MINI REVIEW
published: 10 May 2022
doi: 10.3389/fcomp.2022.835242

Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques

B. Shamreen Ahamed, Meenakshi Sumeet Arya and Auxilia Osvin Nancy V*

Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai, India

The technological advancements in today's healthcare sector have given rise to many innovations for disease prediction. Diabetes mellitus is one of the diseases that has been growing rapidly among people of different age groups; there are various reasons and causes involved. All these reasons are considered as different attributes for this study. To predict type-2 diabetes mellitus disease, various machine learning algorithms can be used. The objective of using the algorithm is to construct a predictive model to critically predict whether a person is affected by diabetes. The classifiers taken are logistic regression, XGBoost, gradient boosting, decision trees, ExtraTrees, random forest, and light gradient boosting machine (LGBM). The dataset used is PIMA Indian Dataset sourced from UC Irvine Repository. The performance of these algorithms is compared in reference to the accuracy obtained. The results obtained from these classifiers show that the LGBM classifier has the highest accuracy of 95.20% in comparison with the other algorithms.

Keywords: prediction, machine learning, classifiers, accuracy, comparison

OPEN ACCESS

Edited by:
Bai Xue,
Institute of Software (CAS), China

Reviewed by:
Bohua Zhan,
Institute of Software (CAS), China
Qiuye Wang,
Institute of Software (CAS), China

Correspondence:
B. Shamreen Ahamed
shamu1502@gmail.com

Specialty section:
This article was submitted to Theoretical Computer Science, a section of the journal Frontiers in Computer Science

Received: *14 December 2021*
Accepted: *08 April 2022*
Published: *10 May 2022*

Citation:
Ahamed BS, Arya MS and Nancy V AO (2022) Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques.
Front. Comput. Sci. 4:835242.
doi: 10.3389/fcomp.2022.835242

INTRODUCTION

Diabetes mellitus (DM) is considered as a chronic disease that has been affecting people of all age groups. The exact cause of the disease is still unknown. However, some of the factors or causes include age, family history, other relative diseases, pregnancy, fluctuating glucose levels, blood pressure, etc. (Dash et al., 2019). Diabetes is a disease that can be controlled under medication; however, a complete cure through medicines is not possible as of today. Diabetes can belong to one of the four broad categories, such as type-1, type-2, gestational diabetes, or prediabetes (Nibareke and Laassiri, 2020). There are some sub-types classified under these four categories as well. "Type-1 diabetes" is also known as "insulin-dependent diabetes," which occurs when the insulin release cell is damaged and unable to produce insulin (Martinsson et al., 2020). In "type-2" diabetes, adequate amount of insulin is not produced in the body (Wang et al., 2015). This commonly happens at an average above age of 40 years. The "gestational diabetes (GDM)" occurs mostly during pregnancy. The last one among the main four categories, "prediabetes," occurs when the blood sugar level is higher than normal but not as high as type-2 diabetes (Mujumdar and Vaidehi, 2019).

In the recent years, many researchers are using the concept of machine learning to predict the DM disease. Some of the commonly used algorithms include logistic regression (LR), XGBoost (XGB), gradient boosting (GB), decision trees (DTs), ExtraTrees, random forest (RF), and light gradient boosting machine (LGBM). Each classifier has its own advantages over the other classifiers (Prabha et al., 2021). However, the classifier that gives the highest accuracy is determined in implementation.

This study is divided into different sections as follows: Section Related Works represents the related works in DM. Section Theoretical Concepts of the Classifiers determines the theoretical concepts of the various algorithms used. Section Results and Discussion determines the architecture and implementation of the classifiers. Section Conclusion and Future Work explains the conclusions and future works of the study.

RELATED WORKS

The following researchers have used the concept of machine learning for predicting DM disease.

Khaleel and Al-Bakry (2021) have created a model to detect whether a person is affected with DM disease. The concept of machine learning (ML) is used for the detection procedures. The PIMA dataset is used for the study. The algorithms used are LR, Naive bayes (NB), and K-nearest neighbour (KNN). The accuracy obtained are 94, 79, and 69% from these algorithms. The measures such as precision, recall, and F-measure are taken into consideration and LR is considered to produce the highest accuracy.

