An Ensemble Technique for Early Prediction of Type 2 Diabetes Mellitus – A Normalization Approach

A. Prakash[a], O. Vignesh [b], R. Suneetha Rani [c] and S. Abinayaa [d]

[a]Associate Professor, Department of EEE, QIS College of Engineering and Technology, Ongole
[b]Assistant Professor (Sr. G.), Department of ECE, Easwari Engineering College, Chennai
[c]Associate Professor, Department of CSE,QIS College of Engineering and Technology, Ongole
[d]Associate Professor , Department of ECE, QIS College of Engineering and Technology, Ongole
.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Diabetic disease in particular originates due to the presence of blood sugar higher than at normal level. This is mainly due to the inadequate production of insulin. In recent days, it has been noticed that in all parts of the world, there has been an increasing number of people affected by diabetes. In the coming days, it is clear that this problem needs to be given greater importance in ensuring a reduction in the average diabetes affecting people. Many researchers carried out comprehensive research recently in the data mining platform to assess the accuracy of diabetic disease prediction at an early stage. The proposed ensemble model incorporates J48, NBTree, RandomForest, SimpleCART, and RandomTree and it is focussed on improving performance parameters such as accuracy, precision, f-measure, ROC, PRC and computational time. The conclusion shows that obtained accuracy is 4.03% higher than other standalone techniques was obtained using the proposed ensemble model.

Keywords*:* Type 2 Diabetes Mellitus, Machine Learning, Normalization, Ensemble Model, Prediction.

1. Introduction

People today suffer from a variety of diseases in the universe, the most common of which is diabetes. Changes in diet, lack of physical activity, elevated stress due to different issues, and obesity are the viable reasons behind the development of this disease. Pancreas not secreting enough insulin and incompetent insulin use induces DM. The downside of this condition is that the symptoms are very mild and that the diagnosis is delayed and that more serious diseased diseases such as kidney and nervous system complications are gradual. Diabetes can be classified into Type 1 diabetes mellitus (T1DM), Type 2 diabetes mellitus (T2DM), Prediabetes and Gestational Diabetes (Leanne, 2009). Out of the above, T2DM is very normal in most patients and diagnosis at an early stage is very important (Jia, 2004). In 2015, about 415,000,000 people were infected by the disease, and it could go up in the coming years, according to a study released by WHA (World Health Association). Around one in ten adults is estimated to suffer from diabetes. Early detection and diabetes prevention are therefore important to reduce this risk. A high risk T2DM community is to be targeted and an extreme study conducted in order to reduce its impact. The micro and macrovascular complications of diabetes are shown in table 1.

Table 1. Major Complications of Diabetes

Microvascular		**Macrovascular**	
Eye	High blood glucose and high pressure can damage eye blood vessels, causing retinopathy, cataracts and glaucoma	Brain	Increased risk of stroke and cerebrovascular disease, including transient ischemic attack, cognitive impairment etc.
Kidney	High blood pressure damages small blood vessels and excess blood glucose overworks in the kidneys resulting in nephropathy	Heart	High blood pressure and insulin resistance increase risk of coronary heart disease
Neuropathy	Hyperglycemia damages nerves in the peripheral nervous system. This may result in pain and/or numbness. Feet wounds may go undetected, get infected and lead to gangrene.	Extremities	Peripheral vascular disease results from narrowing of blood vessels increasing the risk for reduced or lack of blood flow in legs. Feet wounds are likely to heal slowly contributing to gangrene and other complications.

Therefore, for further development some advanced technology is necessary. Data mining (DM), also known as KDD – Information Discovery in databases (Byoung, 2019) is therefore identified as suitable technology for analysis. Using different technologies linked to advanced intelligent technologies, DM can typically be used to

discover different trends in the broader data sets of balanced or imbalanced values (Wu *et al.,* 2018).Elman's recurrent ANNs was used(Robertson, 2011) to forecasts BGLs, meal consumption and injections of insulin. A freeware AIDA mathematical diabetes simulator has assembled a total of 28 datasets. A comparison was made between logistic regression, artificial networks of nervous networks and decision tree models using common risk factors for predicting diabetes or prediabetes (Meng, 2013). In order to obtain demographic information, family history of diabetes, anthropometric and lifestyle risk factors, a standard questionnaire was administered.A focuson predictive diabetic care research, using the technology of regression was performed(Aljumah., 2013). Support vector machine was used for experimental research. Diabetes mellitus is diagnosed(Saxena, 2014) using K-Nearest Neighbor algorithm. Different k values for 3 to 5 with two different test data were used to calculate the efficiency of the KNN Algorithm.

