data augmentation depending on the original data points. The main advantage of using SMOTE is in the creation of different simulated data points than the original points of data.

ii. **Borderline SMOTE:** Borderline-SMOTE makes generates simulated data between the two classes along the decision boundary.

iii. **ADASYN SMOTE:** ADASYN stands for adaptive synthetic oversampling which is another variation from SMOTE. ADASYN creates synthetic data according to the data density.

iv. **KMeans SMOTE:** It is an effective and straightforward oversampling method based on K-Means clustering and SMOTE that evades noise generation and mitigates imbalanced data in classes.

v. **Gaussian SMOTE:** Gaussian oversampling is based on the Gaussian distribution. The newly generated minority samples are simulated based on the area under Gaussian density function.

2) *Feature Selection*

In statistics, PCC is a bivariate correlation. A threshold of 0.08 is used for PCC and thus the attributes with PCC less than threshold were removed from the dataset.

B. *Classification*

The proposed model DESMLA uses 2 classifiers namely DT and RF.

1. **Decision Tree:** DT is an ML algorithm with a tree-like structure. The internal nodes are represented by the features while the outcome by the leaf nodes. Thus the branches of the tree represent the decision rules
2. **Random Forest:** RF is a collective learning and decision making algorithm that ensembles multiple DTs from a randomly selected subset of the training set and for prediction it depends on the votes from different DTs

IV. Result And Analysis

The experimentation is done in PYTHON 3.0 version on windows 10 environment. The proposed model DESMLA uses the Pima-Indians Diabetes Dataset (PIDD), available in the UCI ML repository. A total of 768 number of patients information along with their corresponding nine unique attributes are there in the dataset out of which 500 are negative and 268 are positive. After applying, SMOTE, Borderline SMOTE, K-Means SMOTE, ADASYN, and Gaussian SMOTE to the original data set the synthetic instances created are as shown in Table I.

TABLE I. Number of Synthetic Instances Created By SMOTE, Boderline Smote, ADASYN, KMEANS and Gaussian SMOTE

Methods	**Instances in train set class 0**	**Instances in train set class 1**	**Instances synthetically formed in minority class**
DESMLA WITH SMOTE	500	268	232
DESMLA WITH BORDERLINE SMOTE	500	268	232
DESMLA WITH ADASYN	500	268	232
DESMLA WITH KMEANS SMOTE	500	268	232
DESMLA WITH GAUSSIAN SMOTE	500	268	232

Fig.2. show the imbalanced data in original dataset. Fig.3. shows the balanced data after treatment of imbalanced data.

In the next sub step of the data preprocessing step, the correlation of the features is extracted using the PCC. Table. II shows the Pearson’s correlation coefficient between input and output attributes. Depending upon the coefficient, the attributes with coefficient lower than threshold were removed from the dataset. Hence, skin thickness, blood pressure, were removed from the dataset and remaining 6 attributes were used for prediction.

TABLE II. Input - Output Attribute Correlation

Attributes	Correlation Coefficients
Glucose level	.4666
BMI	.2926
Insulin	.1305
Pregnancies	.2218
Age	.2383
Skin thickness	.0747
Blood pressure	.0650
Diabetes pedigree function	.1738

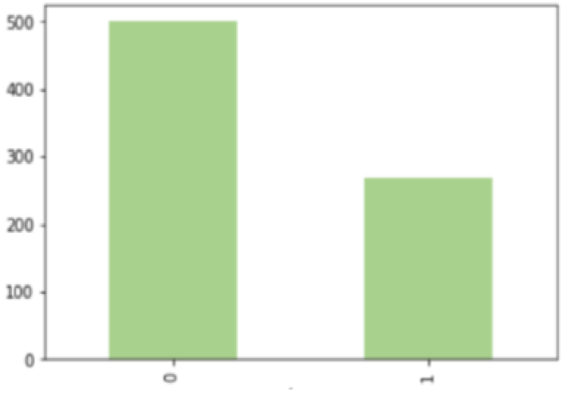

Fig. 2. Class Imbalance Problem in the Original Pima Indians Diabetes Dataset

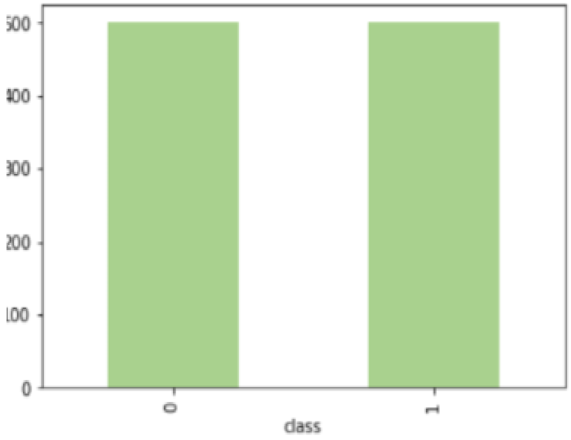

Fig.3. After Applying SMOTE, Borderline SMOTE, ADASYN, KMeans and Gaussian SMOTE Techniques

The proposed model DESMLA is evaluated using accuracy, recall, precision, and F1 score. Table III shows accuracy evaluation of the proposed model DESMLA using DT and RF and by using the oversampling techniques SMOTE, Borderline SMOTE, ADASYN, KMeans and Gaussian smote with the original data using RF and DT classifiers.