Ahmed et al. (2021) have used ML algorithms, namely, DT, KNN, NB, RF, GB, LR, and support vector machine (SVM) for predicting DM. Preprocessing techniques, such as label–encoding–normalization, are used to increase the accuracy. Two different datasets are used. One dataset provides the highest accuracy for SVM with 80.26% and for the second dataset, the highest accuracy is given by DT and RF with 96.81%.

Maniruzzaman et al. (2018) have used the ML technique based on risk-stratification is developed, optimized and evaluated. Features are optimized using six feature selection techniques. Then PIMA Indian diabetes dataset (PIDD) is used. The 10 different classifiers are used. Both RF selection and RF classification techniques yield an accuracy of 92.26%.

Kumari et al. (2021) have used two datasets including PIDD and breast cancer dataset, which were taken from the UC Irvine (UCI) Repository. Three ML classifiers are used for prediction. They are RF, LR, and Naive Bayes. The accuracy obtained is the highest for both datasets with a percentage of 79.08% for PIMA data and 97.27% for breast cancer data using soft voting classifier.

Tigga and Garg (2020) have developed a prediction model for DM disease. A dataset was collected for the study consisting of 952 instances and 18 attributes. The PIMA dataset was also used. The machine learning classifiers used are RF, LR, KNN, SVM, NB, and DT. The accuracy obtained was the highest for RF with a percentage of 94.10% for collected data and 75% for PIMA dataset.

Diwani and Sam (2014) have developed a prediction model using 10-fold-cross-validation on the training and testing data. The Waikato environment for knowledge analysis tool has been used along with Naive Bayes and DTs algorithm. The accuracy obtained is the highest for Naive Bayes with 76.30%.

Butt et al. (2021) have proposed a machine learning based approach for early-stage identification, classification, and prediction of diabetes disease. The PIMA Indian dataset has been used. The classifiers used are RF, multilayer perceptron (MLP) and LR. The accuracy obtained is highest for MLP with 87.26%.

THEORETICAL CONCEPTS OF THE CLASSIFIERS

The various classifiers that are used is explained in the following sub-sections.

Logistic Regression

It is a statistics-based model that uses logical function to develop a binary-dependent variable. The relationship between dependent and independent variables is estimated based on probabilities (Diwani and Sam, 2014). The dependent variable is categorical in this method. Mathematically it is expressed as follows (Kaur and Chhabra, 2014):

$$h_\theta(x) = P(Y = 1|X; theta)$$

The probability that $Y = 1$ given X which is given as "*theta*"

$$\mathrm{P}\left(\mathrm{Y} = 1 \mid \mathrm{X};\ \mathrm{theta}\right) + \mathrm{P}\left(\mathrm{Y} = 0 \mid \mathrm{X}; \mathrm{theta}\right) = 1$$

The XGBoost

It is the implementation of gradient boosted DTs that are created sequentially. An important feature is its weights. Each individual variable is assigned a particular weight that are given to the DTs to obtain the results (Butt et al., 2021). The prediction scores of each individual DT is given by

$$\hat{y}_i = \sum_{k=1}^{K} f_k \epsilon F$$

where the number of trees is denoted by k, the functional space is given as f, and the *possible set available is given as F* (Patil et al., 2019).

Gradient Boosting

Many weak learners are combined into a predictive model typically in the form of DTs (Sehly and Mezher, 2020). It is mainly used when we want to decrease the bias error. A gradient-descent technique is chosen to obtain values of the coefficients (Posonia et al., 2020).