An attemptto apply the method of data mining to analyse the database of diabetes and diagnoses using algorithms (Tejashri *et al.,* 2014) such as Naive Bayes, J48 (C4.5), JRip, neural network, decision trees and KNN, fuzzy logic and genetic algorithms. A novel method for feature selection was proposed(Wang *et al.,* 2015) using the opposite sign-test (OST) algorithm denoted as improved electromagnetism-type mechanism (IEM). The nearest neighbor algorithm is used as a wrapper classifier. Using decision tree and naïve bayes algorithm, (Aiswarya,2015) implemented the pattern analysis process. The noninvasive risk models for the prediction of undiagnosed diabetic disease in Africa have also been introduced(Mbanya., 2015). Some 20 models were evaluated by (Bashir *et al.,*2016) along with the developed HM-BagMoov design, using a collection of seven heterogeneous classificators which overcomes limitations of traditional efficiency bottlenecks. Five separate datasets of cardiac disease, four datasets of breast cancer, two datasets of diabetes, two datasets of liver disease, and one hepatitis dataset from public repositories are used to test the system. Deep Neural Networks and Hybrid Deep Learning Dynamics of Vector Machines was implemented(Kau., 2017) and have provided better results than the other techniques in diabetes prediction. The relative performance of different methods of machine learning (Alghamdi, 2017) was analysed for the prediction of diabetes by using a cardio-respiring fitness record by the use of Decision Tree, Naïve Bayes, Logistic Regression Tree and Random Forest.

An early prediction model was proposed (Ijaz, 2018) for type 2 diabetes (T2D) which consists of a spatial clustering of noise applications based on density (DBSCAN) for the detection of outlier data for removal, SMOTE for the equilibria of class distribution, and a random forest for disease classification. A new data mining model that could improve accuracy, and make it possible for a model to be adopted by several datasets, for predicting T2DM was presented (Wu, 2018). An ML technique was proposed (Sneha .,2019) to apply significant characteristics, design a Machine Learning Prediction Algorithm, and find an optimum classifier to provide the closest results to clinical results. The method proposed focuses on the selection of the attributes of predictive analysis that are used for early detection of DM. Deep Neural Network was implemented (Zhou ,2020) that can be used not only to predict the future onset of diabetes, but also to identify a person's disease type. Given the differences in their treatment methods, type 1 diabetes and type 2 diabetes, this method will help the patient to be treated correctly. A risk nomogram of diabetic retinopathy (DR) was developed (Rhouhui , 2020), where the survey was conducted on 4170 T2 DM patients, with a biochemical indicator examination, and with the data collected, the risk of DR in T2DM patients was evaluated.

Many methods are used to recognise the pattern, forecast, cluster and associate. The highest possible aspects of DM are the quality of data and the application of appropriate methodologies. DM has been used since the beginning to extract useful information from a wide range of data sets in different domains. Further improvements need to be made by constructing an effective model to predict and prevent this disease. T2DM diagnostic structure and WEKA data visualisation are presented in this section (see Figure 1 and Figure 2). The flow of the article is as follows. Section 2 explains the data set selected for analysis. Section 3 explains the pre-processing procedure and Section 4 explains the proposed ensemble technique. The performance assessment and its comparison with various methods are included in Section 5. Section 6 presents the conclusion of the article with additional improvements.

2. Dataset

UCI has a repository of different data sets for the study and application of machine learning algorithms. It is widely used as a primary source of machine learning data sets by researchers, students and educators. For the purpose of our study, we took PIMA Indian Diabetes Dataset from this repository. This dataset is comprised of 768 patients' medical data. Table 2 illustrates the data associated with the chosen dataset. The class variable is the 9th attribute (class) of every data point. For positive and negative diabetes, the result will be 1 and 0 respectively.

