ORIGINAL ARTICLE

A Multilayer Perceptron Model to Predict Risk Factors of Type II Diabetes Mellitus

Brindha Senthil Kumar[1+], Vanlalawmpuia R[2+], Freda Lalrohlui[3], John Zohmingthanga[4], Lalruatpuii Hlawnmual[5], Nachimuthu Senthil Kumar[6] and Lal Hmingliana[7*]

[1]Department of Computer Science Engineering, Mizoram University, Aizawl, Mizoram – 796004, India.
[2]Department of Computer Science Engineering, Mizoram University, Aizawl, Mizoram – 796004, India.
[3]Department of Biotechnology, Mizoram University, Aizawl, Mizoram – 796004, India
[4]Department of Pathology, Civil Hospital Aizawl, Mizoram – 796001, India
[7]Department of Medicine, Civil Hospital Aizawl, Mizoram – 796001, India
[6]Department of Computer Science Engineering, Mizoram University, Aizawl, Mizoram – 796004, India.
[7]Department of Computer Science Engineering, Mizoram University, Aizawl, Mizoram – 796004, India.

ABSTRACT Diet and lifestyle factors are the significant cause of Type II Diabetes Mellitus (T2DM). This case-control study aims to develop a high-precision machine learning model to predict T2DM using Multi-layer Perceptron (MLP) from epidemiological data. The epidemiological data utilized in this work were collected from 500 T2DM patients and 500 healthy individuals using well-structured questionnaire. The balanced dataset consisted of 11 salient features from diet and lifestyle which are high risk factors for T2DM. A MLP model was built using eleven input layers, three-set of hidden layers, one output layer with ReLU as activation function, adam as model optimizer, momentum of 0.99 and learning rate set to 0.0011 to achieve its best accuracy and low error rate. The proposed MLP model produced accuracy of 96%, F1 score of 95%, Receiver operating characteristic (ROC) of 95%, precision of 93%, recall of 98% and misclassification rate of 4.5%. Age, sex, meat intake, saum, smoked meat, salt intake, smoking, sadha, tuibur, chewing paan and alcohol were found to be risk factors to cause T2DM apart from the other environmental factors. These findings indicate that machine learning models are reliable in predicting T2DM disease with its core risk factors for the awareness or early diagnosis of T2DM.

Keywords: Multi-layer Perceptron, Misclassification rate, ROC, Type II Diabetes Mellitus, Lifestyle-diet, Case-control study

Address for correspondence: Lal Hmingliana, Department of Computer Science Engineering, Mizoram University, Aizawl, Mizoram – 796004, India. E-mail: lalhmingliana@mzu.edu.in + First author and second author had equally contributed for this paper.

Submited: 13-Feb-2022 **Accepted:** 17-May-2022 **Published:** 22-Jun-2022

INTRODUCTION

Globally, the number of diabetic people is expected to be 642 million by 2040 (Maniruzzaman, 2020). Poor lifestyle like lack of physical exercise and unhealthy meals are like to play an important role in developing T2DM (Alshammari, 2020). In addition to diet and lifestyle factors, genetic factors (family history), ethnicity (race) also causes T2DM (Lalrohlui, 2020). The incidence of T2DM in American, Chinese, Caucasian, Aborigines (Australia), Pima Indians (Arizona) races is increasing due to sedentary lifestyle, poor diet or change in indigenous traditional diet (Golden, 2019; Steyn 2014). India has 77 million diabetic patients which rank the second in the global list (Alam, 2019). A 15-year longitudinal study had shown prolonged sedentary lifestyle is the major cause of aging-associated diseases such as T2DM, cardiovascular disease, cancer, etc. (*Belikov, 2019;* Thorp, 2011). People with age 45 above are at more risk to acquire T2DM, and the rate of acquiring this disease is an exponential rate in age group 65 and above (Standl, 2019). The major T2DM dietary risk factors are high-fatty foods, processed and red meat, refined grains, carbonated drinks (Sami, 2019). Smokeless-smoking

Access this article online
Website: www.ijfans.org
DOI: 10.4103/ijfans_110-22

How to cite this article: Brindha Senthil Kumar, Vanlalawmpuia R, Freda Lalrohlui, John Zohmingthanga, Lalruatpuii Hlawnmual, Nachimuthu Senthil Kumar and Lal Hmingliana. A Multilayer Perceptron Model to Predict Risk Factors of Type II Diabetes Mellitus. Int J Food Nutr Sci 2022;11: 67-74.

tobacco and alcohol abuse are another non-invasive lifestyle factors that causes T2DM (Maddatu, 2017; Wu, 2014; Carlsson, 2017).

