Machine Learning, Ensembles, and Knowledge Graphs for Diabetes Prediction

Jiayi Jiang

Department of Statistics, College of Letters and Science, University of California, Davis, United States
judjiang@ucdavis.edu

Abstract. This article reviews the progress of machine learning in early prediction and risk identification of diabetes, focusing on three methods: traditional models (such as Logical Regression, SVM, RF, etc.), ensemble learning (such as Bagging, Boosting, Stacking and weighted voting) and Reasoning based on knowledge graph (KG). The traditional model has the advantages of robustness and strong interpretability, which is suitable for small samples and structured features. Ensemble learning further improves the accuracy and generalization ability through the complementarity of heterogeneous models. The knowledge atlas explicitly uses the entity relationship diagram for modeling, which takes into account interpretability and reasoning ability in the absence of features and incomplete knowledge. This article compares the application limitations of various methods in data utilization, feature engineering, category unbalance processing and model calibration. It points out that traditional methods rely on preprocessing and linear assumptions, integrated models need to achieve a balance between complexity and interpretability, and the cost of knowledge atlas construction and maintenance is high, but it is convenient for Physical decision-making support. From clinical applications perspective, it is recommended to strictly prevent information leakage, combine resampling and calibration indicators (AUC/PR-AUC, F1, Brier), introduce interpretation tools such as SHAP, and verify the robustness of the model on independent data. In a word, in order to achieve accurate prevention and intelligent management. It is helpful to seek the best balance between performance, complexity and interpretability according to tasks and resource conditions.

Keywords: Diabetes Detection, Knowledge Graph, Ensemble Learning, Disease Prediction.

1 Introduction

Diabetes is a chronic metabolic disease. The consumption of high-sugar foods such as desserts, milk tea, and processed products has increased markedly. The global incidence of diabetes also has continued to rise.

As pointed out by the World Health Organization (WHO), about 830 million people in the world suffer from diabetes [1]. This infomation suggests that continuous rise in

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_72

blood sugar levels because of body's resistance to insulin. Type 2 diabetes has the highest prevalence rate among adults. These diseases seriously affect the daily life of patients and lead to irreversible results.

As a highly prevalent, chronic, and preventable and controllable disease, in fact, early prediction and risk identification of diabetes have significant public health and clinical value. Early detection can delay disease progression, reduce medical expenses, and provide a basis for individualized intervention, significantly reducing the probability of confirmed diabetes. With the popular development of artificial intelligence and big data, diagnostic methods based on machine learning are becoming increasingly popular.

This review article mainly discusses three directions of intelligent prediction of diabetes, namely traditional machine learning methods, ensemble learning methods, and knowledge graph reasoning models. By comparing the performance of different methods in data utilization, model interpretability, and generalization ability, the applicable scenarios and limitations of various models are summarized.

To summarize the latest progress in improving prediction accuracy, enhancing model interpretability, and supporting clinical decision-making through cutting-edge methods such as ensemble model and knowledge graphs, thereby providing directions and references for the future realization of precise prevention and intelligent management of diabetes.

2 Background

Diabetes is a disease characterized by chronic hyperglycemia. Its pathological mechanism involves insulin secretion disorders and insulin resistance, leading to lipid and protein metabolism disorders and causing damage to multiple systems. Diabetes can be classified into 2 types based on its pathogenesis: type 1 diabetes is caused by the immune destruction of pancreatic islet cells, while type 2 diabetes is characterized by insulin resistance.

The onset of diabetes is multifactorial. Studies have shown that the prevalence of type 2 diabetes has been continuously increasing worldwide due to the widespread adoption of sedentary lifestyles and high-sugar, high-fat diets. In addition, family history, aging, a history of gestational diabetes, and hypertension are important risk factors.

Clinically, the typical manifestations of diabetes can be summarized as "three more and one less", namely polydipsia, polyphagia, polyuria, and weight loss. Moreover, long-term hyperglycemia can also lead to various chronic complications, including diabetic retinopathy (DR), diabetic nephropathy (DN), diabetic foot (DF), neuropathy, and cardiovascular diseases, which are often the main causes of disability and death.

The World Health Organization (WHO) define diabetes as a fasting plasma glucose (FPG) level of $\geq$7.0 mmol/L or a glycated hemoglobin (HbA1c) level of $\geq$6.5% [1]. Based on clinical data, the intelligent prediction and diagnosis of diabetes have become an important research direction in the field of artificial intelligence in medicine, which

also lays the foundation for the diabetes diagnosis method based on machine learning described in the following text.

3 Traditional Methods for Diabetes Diagnosis and Prediction

Traditional machine learning typically refers to a set of classic supervised learning algorithms, including Linear Regression, Logistic Regression, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Decision Tree, Random Forest (RF), Naive Bayes (NB), and k-Nearest Neighbor (k-NN) [2, 3, 4].