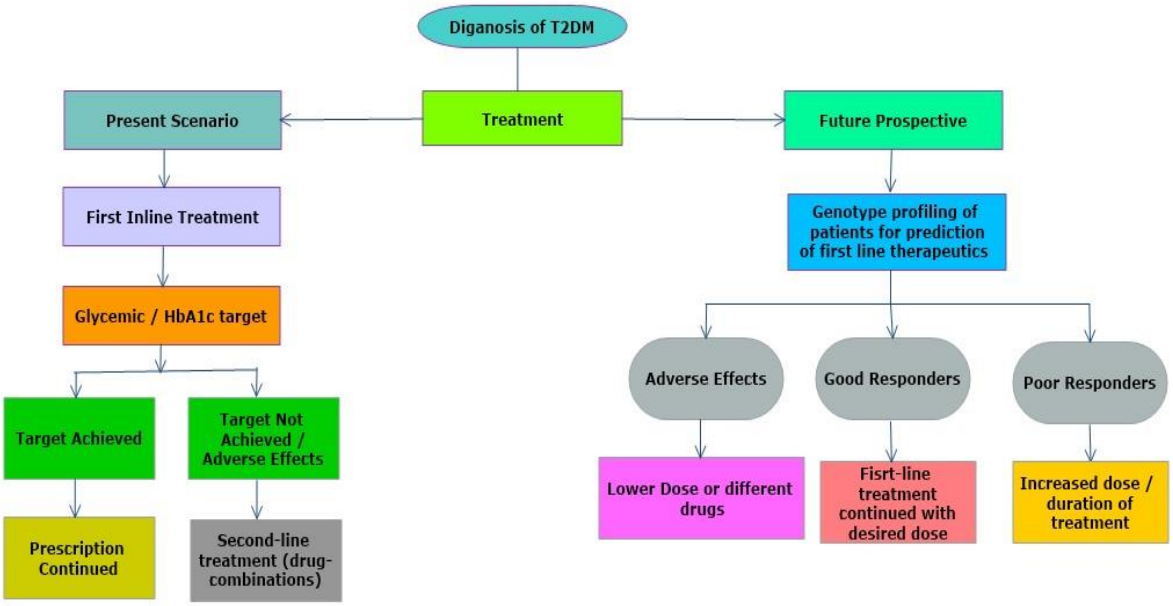

Figure 1. Diagnosis Structure of Type 2 Diabetes Mellitus

Table 2 List of attributes, its variable type and range in the Pima Indian Diabetes Dataset

S. No.	Feature Label	Variable Type	Range
1	Number of times pregnant	Integer	0-17
2	Plasma Glucose Concentration in a 2-hour oral glucose tolerance test	Real	0-199
3	Diastolic blood pressure	Real	0-122
4	Triceps skin fold thickness	Real	0-99
5	2 hours of serum insulin	Real	0-846
6	Body Mass Index	Real	0-67.1
7	Diabetes Pedigree Function	Real	0.078-2.42
8	Age	Real	21-81
9	Class	Integer	0,1

