Application of machine learning in diabetes prediction based on electronic health record data analysis

Zihan Yang

School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, United Kingdom

Abstract. With the application of electronic health records (EHRs) in the medical field, the use of machine learning to predict disease has become one of the important research hotspots in the healthcare industry. This study introduces an improved machine learning model specifically designed to predict diabetes risk, with the aim of improving the accuracy of predictions. The purpose of the study is not only to refine the model, but also to evaluate the performance of the model according to the experimental results. The integrated model was used in this experiment, and the prediction accuracy of diabetes reached 77.7%, showing strong generalization ability on the test data set. These results show that the model performs well at predicting diabetes, but there is still room for further improvement. While presenting the current research results, this study also Outlines future research directions, focusing on further improving the accuracy and reliability of the model. Th is research contributes to the development of machine learning in healthcare, specifically improving disease prediction models through advanced data analysis techniques.

1 Introduction

The adoption of electronic health record (EHR) systems is becoming more widespread, and the use of machine learning in these systems has also grown significantly. This includes predicting the patient's condition, estimating the likelihood of disease, and identifying the numerous factors that have the greatest impact on the patient's health.

This model could not only help patients self-assess their risk of disease, but also reduce the burden on doctors. Because the database contains a large amount of data based on clinical history and medical images, doctors can easily draw on the experience of previous cases to apply specific medical interventions, which means that it can improve the accuracy of diagnosis [1].

With relatively little patient data recorded, the quality of these models varies. This paper aims to explore the impact of various advanced machine learning models on the construction of disease data systems, with a primary focus on accuracy. In this study, taking diabetes as an example, the logistic regression method was used to select the most significant feature

Corresponding author: zhy22zachary@gmail.com

vectors affecting the progression of diabetes. Modeling was performed using four key feature vectors, examining the application of deep learning models, neural network models, and integrated models in predicting the accuracy of diabetes outcomes, and ultimately selecting the best model.

2 Methods

The data used in this experiment comes from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dataset on Kaggle. All patients in the dataset are women aged 21 and above [2].

In this experiment, NLP (Natural Language Processing) techniques were first used to clean the data, ensuring that irrelevant data and noise in the raw text do not affect model performance.

The data was structured by organizing the information into fixed fields (columns) such as age, BMI, and family history of diabetes. This ensured consistency and completeness in the data used by the model.

Logistic regression is used to select four different feature vectors from the data set. Logistic regression is usually easy to implement and intuitively understands the importance of features by estimating the size of the coefficients [3]. The higher the coefficient, the greater the impact on the target vector (whether the patient has diabetes or not). This method is particularly well suited for estimating probabilities in binary classification problems because it uses a logical function to map the output of a linear equation to a range between 0 and 1.

Different machine learning models were evaluated and compared to select the one with the highest predictive accuracy. Models with lower accuracy were optimized to achieve the best possible prediction results.

Visualization was used to help researchers clearly assess the accuracy of the models. For example, we evaluated how the models performed with extreme and intermediate data points.Visualization was used to help researchers clearly assess the accuracy of the models, including how they performed with extreme and intermediate data points. Through the visualization of different models, it was possible to compare which model was the most stable, which performed best when handling extreme data, and which was the most robust, adapting well to varying data requirements.

3 Experimental results

First, logistic regression was used to select four different feature vectors from the dataset [3], and the results are as figure1:

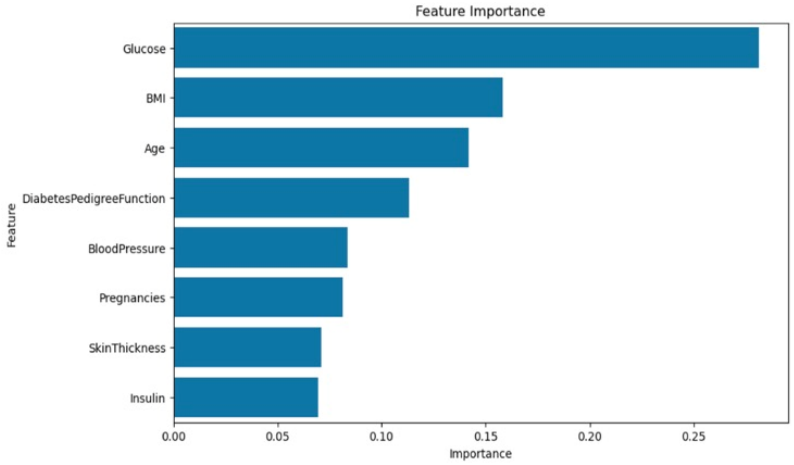

Fig. 1. Using logistic regression to analyze feature vectors. Below the figure.