Data mining techniques were applied on these risk factors to predict T2DM for many years, the latest work have been brief in the following. Deepnet (Deep Learning) has produced highest accuracy, precision, recall and F1 score of 88% from 3-year collected data from vital signs and blood reports in Saudi Arabian population (Alshammari, 2020). A retrospective study in Chinese population showed Multi-layer Perceptron (MLP) Model performance was better with an accuracy of 87% and ROC of 97% using features from blood test and vital signs (Xiong, 2019). Random forest classifier had been utilized on UCI repository diabetic dataset, which yield an accuracy of 92.26% and Area Under the Curve (AUC) of 91.14%, it had replaced the outliers and missing values using median and group median values respectively (Maniruzzaman, 2020). R-based machine learning model on female patients from Pima Indian population (UCI repository) had achieved accuracy of 89% with ROC of 90% using linear Support Vector Machine (SVM) and accuracy of 88% with better ROC of 92% by k-Nearest Neighbour (k-NN) algorithms (Kaur, 2018).

A Random forest classifier with feature extraction using minimum redundancy maximum relevance (mRMR) method had showed accuracy of 80.8% using all 14 features from Luzhou dataset (Zou, 2018). Blood test, hospitalization records, drug claims, insurance claims, emergency claims and ambulatory records were used to build XGBoost model, had AUC of 80.3% in Canadian population (Ravaut, 2019). Ensemble model by combining Random Forest, Naive Bayes Trees and Logistic Model tree classifiers has accuracy of 92.2% and AUC of 92.2% using synthetic minority oversampling technique (SMOTE) on imbalanced cardiorespiratory fitness data (Alghamdi, 2017). Gradient Boosting Machine and logistic regression algorithms had shown area under ROC curve of 84.5% and 84.1%, respectively based on laboratory and demographical data in age group between 18 to 90 years (Lai, 2019).

T2DM risk factors differ between ethnic groups, which had motivated to apply machine learning algorithm on the Mizo population, Northeast India to study their underlying epidemiological risk factors to causes this disease. In Mizoram, about 29.80% of people are getting treated for Type II Diabetes Mellitus (T2DM) (Anjana, 2017). The present work had developed a Multi-layer Perceptron (MLP) model using diet and lifestyle risk factors to predict T2DM in Mizoram urban population. The proposed model had been compared to latest 5 MLP diabetic models. Python Jupyter Notebook Version 3 platform was used to build data models using learning packages from scikit-learn (0.20.4). Data preprocessing, visualization and interpretations were done using Numpy (1.16.6), pandas (0.24.2), seaborn (0.9.1), scipy (1.2.3) and matplotlib (2.2.5) (Hunter, 2007).

METHODS

The primary data for this study was collected from Aizawl (urban) capital of Mizoram. The sampling was done from the patients who had been diabetic for more than 5 years, and healthy individuals with no prior medical conditions. A well-structured questionnaire was designed consisting of key dietary and lifestyle habits unique to the Mizo population. These epidemiological features were age, sex, meat_intake (red and white meat), saum (pork fat), smoked_meat (red and white meat), salt_intake, smoking, sadha, tuibur (smoke-infused tobacco water), chewing_paan and alcohol. Total of 1000 records (500 cases and 500 controls) were collected using the questionnaire after the consent of the participants.

Data distribution was represented using bar chart as all features were discrete variables, except the age, which was only continuous variable in this dataset (Figure 1). Bar charts were created with confidence interval (ci) set to 95% for each feature, the caps on the top of the bars indicate the error bars, and number of iterations was 1000 to compute the confidence interval, so bootstrap was set to 1000. The bars show less than 0.05 error/bar in all risk features with 95% of confidence interval, this clearly showed all the features in the dataset were statistically significant (Figure 1). The distributions of epidemiological and demographical data of diabetic and non-diabetic patients are given in Figure 2. Age, saum, meat_intake, smoked_meat, salt_intake, sadha, tuibur, and chewing_paan features showed prominent difference between cases (diabetic) and controls (non-diabetic). Except for age, other features were showing high-level of Gaussian distribution.