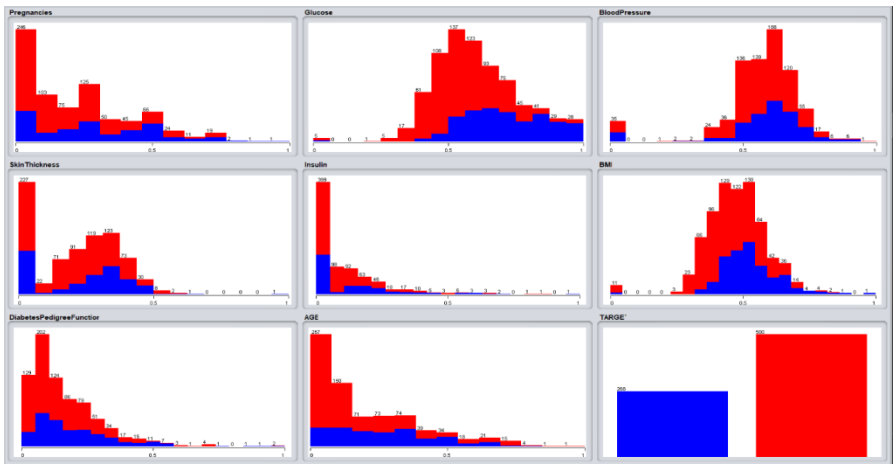

Figure 2. Input data visualization in WEKA

3. Normalisation – A Preprocessing Procedure

Usually, any real-world data contains some kind of noise. So, pre-processing of data is done to reduce it. Each attributes of the data may have a very different range of values. For example, in our PIMA Indian Diabetes dataset, 4th attribute, i.e., the triceps skin fold thickness is in the range of 0-99 whereas the seventh attribute, i.e., diabetes pedigree function has a range of 0.078 to 2.42. This type of variation in the range results in different weighing attributes of the algorithms that use data point distance. Table 3 epitomizes the normalized mean and standard deviation values for all the instances, this is corrected by normalization / standardisation. The main aim is to achieve all attributes below the same minimum, maximum and medium levels in order to normalise the attributes. We have used standardisation features (z-score normalisation) that normalise data by means of mean and standard deviation. It is done using the following formula:

$$z = \frac{x_i - \mu}{\sigma} \qquad (1)$$

where

μ = Mean
σ = Standard Deviation
z = Normalized attribute value
x_i = Original attribute value

Table 3 Statistical Analysis

Feature Label	Before Normalization		After Normalization	
	Mean	**Std. Deviation**	**Mean**	**Std. Deviation**
Number of times pregnant	3.845	3.37	0.226	0.198
Plasma Glucose Concentration in a 2-hour oral glucose tolerance test	120.895	31.973	0.608	0.161
Diastolic blood pressure	69.105	19.356	0.566	0.159
Triceps skin fold thickness	20.536	15.952	0.207	00.161
2 hours of serum insulin	79.799	115.244	0.094	0.136
Body Mass Index	31.993	7.884	0.477	0.117
Diabetes Pedigree Function	0.472	0.331	0.168	0.141
Age	33.241	11.76	0.204	0.196

4. Ensemble Model – Proposed Approach

Ensemble is a technique that combines several machine learning techniques in one ideal predictive model to reduce variance, bias, or improve predictions. Compared to a single model, this approach enables improved predictive performance. In the proposed model, 80% (641 instances) of the total data is used as training data and the remaining 20% (154 instances) as testing data. The following are the ML techniques chosen for ensemble model with forward selection and backward elimination technique (FSBE): J48, NBTree, RandomForest, SimpleCART, RandomTree. The main reason behind using the FSBE technique is that it will fasten the training process, complexity reduction, improvement in accuracy and reduction in over-fitting. The proposed ensemble model is presented below (see Figure 3).

4.1 Implementation, results and discussion

The execution was carried out using WEKA tool, which is one of the most efficient tools for datamining process. The performance analysis of the proposed model was compared with various stand-alone models such as Bayes Net with Hill Climber Search Algorithm, Naïve Bayes (NB), Multi-Layer Perceptron (MLP), AdaBoost with Decision Stump and Bagging with Random Forest Classifier. Table 4 shows the results obtained for various performance metrics.

Table 4 Comparison of Various Performance Metrics

Algorithm	Precision	Recall	F-measure	ROC Area	PRC Area
BayesNet-HCSA [executed]	0.750	0.753	0.751	0.814	0.812
Naïve Bayes [11]	0.741	0.747	0.741	0.819	0.818
MLP [executed]	0.754	0.753	0.735	0.831	0.829
AdaBoost-Decision Stump [executed]	0.755	0.760	0.749	0.820	0.815
Bagging-Random Forest [executed]	0.740	0.747	0.739	0.831	0.832
Proposed Ensemble Model	0.783	0.786	0.783	0.832	0.836