The loss function used is $(y1 - y1')^2$. $y1$ is the actual value and $y1'$ is the final predicted value by this model. So $y1'$ is replaced with $G_n(X)$, which represents the actual target (Ke et al., 2017). It is mathematically expressed as follows:

$$\begin{aligned} G_{n+1}(X) &= G_n(X) + \gamma_n H1(x, e_n) \\ \mathrm{L1} &= \left(y1 - y1^{'}\right)^{2} \\ \mathrm{L1} &= (Y - G_n(x))^2 \end{aligned}$$

Decision Trees

It is a supervised-learning algorithm (Islam et al., 2020). It works with categorical and continuous input and output variables. It is used to represent whether it belongs to classification or regression procedures (Chen and Guestrin, 2016). The types of DTs are as follows: ID3, ID 4.5, CART, and CHAID. The measures used on DT are as follows: Entropy, Gini index, and

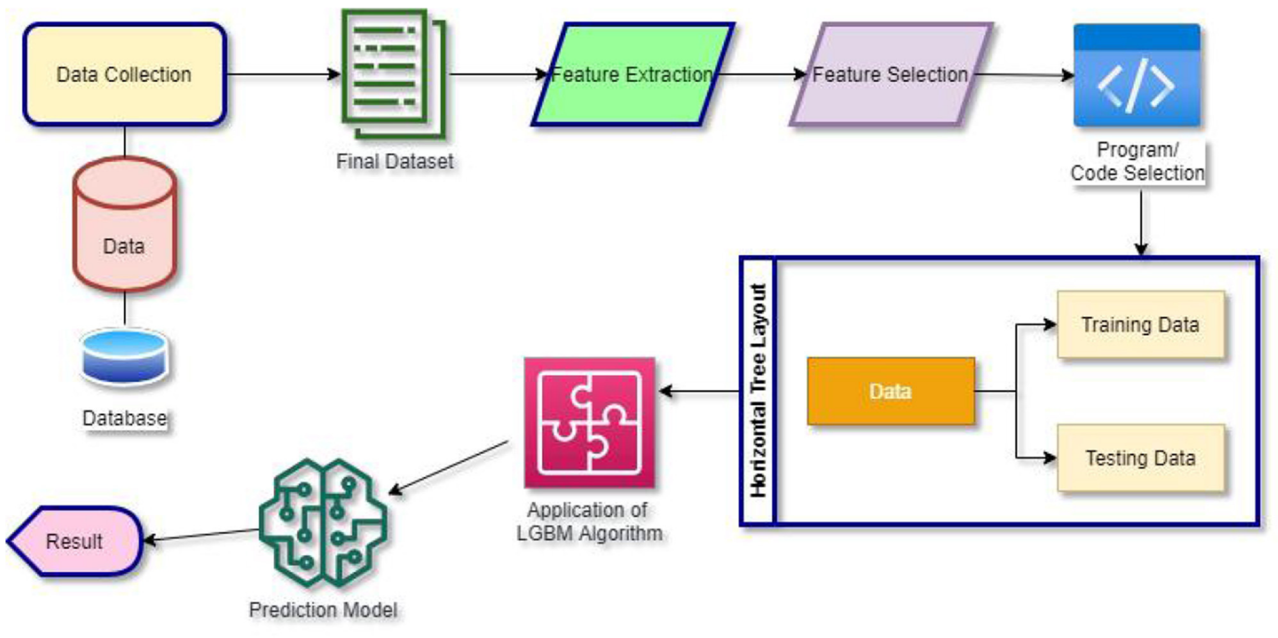

FIGURE 1 | Architectural design.

standard deviation (Khanam and Foo, 2021). It is mathematically calculated as follows (Ambigavathi and Sridharan, 2018):

$$\text{Entropy} = -\sum_{i1=1}^{n1} p_{i1}^{*} log(p_{i1})$$

$$\text{Gini Index} = 1 - \sum_{i1=1}^{n1} {p_{i1}}^{2}$$

Extra Trees

Extra trees (ETs) are also called as "extremely randomized trees classifier." It is a type of "ensemble learning technique" which combines many decorrelated DTs to result as a single tree classification (Chen et al., 2017). It differs from RF in a way in which DTs are built. The entropy is calculated as follows:

$$\text{Entropy(S1)} = \sum_{i1=1}^{c1} -\mathrm{p}_{i1}\log_{2}\mathrm{p}_{i1}$$

where the number of unique class labels is given as $c1$, the proportion of rows with output label is given as p_{i1} (Sisodia and Sisodia, 2018).