During the pre-processing phase, the missing values in the dataset were replaced by group median values and outliers in the age attribute were replaced by median value. All the eleven epidemiological factors were utilized to construct the machine learning models (Figure 3).

The steps are as follows:

1. A two-third of randomly divided dataset was used for training and one-third was reserved for testing the models.
2. Multi-layer perceptron, Random forest, support vector machine, and logistic regression models were built using training dataset.
3. Overfitting of the models was ruled out by checking train and test accuracies.
4. Models were optimized using hyper-parameter tuning.

Figure 1: Bar Chart of the Features in Diabetes Dataset

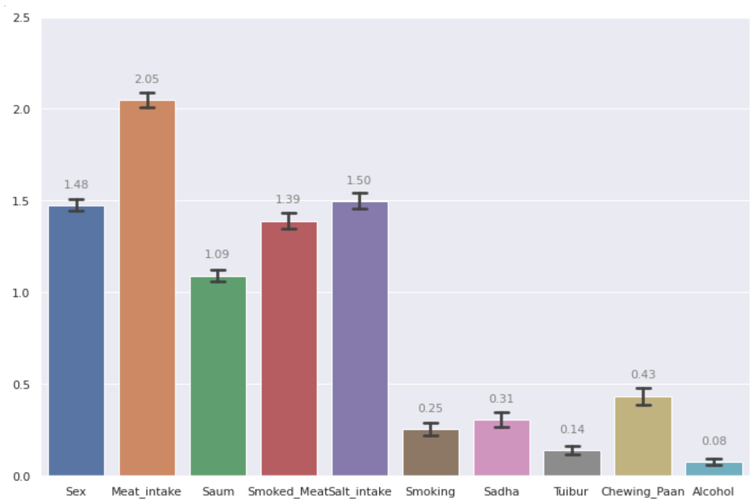


Figure 2: Distribution Plot Between Diabetic and Non-Diabetic with Respect to Risk Factors