(Picture credit : Original)

As shown in Fig. 1, glucose, BMI, age, and the diabetes pedigree function ranked as the top four in feature importance, indicating their significant contribution to the model's predictive performance. By selecting the most important features, the dimensionality of the model can be reduced, improving computational efficiency. Additionally, training the model using these features ensures strong predictive performance while simplifying the model.

Three of the most advanced machine learning methods were used to train the model, including deep learning, neural networks, and ensemble models (combining predictions from support vector machines, logistic regression, linear regression, and deep neural networks). The accuracy of diabetes prediction was evaluated for each model, with differences in accuracy ranging between 5% and 10%.

For the deep learning model, it typically made highly confident predictions. It shows some overfitting, especially for extreme predictions, and shows instability in samples in the middle probability range (0.4-0.6), with results skewed towards the extremes.

For deep neural network models, it provides good predictions for both high - and low-risk samples. However, the model tends to focus too much on extreme values, leading to errors when dealing with more complex samples. After adding the dropout layer, the performance of the model is significantly improved. The use of dropout helps reduce overfitting, resulting in more stable predictions when dealing with unknown data [4]. It also improves the model's predictions for moderate-risk samples, increasing the proportion of samples with probability distributions between 0.2 and 0.8. Although forecasts have become more balanced and reasonable, this increased distribution can also lead to less accurate forecasts for some samples.

For integrated models (combining predictions from support vector machines, logistic regression, linear regression, and deep neural networks), the integration of the strengths of different models reduces the limitations of a single model. This led to more stable performance, particularly when handling complex samples. The ensemble model showed balanced performance across different probability ranges, avoiding extreme predictions and improving the model's generalization ability. Although the ensemble model was more robust, its performance on specific samples may not have been as strong as that of an optimized individual model.

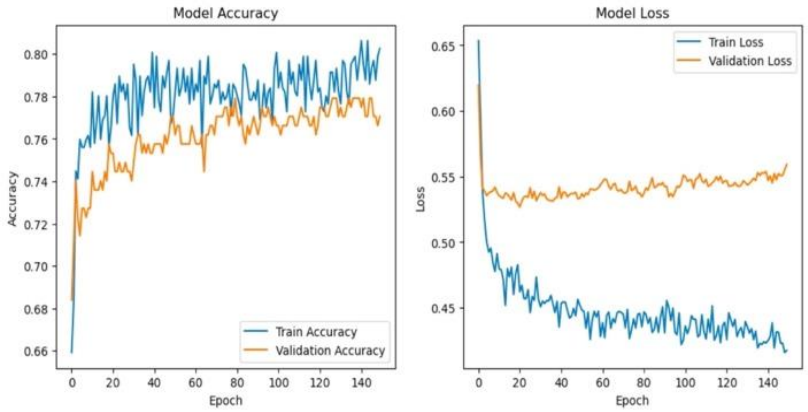

Fig. 2. Prediction Results of the Ensemble Model. Below the figure.

(Picture credit : Original)

As shown in Fig.2, the ensemble model combines the advantages of multiple models, offering greater stability and stronger predictive capabilities, making it better suited for handling complex problems. Through comparison with other models, the ensemble model demonstrated the highest prediction accuracy.

As shown in the graph, the training accuracy steadily increased with each iteration, reaching close to 80%, while the validation accuracy fluctuated around 75%. Although there were some fluctuations in the training accuracy during the iterations, the overall trend was upward. The validation accuracy, on the other hand, remained relatively stable but showed noticeable variations. This may indicate that while the model fits the training data well, its performance on the validation set is not as strong. The significantly higher accuracy on the training set compared to the validation set suggests a degree of overfitting, although it is not severe.The test set accuracy, as indicated at the top of the graph, is 0.77 (77%), demonstrating that the model's performance on the test set is consistent with the validation set, further confirming that the model's generalization ability is decent, but there is still room for improvement.

The loss function curve shows that the model quickly reduced the training loss within the first 20 epochs, indicating a fast early convergence. However, as the number of iterations increased, the reduction in the loss function slowed down, and the validation loss showed little change.