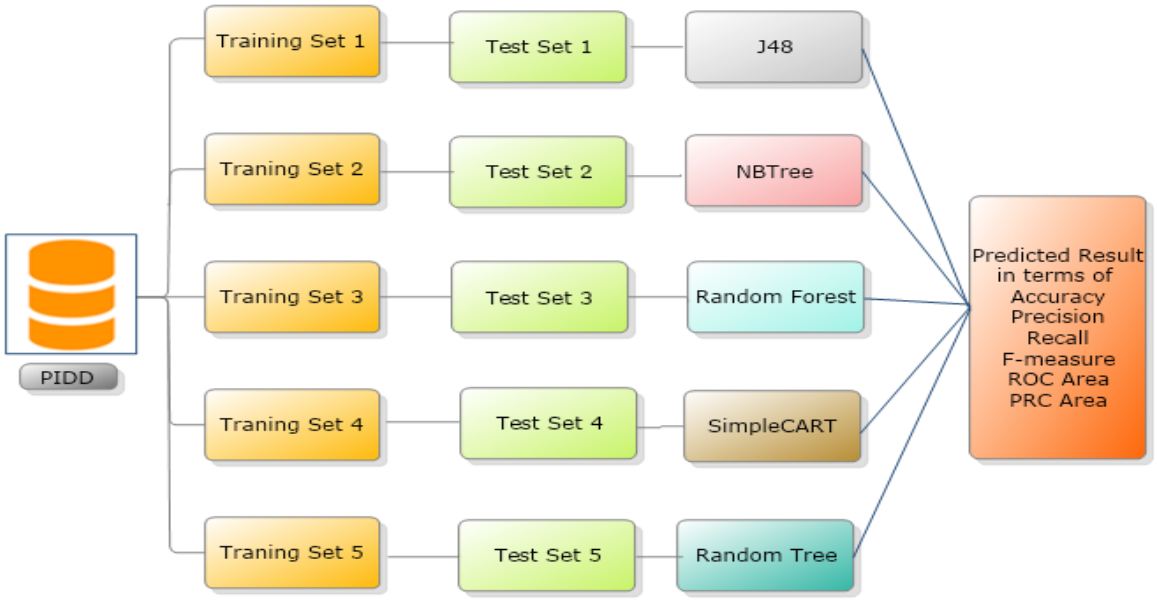

Figure 3. Proposed Ensemble Model

Table 5 shows the comparison of accuracy obtained and time taken to build the model using the standalone classifiers and proposed ensemble model. From table 4, it is inferred that the all the performance metrics obtained by the proposed ensemble model shows better result when compared with other standalone techniques. The parameters precision, recall, f-measure, ROC area and PRC area shows an average improvement of 0.035%, 0.034%, 0.04%, 0.009% and 0.015% than the stand-alone techniques. From table 5, it is inferred that the accuracy obtained by the proposed ensemble model is 79.22%, which shows an average accuracy improvement of 3.78%. The future forecast (Figure 4) representation of all the instances of the dataset is presented, where predictions (forecasts) can be generated for future events based on known past events.

Table 5. Comparison of accuracy and time taken to build the model

Algorithm	Accuracy (%)
Logistic regression [6]	76.13
ANN Model [6]	73.23

Decision Tree (C5.0) [6]	77.87
Improved Electromagnetism-likeMechanism [10]	73.0263
Naïve Bayes [11]	74.67
Deep Neural Network [14]	78.12
MLP [executed]	75.32
BayesNet-HCSA [executed]	75.32
AdaBoost-Decision Stump [executed]	75.97
Bagging-RandomForest [executed]	74.67
Proposed Ensemble Model	**79.22**