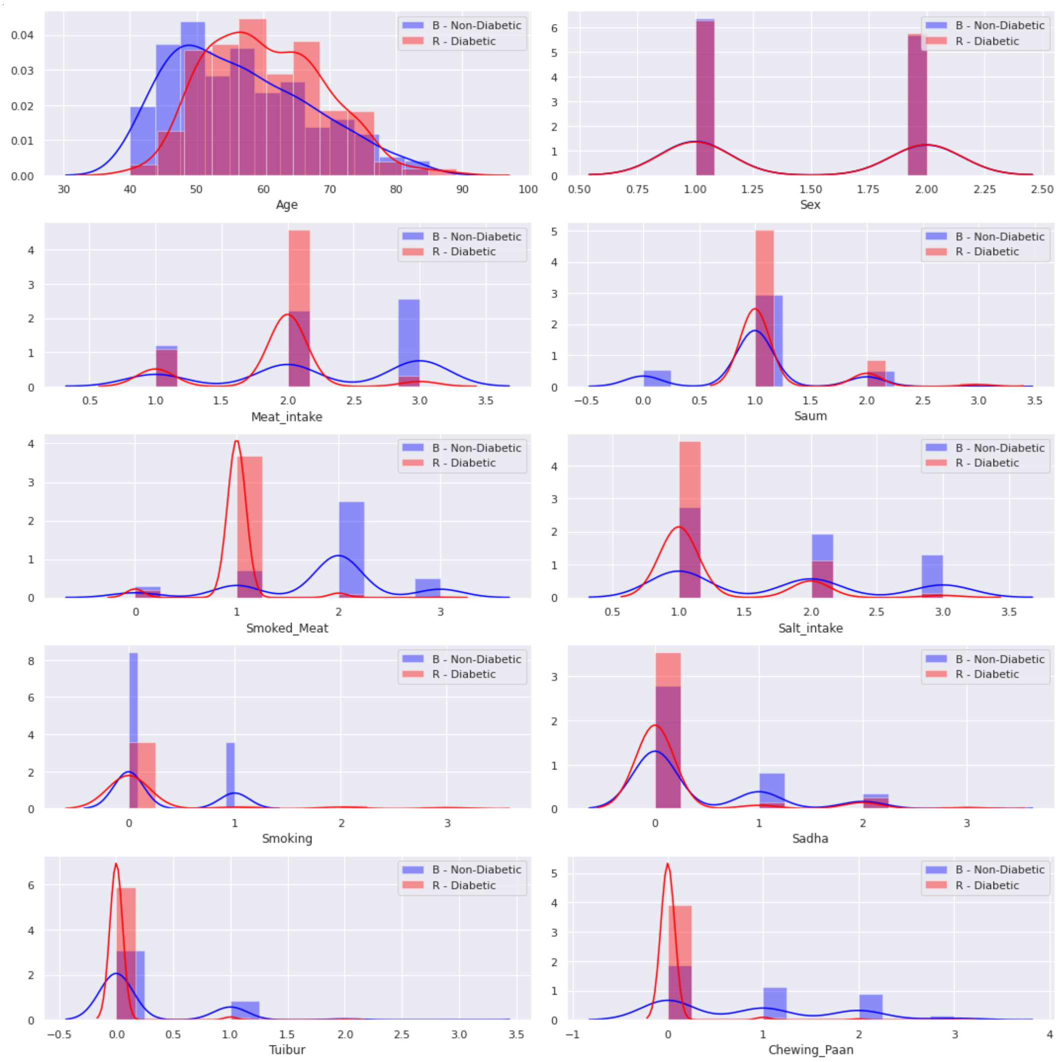
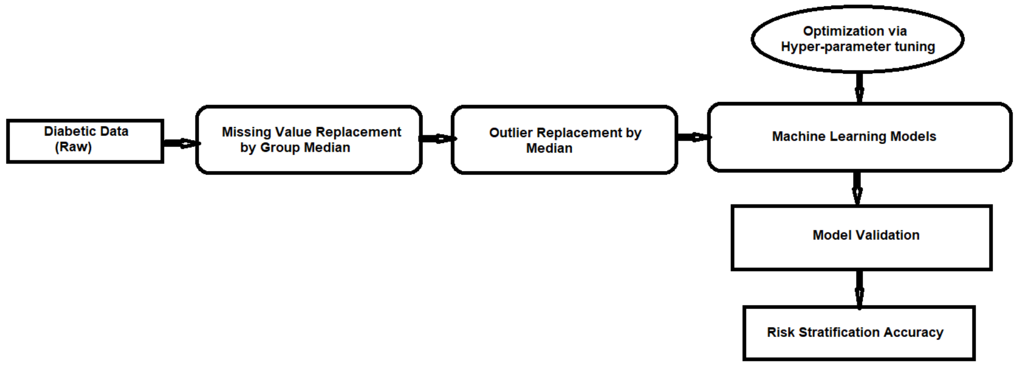


Figure 3: Preparation of Diabetic Data by Replacing Outliers and Missing Value and Evaluating the Models


Optimization via Hyper-parameter tuning
Diabetic Data (Raw)
Missing Value Replacement by Group Median
Outlier Replacement by Median
Machine Learning Models
Model Validation
Risk Stratification Accuracy

5. Model's performances were evaluation using accuracy, F1 score, precision, recall and ROC and low misclassification rate (calculated using root mean squared error).

Python Jupyter Notebook Version 3 platform was used to build data models using learning packages from scikit-learn (0.20.4). Data preprocessing, visualization and interpretations were done using Numpy (1.16.6), pandas (0.24.2), seaborn (0.9.1), scipy (1.2.3) and matplotlib (2.2.5).

RESULTS AND DISCUSSION

Models

Naive Bayes

The present diabetic dataset shows a Gaussian distribution in maximum number of features (Figure 2), so Gaussian Naive Bayes classifier was selected to construct the model. The Naive Bayes model has accuracy of 81%, F1 score of 84% and ROC of 81% (Table 1 and Figure 4). Recall of 96% showed naive bayes model had produced less of false negatives, but has high false positive rates with precision of 74%. The misclassification rate was high with 18.7% (Table 1). Though there was no overfitting of data, the model evaluation parameters were not satisfactory to classify diabetic or non-diabetic for the given risk features (Figure 5).

