Proceedings of CONF-CIAP 2026 Symposium: Applied Mathematics and Statistics
DOI: 10.54254/2753-8818/2026.CH31241

Machine Learning in the Field of Diabetes Prediction

Peixuan Han

School of Computer Science, University of Nottingham Ningbo China, Ningbo, China
ssyph3@nottingham.edu.cn

Abstract. Diabetes prevalence is a serious issue faced by all human beings. Tackling this issue can greatly help to reduce the financial burden on governments of medical expenditure to deal with diabetes patients and improve personal health by avoiding negative effects and complications brought by diabetes. To solve this problem, combining machine learning with the construction of diabetes prediction model has proved to be effective and is expected to develop rapidly in the future. This paper introduces five techniques commonly utilized in the construction of diabetes prediction model based on machine learning in detail, including Feature Ranking, AUC and C-statistics, Remote Framework for Diabetes Prediction, Gradient Boosting Algorithm and Model Evaluation Metrics. Based on these techniques, researchers can efficiently conduct relevant studies towards building diabetes prediction model with optimal machine learning algorithm selections. In the future, by further exploration on enhancing the combinations of prediction model with various machine learning algorithms and medical methodologies, the more accurate and effective ML-based diabetes prediction model can be built.

Keywords: Machine learning, Diabetes prediction, Personal health.

1. Introduction

Diabetes has become a serious problem worldwide in recent years and is expected to grow consistently at a high speed in the future. For example, based on the Centers for Disease Control and Prevention (CDC)'s estimation, by 2045, there will be 693 million people with diabetes all over the world [1]. This puts great pressure on the countries' financial burden and raises people's concerns about personal health. In the previous estimation, CDC predicted more than 327 billion dollars spent to treat diabetes in 2017 [1]. Besides the great financial burden, diabetes has a negative impact on personal health including unexplained weight loss, fatigue and difficulty to concentrate. Most importantly, the diabetes or its complications can even lead to death. To avoid these negative effects, it is important to predict and diagnose diabetes early. By using machine learning approaches, better performance can be achieved in predictive analytics compared to other methods such as statistical data analysis in the situation of larger data sets are provided [2]. Therefore, applying machine learning in the field of diabetes prediction helps to deal with a larger scale of data more efficiently and poses great significance in personal health. A group of researchers conducted data analysis on nearly 140 thousand participants and used machine learning models such as random forest and decision tree to build predictive models for type 2 diabetes [1]. Finally, they showed different

DOI: 10.54254/2753-8818/2026.CH31241

features of different machine learning models, for example, although the highest accuracy was delivered by neural network model, in some certain cases, people would choose decision tree model because of its highest sensitivity [1]. Based on the models, the researchers verified the risk factors discovered previously and found that for type 2 diabetes, sleeping duration and frequency can be new possible risk factors [1]. In order to construct a system with more personalized modeling and prediction process, the researchers described the Artificial Intelligence sub-systems of the SmartWork system which utilize data mining techniques and investigated a number of machine learning algorithms to determine the one with the best performance to be included in the SmartWork system [2]. Their research is also the first to evaluate numerous machine learning models and give personalized guidance towards lifestyle changes and risk prediction of type 2 diabetes occurrence [2]. Based on their analysis and consequences, they found that it is effective to use ensemble methods to predict type 2 diabetes [2]. Some researchers identified the people who have possibility of diabetes by certain risk factors evaluated by several machine learning models [3]. After identification, they compared different machine model's performance and concluded that in the 2 cases tested, applying SMOTE is an effective way to improve the model's performance. The researchers reviewed a number of published research papers to investigate how to detect diabetes at an early stage using machine learning algorithms [4]. After reviewing, they also planned for the future exploration including utilizing different metrics to create more precise diabetic diagnosis.

This article will mainly focus on how to predict diabetes based on machine learning models and algorithms. In the rest of the article, different machine learning models and algorithms will be compared and described in detail, and their effectiveness towards the diabetes prediction will be assessed. By assessing and evaluating, the paper can clearly demonstrate how machine learning helps to predict diabetes at an early stage and the mechanisms that make machine learning methods so effective in this field. In the future, on the basis of this study, people will learn the proper way of using machine learning in the field of diabetes prediction and construct more precise and accurate systems to avoid and intervene in diabetes, thus improving people's personal health and reducing financial burden for medical system.