Then the "information gain" is calculated using the following formula (Ke et al., 2017):

$$\text{Gain(S1, A)} = \text{Entropy(S1)} - \sum_{\mathrm{v}\epsilon \mathrm{Values(A)}} \frac{|\mathrm{S1_V}|}{|\mathrm{S1}|} \text{Entropy}(\mathrm{S1_v})$$

Random Forest

The RF combines the output of multiple DT to reach a single result. The DT is taken as a base and row sampling as well as column sampling. The number of base learners is increased and the variance is decreased or *vice versa*. For cross-validation, K can be used. It is considered as an important bagging method (Mamuda and Sathasivam, 2017).

Random Forest =DT (base learner) + bagging (Row sampling with replacement) + feature
bagging (column sampling) + aggregation
(mean/median, majority vote)

Light Gradient Boosting Machine

The performance of LGBM is considered to be high-performance and is represented as "GB framework" based on DT algorithm (Ahamed and Arya, 2021). It is majorly used for classifying and ranking. It splits the tree leaf-wise with best-fit. It can be measured using the data improvement technique and can be given by calculating the variance after segregating (Zhu et al., 2020). It can be represented as follows:

$$\mathrm{Y1} = \mathrm{Base\_Tree\ (X1)} - \mathrm{lr1}^{*}\mathrm{Tree1\ (X1)} - \mathrm{lr1}^{*}\mathrm{Tree2\ (X1\ )}\ldots.$$

System Architecture

The data needed for the study are initially collected and stored in the database. The dataset PIMA is taken from UCI Repository for execution. The dataset is then pre-processed using different exploratory data analysis techniques. The dataset is divided into "training data" and "testing data." The various algorithms mentioned are then compared and the best working algorithm producing the highest accuracy is taken as the best predictive model for predicting DM disease. The architectural structure depicted in **Figure 1**.

TABLE 1 | Accuracy percentage.

Dataset	Logistic regression	XGB classifier	Gradient boosting classifier	Decision tree	Extra trees classifier	Random forest	LGBM
PIMA Indian dataset	75.20%	83.30%	94.10%	94.40%	94.60%	94.80%	95.20%

RESULTS AND DISCUSSION

The results and accuracy percentage calculated are given in the form of a table (**Table 1**).

The algorithms considered are LR, XGB, GB, DT, ET, RF, and LGBM. The accuracy obtained is the highest for LGBM with 95.2%.

CONCLUSION AND FUTURE WORK

These discussions here were considered and we identified that "LGBM algorithm" worked best for the dataset taken by producing an accuracy that was higher in comparisons with the other algorithms. However, in future, different dataset can be taken and compared with the different classifiers to classify which algorithm can produce the best result. Also, the parameters using in LGBM can be further finetuned and an advanced LGBM algorithm can be used and the prediction accuracy percentage can be increased.

AUTHOR CONTRIBUTIONS

BA and MA: material preparation, data collection, analysis, resources, and writing—review and editing. BA: first draft of the manuscript and investigation. MA: conceptualization, supervision, and visualization. AN: coding and idea of research. All authors contributed to the study conception and design, involved in the idea for the article, performed the literature search, data analysis, drafted, and critically revised the work.

REFERENCES

Ahamed, B. S., and Arya, M. S. (2021). Prediction of Type-2 diabetes using the LGBM classifier methods and techniques. *Turk. J. Comput. Math. Educ*. 12, 223–231. Available online at: https://www.proquest.com/docview/2622815314

Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., et al. (2021). Machine learning based diabetes prediction and development of smart web application. *Int. J. Cogn. Comp. Eng*. 2, 229–241. doi: 10.1016/j.ijcce.2021.12.001

Ambigavathi, M., and Sridharan, D. (2018). "Big data analytics in healthcare," in *IEEE Tenth International Conference on Advanced Computing (ICoAC)*, 269–276. doi: 10.1109/ICoAC44903.2018.8939061

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., and Sherazi, H. H. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *J. Healthc. Eng*. 2021, 9930985. doi: 10.1155/2021/9930985

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi: 10.1145/2939672.2939785

Chen, W., Chen, S., Zhang, J. H., and Wu, T. (2017). "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)*, 386–390. doi: 10.1109/ICSESS.2017.8342938