TABLE III. ACCURACY COMPARISONS

METHODS	DT	RF
With original dataset	70.12	**77.27**
DESMLA with SMOTE	65.58	**78.2**
DESMLA with Borderline SMOTE	69.48	**79.87**
DESMLA with ADASYN	67.53	**79.22**
DESMLA with KMEANS SMOTE	72.72	**81.07**
DESMLA with Gaussian SMOTE	75.97	**80.52**

From Table. III, it is clearly seen that DESMLA with RF gives a better prediction than DESMLA with DT even for imbalanced data, it is because RF is more robust than a single DT. In addition to this treating the imbalance nature reduces the biasness towards the majority class. Further, the proposed DESMLA using RF with KMeans Smote gives the highest accuracy of 81.07%.

TABLE IV. PRECISION COMPARISONS

METHODS	DT	RF
With original dataset	77	**81**
DESMLA with SMOTE	72	**83**
DESMLA with Borderline SMOTE	78	**88**
DESMLA with ADASYN	75	**85**
DESMLA with KMEANS SMOTE	81	**82**
DESMLA with Gaussian SMOTE	85	**86**

TABLE V. RECALL COMPARISONS

METHODS	DTII	RF
With original dataset	76	**84**
DESMLA with SMOTE	77	**83**
DESMLA with Borderline SMOTE	73	**80**
DESMLA with ADASYN	74	**80**
DESMLA with KMEANS SMOTE	74	**90**
DESMLA with Gaussian SMOTE	76	**79**

Moderately less number of false positives and false negative gives accurate prediction which leads to better precision and recall. From Table IV shows that, DEMLA with Borderline SMOTE using RF gives high precision of 88%. Also from table V, shows that, t DEMLA with KMeans SMOTE and RF gives high recall of 90%.

F1 score is considered as a performance metric whenever there is class imbalance problem in the dataset. The reason behind it is, the model predicts correctly for a majority class (no diabetes in this case). That is why F1 score is used as the evaluation metric.

TABLE VI. F1 SCORE COMPARISONS

METHODS	DT	RF
With original dataset	76	**84**
DESMLA with SMOTE	77	**83**
DESMLA with Borderline SMOTE	74	**83**
DESMLA with ADASYN	75	**84**
DESMLA with KMEANS SMOTE	77	**86**
DESMLA with Gaussian SMOTE	80	**82**

From table VI, it seen that DEMLA with KMeans SMOTE using RF also have a better F1 score measure.

V. CONCLUSION

The proposed method DESMLA is to boost the accuracy of the model by using various sampling techniques to rectify the class imbalance problem of the dataset. DESMLA first treat the class imbalance problem by using SMOTE, Borderline SMOTE, ADASYN, KMeans and Gaussian smote and then by using DT and RF diabetes is predicted.

The proposed procedure performs better for PIDD but have not consider other crucial factors related to gestational diabetes, like family history , metabolic syndrome, the habit of smoking, some dietary patterns, lazy routines etc. Hence in future more advance classifiers can be used to produce better results using more relevant and location oriented data.

REFERENCES

[1] Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018 American Diabetes Association Diabetes Care 2018; 41(Supplement 1): S13–S27. https://doi.org/10.2337/dc18-S002

[2] K. G. Alberti and P. Z. Zimmet, “Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation”, *Diabet Med.*, vol. 15(7), pp. 539–53, 1998.

[3] J. S. Kaddis, B. J. Olack, J. Sowinski, J. Cravens, J. L. Contreras, and J. C. Niland, “Human pancreatic islets and diabetes research”, *JAMA*, vol. 301(15), pp. 1580–1587, 2009, https://doi.org/10.1001/jama.2009.482

[4] O. Raha, S. Chowdhury, S. Dasgupta, P. Raychaudhuri, B. N. Sarkar, P. V. Raju and V. R. Rao, “Approaches in type 1 diabetes research: A status report”, *International journal of diabetes in developing countries*, vol. 29(2), pp. 85–101, 2009. https://doi.org/10.4103/0973-3930.53126

[5] L. Bellamy, J. P. Casas, A. D. Hingorani and D. Williams, “Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis", *Lancet*, vol. 373, pp. 1773–1779 2009. doi: 10.1016/S0140-6736(09)60731-5. PMID: 19465232.

[6] J. B. Meigs, R. B. D’Agostino, P. W. Wilson, L. A. Cupples, D. M. Nathan and D. E. Singer, “Risk variable clustering in the insulin resistance syndrome: The Framingham Ofspring Study”, *Diabetes*, vol. 46, pp. 1594–1600, 1997. doi: 10.2337/diacare.46.10.1594. PMID: 9313755.