System comparisons on the Pima Indians Diabetes dataset show that these methods generally have a robust baseline performance. Khanam et al. evaluated multiple models such as LR, SVM, NB, kNN, and RF, and the overall accuracy was generally above 70%, with LR and SVM being in a relatively superior level, approximately 77%–78%. In the same study, the authors also trained a two-layer fully connected neural network, with 400 epochs of training and a learning rate of 0.01, and ultimately achieved the highest accuracy of 88.6%. The authors conducted feature selection through Pearson correlation screening, and finally retained Glucose, BMI, Insulin, Pregnancies, and Age as input variables [2].

In research on early diabetes prediction, Refat et al. compared various machine learning and deep learning models on a sample of 520 cases. The results showed that XGBoost had significant advantages in multiple indicators, with training and testing accuracy rates of approximately 99.99% and 99.0% respectively. The authors also emphasized the limitations of the experiment, as the sample size was small, and the statistical robustness of the conclusion still needed to be verified with a larger sample [3].

Joshi and Dhakal also used the Pima Indians Diabetes dataset and compared Logistic Regression with Classification Trees. After variable selection using AIC, BIC, and Mallows' Cp, the accuracy of Logistic Regression was approximately 78.26%. The decision tree structure was based on Glucose as the root node, indicating that blood sugar is an important indicator for distinguishing diabetes. BMI and Age also showed significant contributions. Regarding missing values, the authors replaced the 0 values in the records with the median of the corresponding variables and pointed out that this method might introduce statistical instability. Additionally, for feature dimensions with fewer features such as age, BMI, and glucose, their extrapolation ability is still limited [4].

Overall, traditional models have good interpretability and stable baseline performance in the diabetes prediction task. Logistic Regression and SVM are often used as controls or starting points [2, 3, 4]. However, these methods rely heavily on feature engineering and preprocessing, and have limited ability to represent complex nonlinear relationships. In structured data with small samples, ensemble tree models such as XGBoost can often further improve the effect, but their representativeness and generalizability still need to be carefully evaluated [3]. Moreover, when dealing with common missing and abnormal values in clinical data, if not handled properly, the performance and stability of traditional models will also decline [4]. Therefore, recent

research is also exploring more complex modeling machine learning models such as: deep learning, attention mechanisms, and knowledge graphs, hoping to further improve prediction performance while maintaining clinical interpretability.

4 Ensemble Learning

Ensemble Learning enhances prediction accuracy and stability by combining multiple base learners. Common methods include Bagging, Boosting, Stacking, and weighted voting. Generally, Bagging mainly reduces variance, Boosting focuses more on reducing bias, Stacking has a multi-layer structure, and optimizes the integration of heterogeneous base models in more complex machine learning scenarios. Weighted/soft voting aggregates based on the posterior probabilities output by each base model, balancing simplicity and robustness.

In the diabetes prediction scenario, Boosting is often regarded as the baseline and is an extremely effective method for predicting diabetes. Sai et al. integrated LightGBM and AdaBoost to develop a 2-type diabetes discrimination mechanism. In the context of imbalanced sample data, the authors evaluated the discrimination performance of the model using indicators such as ROC/AUC, and reported an overall accuracy of approximately 90.76%, demonstrating the competitiveness of the ensemble based on boosting in this task [5]. It is worth noting that the boosting method is more sensitive to difficult samples, and in practical applications, regularization is needed to prevent overfitting.

Stacking is a multi-layer fusion mechanism, distinct from simple voting or a single boosting process. It uses the prediction generated by the "0th layer" as a new feature and passes it to the "1st layer" meta-learner for re-learning, thereby integrating models with different inductive preferences at higher levels. Abdollahi and Nouri-Moghaddam proposed a hybrid stacking method combined with genetic algorithms (GA): on the one hand, they constructed two-layer Stacking using heterogeneous base models such as RF, SVM, kNN, and ANN; on the other hand, they used GA to search for feature subsets and model combinations to improve generalization ability. The authors reported in their data and settings that the diagnostic accuracy could reach approximately 98.8% - 99%, which outperformed single models and conventional integration [6]. However, the increase in model complexity also meant an increase in interpretability complexity, and it was necessary to be assisted by variable importance or local explanation methods [6].