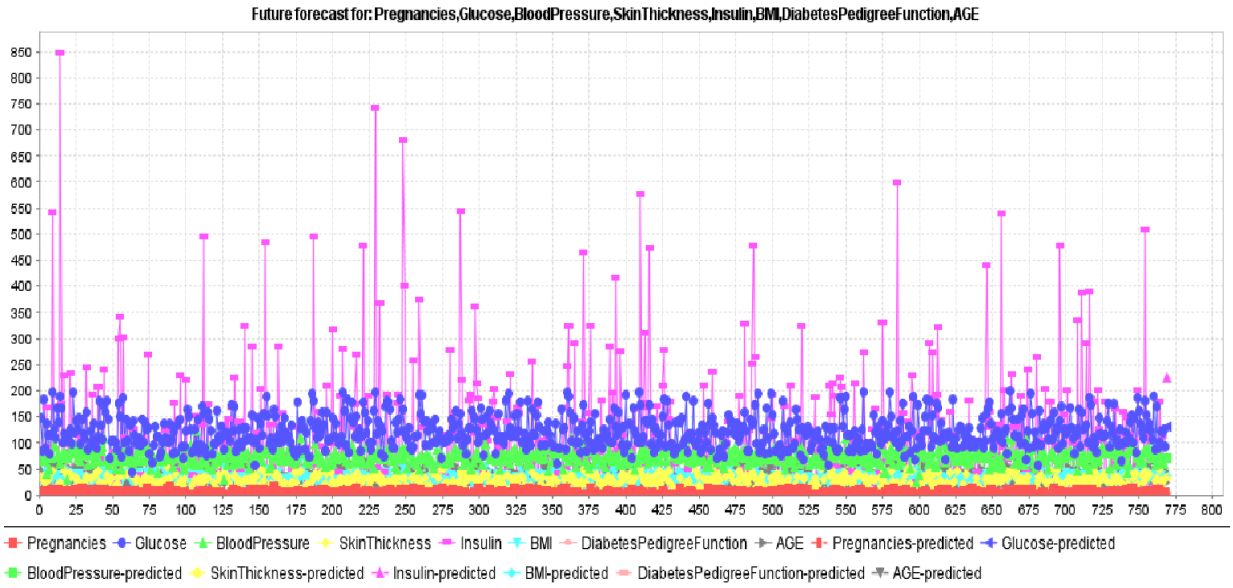

Figure 4. Future forecast representation of all instances

5. Conclusion

An ensemble model is proposed in this research article to predict type-2 diabetes mellitus at an early stage. The proposed model is a combination of J48, NBTree, RandomForest, SimpleCART, RandomTree algorithm. The general PIDD is normalized in order to obtain the instances at an equal range. The highest accuracy obtained with the proposed ensemble model is 79.22 % while the other standalone techniques produce accuracy and other performance metrics lesser than the ensemble model. The entire model was evaluated on a 10-fold cross validation with 80% training data and 20% testing data. As a future enhancement, various machine learning algorithms and deep learning techniques can be implemented to achieve better prediction accuracy.

6. Acknowledgements

The authors would like to thank the management of QIS College of Engineering and Technology for their continuous support in providing the facilities to complete this research activity.

References

1. Alqazzaz S, Sun X, Yang X, Nokes L. (2019) *Automated brain tumor segmentation on multi-modal MR image using SegNet. Comput Vis Media;5(2):209–19.*