2. Techniques overview

2.1. Feature ranking

To construct an efficient diabetes rick prediction model, it is significant to evaluate the importance of features (gender, age, obesity, etc.) and rank them properly such that the researchers can have a good basis for model training and validation.

Firstly, Pearson Correlation is deployed to determine the relationship between features and diabetes, which will return a Pearson correlation coefficient in the value from -1 to 1. For example, in this stage, researchers found that the symptom of polyuria had a solid correlation of nearly 0.7 to diabetes [3].

There are two popular machine learning methods utilized in the stage of feature ranking – Naive Bayes and Random Forest. Naive Bayes determines the probability of a feature given a feature x_i in the class c (assuming the high-level independence of features) by the formula $P x_{i1}, \ldots, x_{in} \left| c = \Pi_{j=1}^{n} P x_{ij} \right| c$. Based on this result, the features' accuracy and efficiency of predicting the output results can be assessed. Random Forest employs Geni impurity to generate decision trees that evaluate the features' capabilities of dividing instances into 2 classes with the best

DOI: 10.54254/2753-8818/2026.CH31241

efficiency [3]. After observing the coefficients given by different methods, the similarities can be found in the ranking order of features [3].

The feature ranking results provide the medical professionals and doctors with opportunities to evaluate the features' risk potential and recognize the relationship between the features and diabetes with the help of machine learning methods such as Naive Bayes and Random Forest. This helps medical professionals to provide proper guidance and intervention to the patient with risk of diabetes more efficiently based on the features and risk factors by paying more attention to the features or risk factors with higher importance (higher feature ranking). To achieve better accuracy in the future, deep learning methods are expected to be utilized in this topic.

2.2. AUC and C-statistics

To build helpful databases for predicting diabetes, it is necessary to process the data in a proper way with the optimal machine learning methods. Choosing suitable machine learning methods needs an evaluation system.

Machine learning methods can be classified as supervised ones and unsupervised methods. The main difference is that, for supervised learning, there are known target variables for the algorithm to be trained to project the values, while the unsupervised learning has no 'determined answer', the algorithm tries to divide the types from unlabeled datasets [4]. An effective method to evaluate both supervised learning methods and unsupervised ones is to employ AUC and C-statistics

After extracting sufficient data, the AUC and C-statistics are employed to assess the performance of different machine learning methods. The probability of generating a positive prediction is calculated by the area under the C-statistics and the curve used to describe the features of the receiver [4]. In this way, the effectiveness of a model can be reflected directly by the AUC value. For example, if the AUC of a specific ML model is less than 0.5, it states that the positive prediction is generated by that model with a probability of less than 50 percent at a time. This indicates that the machine learning model with higher AUC value is more efficient and helpful in discrimination.

The relative AUC difference value is then used to conduct comparisons between different machine learning models. If the paper denotes the AUC of the evaluated ML model as a and the AUC of the ML model for comparison as b, then the relative AUC difference can be calculated by the formula $\frac{a-b}{b} \times 100\%$[4], which means that if researchers get a positive value of relative AUC difference value, the conclusion that the evaluated ML model outperforms the ML model for comparison can be made.

2.3. Remote framework for diabetes prediction

Ideally, diabetes can be predicted at an early stage by the data transferred remotely from personal health devices such as watches and processed within the framework combining with suitable machine learning methods.

After collecting and processing the data, four machine learning algorithms are employed to predict the outcome as diabetes or non-diabetes [5]. The first ML algorithm used in the framework is k-nearest neighbours (k-NN). The k-NN method is regarded as a simple but efficient supervised learning algorithm since it determines the new unknown instances simply by the major labels of their k-nearest neighbours without complex computation. This attribute also renders k-NN method high-speed and efficient. The second ML algorithm is logistic regression (LR). Utilizing a best fit model, LR outputs the probability of a binary result, in this case, diabetes or non-diabetes. Gaussian Naive-Bayes (NB) is another ML algorithm used in the framework. NB is capable of determining

DOI: 10.54254/2753-8818/2026.CH31241

the independence between features included as strong or naïve [5]. As a prediction technique, NB applies a combination of Bayes theorem and Gaussian distribution to get the accurate predictions [5]. The last ML algorithm applied in the framework is support vector machines (SVM) [5]. Assisted by support vectors, SVM divides the data instances linearly with the best estimation. The data instances are plotted in the feature space based on the number of features.