In 2024, Li et al. simultaneously addressed the issues of class imbalance, parameter optimization, and interpretability within the Stacking framework: first, they compared various sampling strategies and used SMOTEENN for resampling and class balance; then, they optimized the key hyperparameters of XGBoost using GA (such as n_estimators, learning_rate, max_depth, etc.); finally, they incorporated LightGBM, RF, etc. into the two-layer Stacking and completed the integration with Random Forest as the meta-learner. In public metrics, this method achieved an AUC of 0.9888, Accuracy of 0.9482, and F1 of 0.9579 [7]. At the same time, through SHAP analysis, core variables such as Glucose, BMI, and Age were identified, balancing performance

and medical interpretability [7]. This approach improved the shortcomings of performance-oriented stacking fusion in terms of interpretability, adding an explanation framework to the overall design to enhance the clinical operability and practicality beyond optimization and resampling.

In summary, in the prediction of diabetes using structured medical data, the fusion model is not "the more complex, the better". Instead, it should be based on the goals and constraints. When the sample size is average, using the improvement method as the baseline is often the most stable. When it is necessary to incorporate multiple sources or heterogeneous data, Stacking can be added and appropriate GA model/feature selection can be conducted. When dealing with class imbalance and strong interpretability, the focus is on resampling, calibration, and SHAP and other explanation tools. Regardless of which paradigm is adopted, the ultimate goal is to balance the performance, complexity, and interpretability of the model.

5 Knowledge Graph

A knowledge graph (KG) is a data model that organizes entities and their relationships in a graph structure. By explicitly connecting entities and relationships, a graph can be constructed to cover multiple types of elements and depict the interactions between variables. The graph structure helps to identify correlations, enhance explanations, and support subsequent reasoning and decision-making [8-10].

In Wang's study, a knowledge graph was used to predict the risk of diabetes complications. The study did not use EMR because there was insufficient evidence. Instead, it used the knowledge graph of evidence-based medicine EMR and analyzed the data using Meta. This allowed the data to be sorted in order of risk levels as the framework for building the knowledge graph [8]. The authors mentioned that in the study, evaluations and analyses were conducted in terms of social, psychological and behavioral aspects, which made the study of type 2 diabetes complications more complete. Thus, more factors were obtained, including various diabetic retinopathy (DR), diabetic nephropathy (DN), diabetic foot (DF), depression (DD), bariatric surgery, myopia, lipid-lowering drugs, duration of lipid-lowering drugs, blood sugar control, disease duration, glycated hemoglobin, fasting blood sugar, hypertension, gender, smoking, etc. While completing the knowledge graph, Meta also enabled the quantification of each factor, divided into OR odds ratio and RR relative risk. It is worth noting that in the prediction model, the logistic regression was added to the model. Therefore, the knowledge graph is not static and can be personalized for risk prediction based on each individual's risk. Specifically, when applied to each patient, the relevant information of the patient is input, and the interactive system of Shiny can automatically predict the risk value, and the size of the nodes is used to explain and emphasize the size of the risk [8]. This is a very good method of combining models and graphs, and such a model has the strongest explanatory ability. This means that the higher the risk, the larger the node, and the visual effect is very obvious. In clinical practice, it can help doctors make references and assist in predicting the potential disease risks of the authors.

Another advantage of the knowledge graph is that it performs better than traditional models when dealing with missing features. Often, during data collection, some variables of the samples have empty features and no recorded values. Traditional models such as SVD, RF, and NN have limitations when encountering missing feature values. However, in the article by Li et al. on using knowledge graph reasoning to achieve risk prediction for diabetic macular edema (DME), they can solve the problem of sample missing features [9]. The motivation of the author's research is that traditional machine learning has a decline in accuracy when encountering missing values, but the reasoning of the knowledge graph can "any number of known feature data" for prediction [9]. They used 507 cases from the Ophthalmology Hospital of Tianjin Medical University as samples. During the data cleaning stage, 39 influencing factors were identified and further refined into 116. Using two nodes, namely whether or not having DME and 116 risk factors, they were matched. Therefore, the entire knowledge graph has a total of 223 relationship entries. The author first used Statistical rule reasoning, Correlation enhancement, and Improved correlation enhancement + generalized closeness for correlation tests. The results of these three tests were scored using AUC, Precision, and Ranking score. Finally, it was found that Improved correlation enhancement + generalized closeness performed better. To make the prediction more practical, the author also listed the prediction output based on the probability level, providing reminders for low risk, follow-up, further examination, and specialized examination. The application of the knowledge graph is very strong. By inputting any known disease history, the system can obtain the recommended results for DME [9].