Logistic Regression

Performance of logistic regression model was better than Naive Bayes classifier as the accuracy, F1 score, ROC were 92 and recall was 97% (Table 1 and Fig 4). The misclassification rate was less with 8.4% in compared to naive bayes classifier, no overfitting, it has set back of low precision of 87% (Figure 5).

Support Vector Machine

Support vector machine classifier has misclassification rate of 8.1%, which was slightly less than logistic regression misclassification rate. The accuracy, F1 score and ROC of 92%, with dis-satisfaction in high false positive rates showed precision of 87% (Table 1 and Figure 4). Support vector machine showed no evidence of overfitting in this model, but still this performance does not support to prove these epidemiological risk factors causes diabetes in Mizoram population (Figure 5).

Multi-layer Perceptron

The MLP model was built using 11 input layer, three hidden layers and one output neural network layers. The number of neurons in three hidden layers was 25, 12, and 6 respectively. Rectified Linear Unit (ReLU) was used as activation function, 'adam' for the weight optimization, initial learning rate was 0.0011, momentum was 0.99, and batch size was 35. These hyper parameters were optimized to achieve the highest accuracy of 96%, F1 score of 95%, precision of 93%, recall of 98%, ROC of 95% and misclassification rate of 4.54% (Table 1 and Fig 4). The training accuracy of 98% and testing accuracy of 96% clearly showed there was not overfitting of data in MLP model (Fig 5.). Compared to support vector machine, logistic regression and naïve bayes models, MLP model had scored high satisfactory results in terms of best accuracy and low misclassification rate (Table 1 and Figure 4).

Random Forest

Random forest classifier had clearly shown overfitting of data as the training accuracy was higher than testing accuracy (Figure 5). This model had memorized the data points thereby the evaluation parameters had not been further considered for comparison. We need more data to feed the model, which can significantly overcome overfitting.

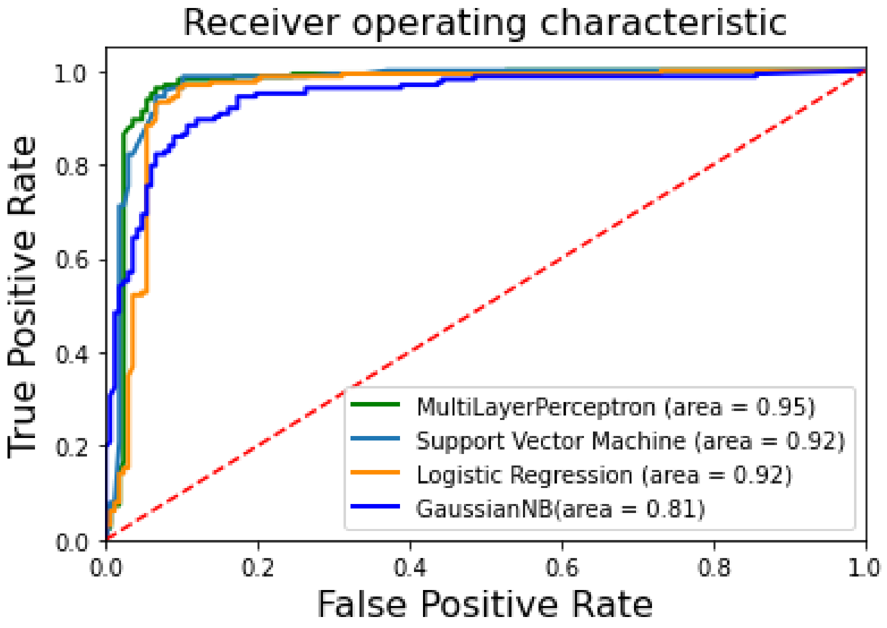
Comparison of MLP Models

To compare the performance of the proposed MLP model, the latest five models which had used MLP to classify diabetes dataset (T2DM) were considered. These five MLP models

Table 1: Evaluation Metrics of Four Classifiers

Models	Model Performance Parameters					
	Accuracy%	F1 Score%	Precision%	Recall%	Mis-classification Rate %	ROC%
Multi-layer Perceptron	96	95	93	98	4.5	95
Support Vector Machine	92	92	87	98	8.1	92
Logistic Regression	92	92	87	97	8.4	92
Naive Bayes	81	84	74	96	18.7	81