2. Aiswarya, I., Jeyalatha, S., & Sumbaly, R. (2015). *Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process*, 5, 1-14.
3. Alghamdi, M., Mallah, MA., Keteyian, S., Brawner, C., Ehrman, J., &Sherif, S. (2017). *Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project, PLoS ONE, 12, 1-15.*
4. Aljumah, AA., Ahamad, MG., and Siddiqui, MK. (2013). *Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University–Computers and Information Sciences, 25, 127–136.*
5. Balamurugan, K., Uthayakumar, M., Sankar, S., Hareesh, U.S. and Warrier, K.G.K., 2019. Predicting correlations in abrasive waterjet cutting parameters of Lanthanum phosphate/Yttria composite by response surface methodology. Measurement, 131, pp.309-318.
6. Bashir, S., Qamar, U., & Khan, FH. *IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework, Journal of Biomedical Informatics, 59, 185-200.*
7. Bhasha, A.C., Balamurugan, K. End mill studies on Al6061 hybrid composite prepared by ultrasonic-assisted stir casting. Multiscale and Multidiscip. Model. Exp. and Des. (2020). https://doi.org/10.1007/s41939-020-00083-1
8. Byoung, G.C., Rha, S.W., Kim, S.W., Kang, J.H., Park, J.Y., & Noh, Y.K. (2019). *Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks, Yonsei Medical Journal, 60(2), 191-199.*
9. ChinnamahammadBhasha, A., Balamurugan, K. Studies on Al6061nanohybrid Composites Reinforced with SiO_2/3x% of TiC -a Agro-Waste. Silicon (2020). https://doi.org/10.1007/s12633-020-00758-x
10. Ijaz, MF., Alfian, G., Syafrudin, M., & Rhee, J. (2018). *Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE) and random forest, Applied Sciences, 8, 1-22.*
11. Jia, J. (2004). *The study of the application of a web-based chatbot system on the teaching of foreign languages. Society for Information Technology and Teacher Education International Conference. Atlanta, Georgia, United States of America: Association for the Advancement of Computing in Education (AACE).*
12. Kaul, D., Raju, H., & Tripathy, BK. (2017). *Comparative analysis of pure and hybrid machine learning algorithms for risk prediction of diabetes mellitus, Helix, 7, 2029-2033.*
13. Latchoumi, T.P. and Kannan, V.V., 2013. Synthetic Identity of Crime Detection. International Journal, 3(7), pp.124-129
14. Latchoumi, T.P. and Parthiban, L., 2016. Secure Data Storage in Cloud Environment using MAS. Indian Journal of Science and Technology, 9, pp.24-29.
15. Leanne, B., Casa, JP., Hingorani, AD., & Williams, D. (2009). *Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis, Lancet, 373 (9677), 1773-1779.*
16. Loganathan, J., Latchoumi, T.P., Janakiraman, S. and parthiban, L., 2016, August. A novel multi-criteria channel decision in co-operative cognitive radio network using E-TOPSIS. In *Proceedings of the International Conference on Informatics and Analytics* (pp. 1-6). https://doi.org/10.1145/2980258.2982107
17. Mbanyaa, V., Hussain., A., & Kengne, AP. (2015). *Application and applicability of non-invasive risk models for predicting undiagnosed prevalent diabetes in Africa: A systematic literature search, Primary Care Diabetes, 9, 317-329.*
18. 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 Journal of Medical Sciences, 29, 93-99.*

19. Robertson, G., Lehmann, ED., Sandham, W., & Hamilton, D. (2011). *Blood glucose prediction using artificial neural networks trained with the AIDA Diabetes Simulator: A Proof-of-Concept pilot study. Journal of Electrical and Computer Engineering, Hindawi Publishing Corporation, 681786, 1-12.*

20. Ruohui, M., Shi, R., Hu, Y., & Hu, F. (2020). *Nomogram based prediction of the risk of diabetic retinopathy: A retrospective study, Journal of Diabetes Research, 7261047, 1-13.*

21. Saxena, K., Khan, Z., & Singh, S. (2014). *Diagnosis of diabetes mellitus using K Nearest Neighbor algorithm, International Journal of Computer Science Trends and Technology, 2, 36-43.*

22. Sneha, N., &Gangil, T. (2019). *Analysis of diabetes mellitus for early prediction using optimal features selection, Journal of Big Data, 6, 1-19.*

23. Tejashri Giri, N., & Todamal, SR. (2014). *Data mining approach for diagnosing type 2 diabetes. International Journal of Science Engineering and Technology, 2, 191-194.*

24. Wang, KJ., Adrian, AM., Chen, KH., & Wang, KM. (2015). *An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus, Journal of Biomedical Informatics, 54, 220–229.*

25. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). *Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, 10, 100–107.*

26. Wu, H., Yang, S., Huang, Z., He., J., & Wang, X. (2018). *Type 2 diabetes mellitus prediction model based on data mining, Information in medicine unlocked, 10, 100-107.*

27. Zhou, H., Myrzashova, R., & Zheng, R. (2020). *Diabetes prediction model based on an enhanced deep neural network, EURASIP Journal on Wireless Communication Networks, 148, 1-13.*