Although the predictions of the knowledge graph have shown a relatively stable predictive ability in the absence of features, in medicine, inevitable structural deficiencies and feature disappearance will occur. This is because new knowledge is constantly emerging and the prior relationships are incomplete, so the graph will always be incomplete at some point. Therefore, a complete knowledge graph, with completed relationships, can enable sustainable predictions and allow more traditional models to be incorporated into the predictions. Singh [10] mentioned in the article that the existing medical datasets cannot fully and comprehensively summarize the relationships using traditional tables. The author's solution is to merge two diabetes knowledge graphs, KG-1, which comes from the previously mentioned traditional PIDD and KG-2, the medical ontology library. By combining clinical patient data and the semantic ontology perspective, a new merged graph was obtained [10]. The author stated that using link prediction (llink prediction) can complete the knowledge graph, and trained a dedicated annotated dataset, Diabetes-KG, to evaluate the new prediction algorithm [10]. To ensure the baseline level of the dataset, the author used mainstream embedding LP models, TransE, ComplEx, RotatE, and pRotatE to test Diabetes-KG. Using three common metrics - Hits@k, Mean Rank (MR), and Mean Reciprocal Rank (MRR) for comparison. The results showed that pRotatE performed best in MRR (0.74) and Hits@10 (85.87%) indicators. Subsequently, the author applied the same model to Diabetes-KG for empirical analysis. These embedding models also achieved stable prediction performance on the diabetes knowledge graph. This article proves that the

predictive ability of the knowledge graph is more advantageous than traditional models [10].

6 Limitation

6.1 Traditional Machine Learning Models

Traditional methods such as Logical Regression, SVM, and RF support vector machine and random forest are stable and easy to explain on small sample data. However, they are highly dependent on statistical hypotheses, manual feature selection and data preprocessing. It is difficult to capture complex nonlinear relationships due to these models usually assume that there is a linear relationship between features. For clinical data with missing values, label imbalances or heterogeneous characteristics, the robustness of these algorithms is poor.

6.2 Ensemble Models

Ensemble models such as Bagging, Boosting and Stacking are superior to traditional models in terms of accuracy and generalization performance. At the same time, they have poor interpretability, complex structure and large computing volume. These models require a large number of samples to prevent overfitting, and the diabetes data is still very limited. All of these will cause limiting their feasibility in practical clinical application.

6.3 Knowledge Graph Models

When certain characteristics are missing, the knowledge model can identify potential relationships related to diseases and maintain predictive ability. It is also worth noting that the construction cost is very high. Most medical knowledge atlases rely on manual creation, which is slow to update and has limited coverage. Due to the continuous development of medical knowledge, even if there is a complete algorithm, it still needs to be verified again.

7 Conclusions

This article reviews the research on diabetes prediction and diagnosis based on machine learning, mainly focusing on: traditional machine learning methods, fusion models, and knowledge graph reasoning methods. Traditional models have advantages in interpretability and robustness, and are suitable for small sample scenarios. The fusion model significantly improves prediction accuracy through integration optimization algorithms. Knowledge graphs can perform individualized reasoning under unstructured and missing data conditions.

Overall, with the development of artificial intelligence technology and the increase of medical data, diabetes prediction will no longer be limited to the improvement of algorithm performance, but will focus more on interpretation, verification, and practicality. This helps to achieve early screening and intervention, provide data support for clinical practice, and assist in the prevention and control of diabetes.

References

1. World Health Organization: Diabetes: coverage of treatment with glucose-lowering medication. The Global Health Observatory, [Online]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2379, last accessed 2025/11/6
2. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. ICT Express 7(4), Feb. 2021
3. Refat, M.A.R., Amin, Md.A., Kaushal, C., Yeasmin, M.N., Islam, M.K.: A Comparative Analysis of Early Stage Diabetes Prediction Using Machine Learning and Deep Learning Approach. IEEE Xplore, Oct. 01, 2021
4. Joshi, R.D., Dhakal, C.K.: Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. International Journal of Environmental Research and Public Health 18(14), 7346, Jul. 2021
5. Sai, M.J., Chettri, P., Panigrahi, R., Garg, A., Bhoi, A.K., Barsocchi, P.: An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. International Journal of Computational Intelligence Systems 16(1), Feb. 2023
6. Abdollahi, J., Nouri-Moghaddam, B.: Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction. Iran Journal of Computer Science 5, Mar. 2022
7. Li, W., Peng, Y., Peng, K.: Diabetes Prediction Model Based on GA-XGBoost and Stacking Ensemble Algorithm. PLoS ONE 19(9), e0311222, Sep. 2024
8. Wang, L., Xie, H., Han, W., Yang, X., Shi, L., Dong, J., Jiang, K., Wu, H.: Construction of a knowledge graph for diabetes complications from expert-reviewed clinical evidence. Computer Assisted Surgery 25(1), 29–35, 2020
9. Li, Z.-Q., et al.: Prediction of Diabetic Macular Edema Using Knowledge Graph. Diagnostics (Basel) 13(11), 1858, May 2023
10. Singh, S., Siwach, M.: Evaluating Diabetes Dataset for Knowledge Graph Embedding Based Link Prediction. Data & Knowledge Engineering 157, 102414, May 2025