Figure 4: ROC Curve of Four Models



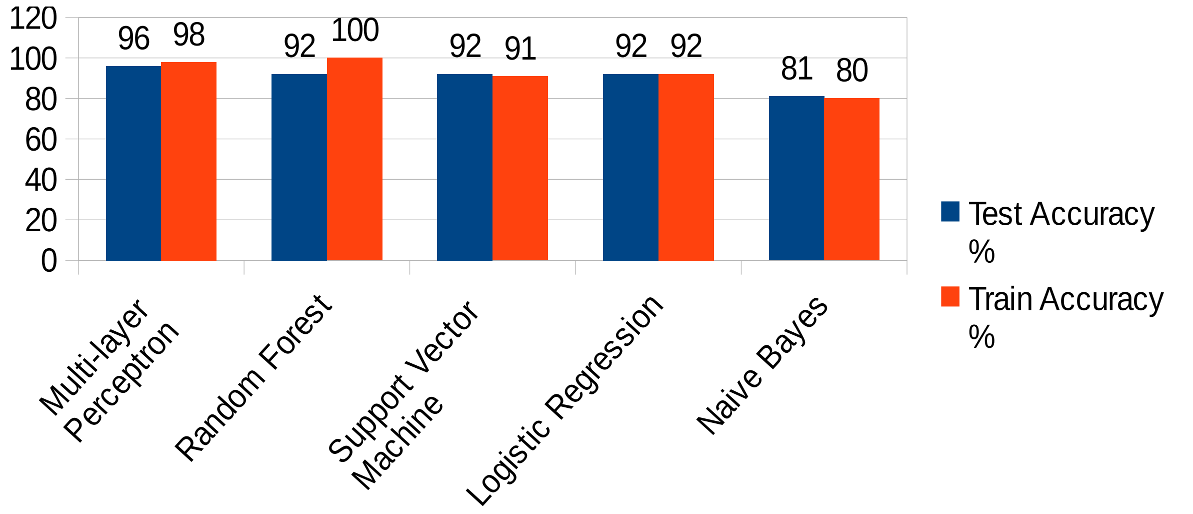
had used both public and private datasets. Jahangir et al (Jahangir, 2017), Mohapatra et al (Mohapatra, 2019), and Mishra et al (Mishra, 2020) had produced F1 score of 87%, 84% and 87%, respectively using Pima Indian Diabetes dataset, but our proposed model had produced higher F1 score of 95% in Mizoram population. Our MLP model has high accuracy of 96% in Mizo urban population when compared Xiong et al has accuracy of 87% in Chinese population [14]. Multi-layer perceptron neural network (MLPNN) model has an average performance with metrics: accuracy, f1 score and precision of 82% each when compared to our proposed model (Verma, 2020). The proposed model has highest precision of 93% when compared to all other MLP models, this clearly showed the proposed model has very low misclassification error and false positive rates (Table 2). In terms of accuracy Enhanced and Adaptive-Genetic Algorithm-Multilayer Perceptron (EAGA-MLP) has accuracy of 98%, which was mildly elevated accuracy than the proposed model. But EAGA-MLP model has low precision and F1 scores as compared to our MLP model (Table 2).

Age, sex, meat intake, saum, smoked meat, salt intake, smoking, sadha, tuibur, chewing paan and alcohol were significant epidemiological risk factors which cause diabetes in Mizo ethnic population. These factors can be unique for different populations and needs to be integrated with genetic changes to understand the main and confounding factors for T2DM (Lalrohlui, 2020). This was well-evidently shown using the proposed MLP model which has highest accuracy of 96% and the lowest misclassification rate of 4.5% (Table 1). Overall, the proposed model ensures that these epidemiological risk factors can be avoided for future management and prevention

Table 2: MLP model performance of existing state-of-art

S.No.	Models	Dataset	Accuracy	Precision	F1 score
1.	AutoMLP [22]	Pima Indian Diabetes dataset	89%	86%	87%
2.	MLP [14]	Chinese Urban Population	87%	-	-
3.	MLP [23]	Pima Indian Diabetes dataset	78%	83%	84%
4.	EAGA-MLP [24]	Pima Indian Diabetes dataset	**98%**	80%	87%
5.	MLPNN [25]	UCI repository	82%	82%	82%
6.	**Proposed MLP**	**Mizoram Urban population**	**96%**	**93%**	**95%**

Figure 5: Train Vs Test Accuracy of Five Models



of T2DM. It has a great potential to improve the understanding of factors those were etiological factors to cause T2DM.