With the four machine learning algorithms, the likelihood of diabetes can be demonstrated accurately based on the information collected by personal health devices and the medical professionals will get alerts to intervene at proper time, avoiding further development of the diabetes, greatly boosting the personal health condition of people with potential risk.

2.4. Gradient boosting algorithm

Another useful supervised machine learning algorithm to help predict the binary (diabetes or non-diabetes) results for the diagnosis of diabetes is the Gradient Boosting algorithm [6].

On the basis of gradient descent method, the Gradient Boosting algorithm shows an excellent classification abilities comparable to regression for this prediction model [6]. Even less competent learner becomes stronger by Gradient Boosting's unique feature of declining variance and bias [6]. To realize this feature, after preprocessing the data given by some basic models, Gradient Boosting algorithm creates additional trees based on the theory of Optimization Technique which utilizes differential loss function to calculate the coefficients' values of the function when the cost function's value is minimized. Through this way, though more time is consumed compared to Random Forest, the randomly built trees in Random Forest gets correction sequentially based on previous trees' error rates, which integrates the less competent learners into a single strong learner, enhancing the accuracy and precision of the prediction for the diagnosis of diabetes.

2.5. Model evaluation metrics

To efficiently evaluate the effectiveness of a specific machine learning method used in the diabetes prediction model, four metrics are frequently used to calculate the accuracy (the proportion of the correct predictions out of the total predictions) and precision (the proportion of the instances labelled as positive by classifier out of the total predictive positive instances) of the ML method: True Positives (TP), False Positives (FP), False Negatives (FN), True Negatives (TN). They are defined as follow respectively [7]:

TP represents people who were predicted to have diabetes and finally diagnosed with diabetes. FP represents people who were predicted to have diabetes but finally diagnosed with non-diabetes. FN represents people who were predicted to be non-diabetes but finally diagnosed with diabetes. TN represents people who were predicted to be non-diabetes and finally diagnosed with non-diabetes.

The accuracy and precision of a specific machine learning diabetes prediction model can be calculated as [7]:

$$Accuracy = (TN + TP)/(TN + TP + FN + FP) \tag{1}$$

$$Precision = TP/(TP + FP) \tag{2}$$

3. Analysis

The techniques introduced previously in this paper have now been widely used in building machine learning diabetes prediction models. To build an efficient and effective prediction model, firstly, it is

DOI: 10.54254/2753-8818/2026.CH31241

important to conduct the feature ranking [3]. The prediction model can be combined with remote framework to make predictions for diabetes anytime and anywhere and ensure on time intervention [5]. The research has shown that most machine learning algorithms perform better than the non-ML algorithms in the diabetes prediction model [4], implying that choosing accurate, precise and efficient ML algorithms for building prediction model plays an important role. For predicting the binary results for the diabetes prediction (diabetes or non-diabetes), it is highly recommended to implement Gradient Boosting algorithm for the reason that in the research comparing Gradient Boosting algorithm with other two commonly used classification algorithms Logistic Regression and Naive Bayes, Gradient Boosting demonstrated a higher reliability and correctness of discriminating data [6]. But this may not apply to all conditions and other machine learning models, to get the optimal solution for specific task in building diabetes prediction model, some machine learning model evaluation methods can be used to compare different ML methods. For example, a group of researchers employed AUC and C-statistics to assess different ML methods' performance in the context of processing data and found that for microvascular outcomes, neural network has the best performance with a mean AUC of 0.87, and for macrovascular outcomes, ensemble methods have the best performance with a mean AUC of 0.70 and if both conditions are considered, overall, the optimal machine learning method is random forest, which shows that it is recommended to process the data for predicting diabetes using random forest as the machine learning model [4]. To calculate the effectiveness of a ML algorithm in diabetes prediction model more directly, the model evaluation metrics can be utilized to compute the accuracy and precision of the ML algorithm [7]. After computation, the developers are able to choose the best ML algorithm based on the unique requirements of the specific parts of the diabetes prediction model to ensure the best effectiveness, efficiency and reliability.

4. Discussion

4.1. Challenges

Although a lot of research conducted, at present, people are still facing several difficulties of building an ideal diabetes prediction model with machine learning algorithms. For example, there is a lack of research on evaluating type 2 diabetes prediction model on elder office workers before, which makes the relevant studies hard to compare the performance of their machine learning algorithms with previous studies, leading to difficulties of evaluating the machine learning models in a proper way by comparing with other models [2]. In previous research, often the researchers only focused on diabetes risk factors analysis [2]. There are also some problems on the limited dataset size [2]. The relatively small dataset for type 2 diabetes among elder office workers inevitably places limitations on the research since the ML algorithm has better performance when there are sufficient training samples. For other specific targeted people, the similar situation and challenges also exist, if the certain dataset does not have sufficient training samples, it is very difficult to properly evaluate and choose optimal machine learning algorithms to build effective diabetes prediction model. In the future, more studies focusing on different special groups of people and different machine learning algorithms are encouraged.