[7] V. Anna, H. P. van der Ploeg, N. W. Cheung, R. R. Huxley and A. E. Bauman, “Socio-demographic correlates of the increasing trend in prevalence of gestational diabetes mellitus in a large population of women between 1995 and 2005”, *Diabetes Care*, vol. 31(12), pp. 2288–2293, 2008.

[8] P. Swain and H. Hauska, “The Decision Tree Classifier: Design and Potential.” *IEEE Transactions on Geoscience Electronics*, vol. 15(3), pp. 142-147, 1977.

[9] J. R. Quinlan, “Induction of Decision Trees”, Machine *Learning*, 1, pp. 81-106, 1986.

[10] S. R. Safavian and D. Landgrebe, “A Survey of Decision Tree Classifier Methodology.” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21(3), pp. 660-674, 1991.

[11] A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, “Overview of use of decision tree algorithms in machine learning”, *IEEE Control and System Graduate Research Colloquium (ICSGRC)*, pp. 37-42 (2011). https://doi.org/10.1109/ICSGRC.2011.5991826

[12] L. Breiman, “Bagging predictors”, *Machine Learning*, vol. 24(2), pp. 123–140, 1996.

[13] L. Breiman, “Random Forest”, *Machine Learning*, vol. 45, pp. 5–32, 2001.

[14] N. V. Chawla, K. Bowyer, L. O. Hall and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Artificial Intelligence Research*, vol. 16(1), pp. 321-357, 2002. DOI:10.1613/jair.953

[15] H. Han, W. Y. Wang, B. H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning.” In: Huang DS, Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11538059_91

[16] H. Haibo, Y. Bai, E. A. Garcia and L. Shutao, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence*), pp. 1322-1328, 2008, DOI: 10.1109/IJCNN.2008.4633969.

[17] F. Last, G. Douzas, F. Bacao, “Oversampling for Imbalanced Learning Based on K-Means and SMOTE”, *Information Sciences*, 465, pp. 1-20, 2017.

[18] P. Tingting, Z. Junhong , J. Y. Wei Wu, “ Learning imbalanced datasets based on smote and gaussian distribution” [online]: available-https://www.sciencedirect.com/science/article/abs/pii/S0020025519310187

[19] T.M. Alam, M.A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T.I. Baig, et al., “A model for early prediction of diabetes”, Inform. Med. Unlocked, vol. 16, p. 100204, 2019. https://doi.org/10.1016/j.imu.2019.100204

[20] S. Perveen, M. Shahbaz, A. Guergachi and K. Keshavjee, “Performance analysis of data mining classification techniques to predict diabetes”, *Procedia Computer Science*, vo. 82, pp. 115-121, 2016.

[21] D. Sisodia, D.S. Sisodia, “Prediction of diabetes using classification algorithms”, *Procedia Comput. Sci.*, vol 132, pp.1578–1585, 2018.

[22] S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 141-146, 2021. doi: 10.1109/ICACCS51430.2021.9441935.

[23] N. P. Tigga and S. Garg, “predicting type 2 Diabetes using Logistic Regression”, In: Nath, V., Mandal, J.K. (eds) Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems. Lecture Notes in Electrical Engineering, vol 673. Springer, Singapore. https://doi.org/10.1007/978-981-15-5546-6_42

[24] S. A Diwani and A Sam, “Diabetes forecasting using supervised learning techniques”, *Adv. Comput. Sci.: Int. Journal*, vol. 3(5), 2014.

[25] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, “Predicting Diabetes Mellitus with Machine Learning”, *Front. Genet.*, vol.9, 2018.

[26] J. P Kandhasamy and S. Balamurali, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus”, *Procedia Comput. Sci.*, vol. 47, pp. 45-51, 2015.

[27] N. Yuvaraj and K. R. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster”, *Cluster Computing*, vol.22, pp. 1-9, 2017.

[28] A. N. Boruah, S. Kumar Biswas, S. Bandyopadhyay and S. Sarkar, "Expert System to Manage Parkinson Disease by Identifying Risk Factors: TD-Rules-PD," *2020 International Conference on Computational Performance Evaluation (ComPE), 2020*, pp. 001-006, doi: 10.1109/ComPE49325.2020.9200075.

[29] Z. Tafa, N. Pervetica and B. Karahoda, "An intelligent system for diabetes prediction," *2015 4th Mediterranean Conference on Embedded Computing (MECO), 2015*, pp. 378-382, doi: 10.1109/MECO.2015.7181948.

[30] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction”, *ICT Express*, vol. 7(4), pp. 432-439, 2021.

[31] A. N. Boruah, S. K. Biswas, S. Bandyonadhyay and S. Sarkar, "An Expert System for Identification of Key Factors of Parkinson's Disease: B-TDS-PD," *2020 IEEE India Council International Subsections Conference (INDISCON)*, 2020, pp. 37-42, doi: 10.1109/INDISCON50162.2020.00020.