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *J. Big Data* 6, 54. doi: 10.1186/s40537-019-0217-0

Diwani, S. A., and Sam, A. (2014). Diabetes forecasting using supervised learning techniques. *Adv. Comp. Sci. Int. J*. 3, 10–18. Available online at: http://www.acsij.org/acsij/article/view/156

Islam, M. S., Qaraqe, M. K., Abbas, H. T., Erraguntla, M., and Abdul-Ghani, M. (2020). "The prediction of diabetes development: a machine learning framework," in *2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering, MECBME 2020* (IEEE Computer Society). doi: 10.1109/MECBME47393.2020.9292043

Kaur, G., and Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. *Int. J. Comp. Appli*. 98, 13–17. doi: 10.5120/17314-7433

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: a highly effiicient gradient boosting decision tree," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

Khaleel, F. A., and Al-Bakry, A. M. (2021). Diagnosis of diabetes using machine learning algorithms. *Mater. Today Proc*. doi: 10.1016/j.matpr.2021.07.196

Khanam, J. J., and Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Exp*. 7, 432–439. doi: 10.1016/j.icte.2021.02.004

Kumari, S., Kumar, D., and Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int. J. Cogn. Comp. Eng*. 2, 40–46. doi: 10.1016/j.ijcce.2021.01.001

Mamuda, M., and Sathasivam, S. (2017). "Predicting the survival of diabetes using neural network," in *Proceedings of the AIP Conference Proceedings* (Bydgoszcz), 40–46. doi: 10.1063/1.4995878

Maniruzzaman, M., Rahman, M., Al-MehediHasan, M., Suri, H. S., Abedin, M., El-Baz, A., et al. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *J. Med. Syst*. 42, 92. doi: 10.1007/s10916-018-0940-7

Martinsson, J., Schliep, A., Eliasson, B., and Mogren, O. (2020). Blood glucose prediction with variance estimation using recurrent neural networks. *J. Healthc. Inform. Res*. 4, 1–18. doi: 10.1007/s41666-019-00059-y

Mujumdar, A., and Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Proc. Comp. Sci*. 165, 292–299. doi: 10.1016/j.procs.2020.01.047

Nibareke, T., and Laassiri, J. (2020). Using big data-machine learning models for diabetes prediction and flight delays analytics. *J. Big Data* 7, 78. doi: 10.1186/s40537-020-00355-0

Patil, M. K., Sawarkar, S. D., and Narwane, M. S. (2019). Designing a model to detect diabetes using machine learning. *Int. J. Eng. Res. Technol*. 8, 333–340. Available online at: https://www.ijert.org/designing-a-model-to-detect-diabetes-using-machine-learning

Posonia, A. M., Vigneshwari, S., and Rani, D. J. (2020). "Machine learning based diabetes prediction using decision tree J48," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 498–502. doi: 10.1109/ICISS49785.2020.9316001

Prabha, A., Yadav, J., Rani, A., and Singh, V. (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. *Comp. Biol. Med*. 136, 104664. doi: 10.1016/j.compbiomed.2021.104664

Sehly, R., and Mezher, M. (2020). "Comparative analysis of classification models for pima dataset," in *International Conference on Computing and Information Technology (ICCIT-1441)*, 1–5. doi: 10.1109/ICCIT-144147971.2020.9213821

Sisodia, D., and Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Proc. Comp. Sci*. 132, 1578–1585. doi: 10.1016/j.procs.2018.05.122

Tigga, N. P., and Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Proc. Comp. Sci*. 167, 706–716. doi: 10.1016/j.procs.2020.03.336

Wang, F., Stiglic, G., Obradovic, Z., and Davidson, I. (2015). Guest editorial: special issue on data mining for medicine and healthcare. *Data Min. Knowl. Disc*. 29, 867–870. doi: 10.1007/s10618-015-0414-1

Zhu, T., Li, K., Chen, J., Herrero, P., and Georgiou, P. (2020). Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *J. Healthc. Inform. Res*. 4, 308–324. doi: 10.1007/s41666-020-00068-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.