4.2. Outlook

To improve the diabetes prediction model, it is reasonable to combine the predictive data mining methods equipped with various machine learning algorithms with some steps of building a decision

DOI: 10.54254/2753-8818/2026.CH31241

model, for example, prognosis, diagnosis and arrangement of treatment. Through testing and verifying these predictive data mining methods, it is practical to utilize them in clinical information system to construct a predictive model that generates wanted prediction based on the patient-specific information from the system [8]. This will help to prevent diabetes complications by recognizing the link between diabetes complications and exposure to individual risk factors and providing corresponding medical support during diabetes patients' treatments [8]. It is also significant to pay more attention to constructing predictive model for early-stage detection of diabetes risk [9]. Integrating diverse medical methodologies with diabetes prediction model has been proved successful from research about building predictive models based on tongue features and ML methods [10]. A significant higher accuracy can be achieved by combining tongue features with diabetes prediction model compared to prediction model only equipped with classical ML algorithms [10], indicating a bright future for diabetes prediction model powered by ML algorithms with further exploration of the diverse medical methodologies and features.

5. Conclusion

This paper focuses on the machine learning in the field of diabetes prediction. Firstly, the definition and negative effects of the diabetes are introduced, which clearly shows that why it is important to import some advanced technology such as machine learning to deal with the diabetes. The introduction part also gives some achievements realized by the machine learning methods, indicating that machine learning can be a helpful method in diabetes prediction. Then the techniques commonly used in implementing machine learning to build diabetes prediction model are explained in detail, including principles of ML algorithms, methods for evaluating different ML algorithms and frameworks to build diabetes prediction model based on ML methods. Based on the techniques introduced previously, the paper analyses the advantages of applying machine learning algorithms to diabetes prediction model and how to choose the optimal ML algorithm for different part of the diabetes prediction model.

At the end of the paper, some challenges towards the implementation of machine learning in diabetes prediction are discussed and possible solutions to the existent problems are raised. Then, the paper mentions some possible improvements that can be made to enhance the ML-based diabetes prediction model, such as integrating diverse medical methodologies. The future development of diabetes prediction model based on machine learning is believed to be realized by further combinations of prediction model with more advanced machine learning algorithms and currently underexplored medical methodologies.

References

[1] Xie, Z.D., et al. (2019) Building risk prediction models for type 2 diabetes using machine learning techniques. Preventing chronic disease 16: E130.

[2] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N. and Moustakas, K., (2021) Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction, " in IEEE Access, vol. 9, pp. 103737-103757, doi: 10.1109/ACCESS.2021.3098691.

[3] Dritsas, E., Trigka, M. (2022) Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. Sensors. 22, 5304. https: //doi.org/10.3390/s22145304

[4] Tan, K.R., et al. (2023) Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review. Journal of Diabetes Science and Technology 17.2: 474-489.

[5] Ramesh, J., Raafat A., and Assim S., (2021) A remote healthcare monitoring framework for diabetes prediction using machine learning." Healthcare Technology Letters 8.3: 45-57.

DOI: 10.54254/2753-8818/2026.CH31241

[6] Birjais, R., et al. (2019) Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences 1.9: 1112.
[7] Chou, C.Y., Hsu, D.Y., Chou, C. H. (2023) Predicting the Onset of Diabetes with Machine Learning Methods. J. Pers. Med. 13, 406. https: //doi.org/10.3390/jpm13030406
[8] Dagliati, A., et al. (2018) Machine learning methods to predict diabetes complications. Journal of diabetes science and technology 12.2: 295-302.
[9] Firdous, S., Wagai, G. A., Sharma, K. (2022) A survey on diabetes risk prediction using machine learning approaches. Journal of Family Medicine and Primary Care 11(11): p6929-6934 | DOI: 10.4103/jfmpc.jfmpc_502_22
[10] Li, J., et al. (2021) Establishment of noninvasive diabetes risk prediction model based on tongue features and machine learning techniques." International Journal of Medical Informatics 149: 104429.