Due to the limited amount of data, even with the ensemble model, which had the best predictive accuracy, the prediction accuracy was only around 77.7%. Factors such as the lack of important feature vectors contributed to this limitation. According to some studies [5], diabetes is closely linked to dietary habits, sleep patterns, and daily sugar intake. However, current EHR systems lack quantifiable data in these areas, which significantly impacts the results. Compared to existing literature, the results of this experiment are largely consistent, further proving that the model requires more data to improve prediction accuracy.

4 Discussion

4.1 Data security and openness

Although machine learning offers convenience in disease prediction, allowing non-medical individuals to easily quantify their health data for predictions, the collection of such data

presents significant challenges. One major issue is ensuring the security and openness of the data [6]. To minimize potential harm from prediction results, system developers must ensure the safety and reliability of these systems. For instance, during the training process, biases at all levels must be eliminated [7].

4.2 Future directions

To predict outcomes for unknown data and effectively train models, deep learning typically requires large datasets. When available datasets are limited, this becomes a significant challenge [8]. For this experiment, the datasets used for training were sourced from GitHub and Kaggle (an online platform for data science and machine learning widely used for data competitions, learning, and collaboration). However, there are very few datasets for common diseases (such as diabetes, hypertension, asthma), resulting in lower prediction accuracy for the models.

4.3 Proposed solutions

One potential solution is the widespread adoption of IoMT (Medical Internet of Things) devices that can be used to collect and transmit patient-related data [9]. These devices can monitor a variety of physiological data (such as heart rate, blood pressure, blood sugar levels, and body temperature). Ideally, by integrating with electronic medical record systems, they can continuously monitor the user's vital signs and provide recommendations for improving health to prevent disease from occurring. In addition, these devices can intervene in emergency situations, such as alerting a fall or cardiac arrest [10].

5 Conclusion

The main goal of this study was to improve the accuracy of disease prediction using electronic health record (EHR) data by selecting suitable algorithms. The experimental results show that the integrated model has high stability and strong prediction ability in disease prediction. For complex problems, integrated models often provide more reliable predictions, making them a reliable choice.

This model has strong performance, but the data sources are too limited, and there are no better alternatives in similar predictive models. Future directions include using machine learning from patient data from different national Centers for Disease Control to predict a patient's condition in advance and suggest appropriate treatment.

The experimental approach faces certain challenges, such as algorithmic bias and ethical review, but advances in electronic medical record systems are undeniable. They play a vital role in increasing global life expectancy and facilitating the efficient allocation of healthcare resources.

References

1. S.C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digital Medicine, **3**, 136 (2020)
2. Krishna, A. Yuva; Kiran, K. Ravi; Sai, N. Raghavendra; Sharma, Aditi; Praveen, S. Phani; Pandey, Jitendra, Ant Colony Optimized XGBoost for Early Diabetes

Detection: A Hybrid Approach in Machine Learning. *Journal of Intelligent Systems & Internet of Things*, **10**, 76 (2023)

3. B. Vrigazova, Novel approach to choosing principal components number in logistic regression. ENTRENOVA - ENTerprise REsearch InNOVAtion, **7**, 12 (2021)
4. S. T. Ahmed, K. Danouchi, C. Münch, G. Prenat, L. Anghel, and M. B. Tahoori, SpinDrop: Dropout-Based Bayesian Binary Neural Networks with Spintronic Implementation. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, **13**, 150–164 (2023)
5. M. A. Ceran, M. G. Keser, M. Bektas, N. Unusan, and B. S. Eklioglu, The impact of dietary habits on sleep deprivation and glucose control in school-aged children with type 1 diabetes: A cross-sectional study. Children, **11**, 779 (2024)
6. M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, **3**, 73 (2022)
7. M. M. Taye, Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*, **12**, 91 (2023)
8. M. D. McCradden, S. Joshi, M. Mazwi, and J. A. Anderson, Ethical limitations of algorithmic fairness solutions in health care machine learning. The Lancet Digital Health, **2**, E221–E223 (2020)
9. A.V.L.N. Sujith, G.S. Sajja, V. Mahalakshmi, S. Nuhmani, and B. Prasanalakshmi, Systematic review of smart health monitoring using deep learning and Artificial intelligence. Neuroinformatics, **2021**, 100028 (2021)
10. K. Naik, R. K. Goyal, L. Foschini, C. W. Chak, C. Thielscher, H. Zhu, J. Lu, J. Lehar, M. A. Pacanowski, N. Terranova, N. Mehta, and N. Korsbo, Current status and future directions: The application of artificial intelligence/machine learning for precision medicine. Clinical Pharmacology & Therapeutics(2023)