CONCLUSION AND FURTHER STUDY

This study had aimed to find the best supervised machine learning algorithm for prediction of T2DM using epidemiological risk factors from Mizoram Urban population. The results showed that proposed MLP classifier has maximum ROC of 95%, and accuracy of 96% with low misclassification rates among logistic regression, naive bayes and support vector machine classifiers. This extracted knowledge will aid in early detection and diagnosis of T2DM, as the risk factors were truly invasive. In India, the race/ethnicity play a major role in T2DM including the lifestyle and dietary risk factors. For better understanding of these risk factors, we further need to identify the driving and confounding factors for causing T2DM in Mizo ethnicity.

AUTHORS CONTRIBUTION

Brindha Senthil Kumar, Vanlalawpuia R and Lal Hmingliana were the main researcher who played a role in formulating articles and processing data. Freda Lalrohlui, John Zohmingthanga and Lalrutapuii Hlawnmual played a role in the data collection and background formulation. Nachmimuthu Senthil kumar and Lal Hmingliana helped to formulate a framework for learning and discussion.

FUNDING

Not Applicable

ACKNOWLEDGMENT

The authors thank DBT, New Delhi for establishing the Bioinformatics Infrastructural Facility (No. BT/BI/12/060/2012 - BTISNeT), Mizoram University which was utilized for completing this work. The authors also thank the patients

who gave consent for the questionnaire data and samples as well as to the Lab technicians for their contribution.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests.

ETHICS APPROVAL FOR INVOLVING HUMAN PARTICIPANTS

The study protocol has been approved by ethical committees of Civil Hospital, Aizawl (B.12018/1/13-CH (A)IEC/39 dtd. 23/12/2015) and Human Ethical Committee, Mizoram University (MZU/IHEC/2015/006 dtd. 14/12/15).

INFORMED CONSENT

All participants signed the written informed consent.

REFERENCES

Alam, S., Hasan, M. K., Neaz, S., Hussain, N., Hossain, M. F. and Rahman, T. (2021). Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management. *Diabetology*, 2(2), 36-50.

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS ONE*, 12(7), e0179805.

Alshammari, R., Atiyah, N., Daghistani, T. and Alshammari, A. (2020). Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet. *Online Journal of Public Health Informatics*, 12(1), e11. doi:10.5210/ojphi.v12i1.10611.

Anjana, R. M., Deepa, M., Pradeepa, R., Mahanta, J., Narain, K., Das, H. K., et al. (2017). ICMR–INDIAB Collaborative Study Group. Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR-INDIAB population-based cross-sectional study. *Lancet Diabetes Endocrinol.*, 5(8), 585-596. https://doi:10.1016/S2213-8587(17)30174-2.

Belikov, A. V. (2019). Age-related diseases as vicious cycles. *Ageing Research Reviews*, 49, 11-26.

Carlsson, S., Andersson, T., Araghi, M., Galanti, R., Lager, A., Lundberg, *et al.* (2017) P, Norberg, M, Pedersen, NL, TrolleLagerros, Y, Magnusson, C (Karolinska Institute; Stockholm County Council; Stockholm; Skåne University Hospital, Malmö; Umeå University, Umeå; Sweden). Smokeless tobacco (snus) is associated with an increased risk of type 2 diabetes: results from five pooled cohorts. *J Intern Med.* 281(4), 398-406. doi: 10.1111/joim.12592.

Golden, S. H., Yajnik, C., Phatak, S., Hanson, R. L. and Knowler, W. C. (2019). Racial/ethnic differences in the burden of type 2 diabetes over the life course: a focus on the USA and India. *Diabetologia*, 62(10), 1751-1760. doi: 10.1007/s00125-019-4968-0.

Hunter, J. D. (2007). Matplotlib: A 2D graphic environment. *Computing in Science and Engineering*. 9(3), 90-95.

Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K. and Nawaz, R. (2017). An expert system for diabetes prediction using auto tuned multi-layer perceptron. 2017 Intelligent Systems Conference (IntelliSys). 722-728. doi: 10.1109/IntelliSys.2017.8324209.

Kaur, H. and Kumari, V. (2018). Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach. *Applied Computing and Informatics*. doi.org/10.1016/j.aci.2018.12.004.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A. and Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord.*, 19**,** 101. https://doi.org/10.1186/s12902-019-0436-6.

Lalrohlui, F., Sharma, V., Sharma, I., Singh, H., Kour, G., Sharma, S., *et al.* (2020). Candidate gene association study of UCP3 variant rs1800849 with T2D in Mizo population of Northeast India. *Int J Diabetes Dev Ctries*. 40, 513-517. doi.org/10.1007/s13410-020-00812-9

Lalrohlui, F., Sharma, V., Sharma, I., Singh, H., Kour, G., Sharma, S., *et al.* (2020). MACF1 gene variant rs2296172 is associated with T2D susceptibility in Mizo population from Northeast India. *Int J Diabetes Dev Ctries*. 40, 223-226. https://doi.org/10.1007/s13410-019-00788-1.

Maddatu, J., Anderson-Baucum, E. and Evans-Molina, C. (2017). Smoking and the Risk of Type 2 Diabetes. *Translational Research*. 184, 101-107. doi: 10.1016/j.trsl.2017.02.004.

Maniruzzaman, M., Rahman, M. J., Ahammed, B., Abedin, M. M. (2020), Classification and prediction of diabetes disease using machine learning paradigm. *Health Inf Sci Syst.*, 8(1), 7. doi: 10.1007/s13755-019-0095-z.

Mishra, S., Tripathy, H. K., Mallick, P. K., Bhoi, A. K. and Barsocchi, P. (2020). EAGA-MLP-An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis. *Sensors (Basel)*. 20(14), 4036. https://doi: 10.3390/s20144036.

Mohapatra, S. K., Swain, J. K. and Mohanty, M. N. (2019). Detection of Diabetes Using Multilayer Perceptron. International Conference on Intelligent Computing and Applications. *Advances in Intelligent Systems and Computing*. 846. doi.org/10.1007/978-981-13-2182-5_11.

Ravaut, M., Sadeghi, H., Leung, K. K., Volkovs, M. and Rosella, L. C. (2019). Diabetes mellitus forecasting using population health data in Ontario, Canada. *Proceedings of Machine Learning Research*. 85, 1-18.

Sami, W., Ansari, T., Butt, N. S. and Hamid, M. R. A. (2017). Effect of diet on type 2 diabetes mellitus: A review. *Int J Health Sci (Qassim)*. 11(2), 65-71.

Standl, E., Khunti, K., Hansen, T. B. and Schnell, O. (2019). The global epidemics of diabetes in the 21st century: Current situation and perspectives. *European Journal of Preventive Cardiology*. 26(2_suppl), 7-14.

Steyn, N. P., Mann, J., Bennett, P. H., Temple, N., Zimmet, P., Tuomilehto, J., Lindstrom, J. and Louheranta, A. (2004). Diet, nutrition and the prevention of type 2 diabetes. *Public Health Nutrition*. 7(1A), 147-165.

Thorp, A. A., Owen, N., Neuhaus, M. and Dunstan, D. W. (2011). Sedentary behaviors and subsequent health outcomes in adults a systematic review of longitudinal studies, 1996-2011. *American Journal of Preventive Medicine*. 41(2), 207-215.

Verma, G. and Verma, H. (2020). A multilayer perceptron neural network model for predicting diabetes. *International Journal of Grid and Distributed Computing*. 30, 1018-1025.

Wu, Y., Ding, Y., Tanaka, Y. and Zhang, W. (2014). Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. *Int J Med Sci*., 11(11), 1185-1200.

Xiong, X. L., Zhang, R. X., Bi, Y., Zhou, W. H., Yu, Y. and Zhu, D. L. (2019). Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-sectional Retrospective Study in Chinese Adults. *Curr Med Sci.,* 39(4), 582-588.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Front Genet*. 9, 515. doi:10.3389/fgene.2018.00515.