ISSN 2959-6157

Application of Machine Learning in the Field of Diabetes

Churong Wei[1, *]

[1]Shanghai International High School of BANZ, Shanghai, China
*Corresponding author: weichurong@outlook.com

Abstract:

Diabetes, a prevalent chronic disease worldwide, poses significant challenges to healthcare systems, with its cases steadily increasing. Machine learning has emerged as a promising tool in diabetes research, offering opportunities for prediction and diagnosis through analysis of vast clinical datasets. This paper systematically reviews and analyzes the latest research progress in ML applications for diabetes, encompassing risk prediction and clinical diagnosis. Various ML techniques, including neural networks, supervised learning, semi-supervised learning, deep learning, and ensemble learning, are explored in their application to diabetes research. While ML has shown promise in improving diagnostic accuracy and efficiency, several challenges remain, such as data quality, model interpretability, and computational requirements. Recommendations for future research focus on addressing these challenges to further advance ML's effectiveness in diabetes management. Through this review, we aim to provide valuable insights for researchers and clinicians, promoting the continued development and application of ML technology in diabetes care, improving efficiency and accuracy.

Keywords: Machine Learning; Diabetes; Neural Network; Deep Learning.

1. Introduction

Diabetes represents a formidable global health challenge, with its prevalence on an upward trajectory that threatens to overwhelm healthcare systems worldwide. As of 2020, the World Health Organization reported a staggering 415 million individuals grappling with this chronic condition, a number anticipated to burgeon to 625 million by 2030. Beyond the immense toll it exacts on individual well-being, diabetes imposes substantial burdens on healthcare infrastructures, necessitating innovative approaches for effective management and treatment.

In this context, the advent of machine learning stands as a beacon of hope, offering unprecedented opportunities for revolutionizing diabetes care. Machine learning algorithms, fueled by vast troves of clinical data, hold the potential to unearth intricate patterns obscured to human observation alone. By harnessing these insights, healthcare professionals can enhance the accuracy and efficiency of diabetes prediction, diagnosis, and treatment while concurrently mitigating costs.

The primary objective of this paper is to meticulously distill and evaluate the latest strides in machine learning research within the realm of diabetes. By surveying a breadth of literature, encompassing but not limited to predictive modeling and diagnostic support, we endeavor to furnish a comprehensive understanding of the landscape. Furthermore, we aim to delineate the diverse array of machine learning algorithms employed in diabetes research, delineating their respective strengths and limitations.

In scrutinizing the manifold applications of machine learning, we aspire to furnish valuable insights that resonate with researchers and clinical practitioners alike. By elucidating the current state-of-the-art methodologies and their implications for diabetes management, this paper endeavors to catalyze the adoption and evolution of machine learning technologies within the field.

Looking ahead, we anticipate a burgeoning wave of innovation, propelled by ongoing advancements in machine learning techniques and the proliferation of healthcare data. However, we also recognize the attendant challenges, including the need for robust validation frameworks, the ethical implications of data utilization, and the imperative of ensuring equitable access to technology-driven solutions.

According to variety of researches, among all attributes which should be considered, the importance of the attributes from high to low is: blood glucose value, BMI, hereditary index, age, blood pressure, number of pregnancies, insulin content, and sebum thickness. From the point of view of preventing diabetes, more attention should be paid to whether the blood glucose value is in a stable range, as the risk of diabetes is higher as we age, weight control is particularly important, due to genetic factors, we must not have a psychological burden, to maintain a good spirit, an active lifestyle and healthy habits [1].

Ultimately, it is our fervent aspiration that this paper serves as a beacon of guidance, galvanizing concerted efforts toward harnessing the transformative potential of machine learning in the relentless battle against diabetes. Through collaborative endeavors and steadfast dedication, we envisage a future wherein machine learning not only augments clinical decision-making but also fosters profound improvements in patient outcomes and quality of life.

2. Application of Machine Learning in the Field of Diabetes

2.1 Current Status Description

With the rapid advancement of computer technology and machine learning in recent years, research on the application of machine learning in the field of diabetes has also made significant progress, covering multiple aspects such as disease prediction and diagnosis. Specifically, it can be divided into the following categories:

1. Risk prediction of diabetes: By analyzing large samples of data, predicting the risk of individuals developing diabetes can be achieved by models analyzing multiple factors such as gender, age, family history, and lifestyle, to help doctors better identify high-risk groups and take preventive measures.

2. Clinical diagnosis of diabetes: Machine learning models can assist doctors in diagnosing patients by analyzing large amounts of clinical data, judging whether they meet the criteria for diabetes based on various indicators, and providing reference judgments.

Overall, machine learning models have made significant progress in the field of diabetes, and it is expected that they will provide significant support to this field in the future, improving diagnostic accuracy, speeding up diagnosis, and reducing medical costs.

2.2 Existing Research Analysis

2.2.1 Neural Networks

Neural networks are computational models that mimic the human nervous system, consisting of a large number of artificial neuron nodes. They process information through connections and weights between them, usually divided into multiple layers, including input layer, hidden layer, and output layer. Each neuron receives inputs from the previous layer's neurons and produces outputs, which are then passed to the next layer's neurons. Neural networks adjust the weights between neurons through training to efficiently and accurately process input data.

In recent years, with the advancement of computer performance and the prevalence of big data, neural network technology has made significant progress, gradually becoming a core technology in many fields. Neural networks have the advantage of strong adaptability, being able to learn complex patterns and relationships from data, suitable for more complex problems. However, when it comes to more complex task analysis, neural network models require a large amount of time and computational resources. Moreover, the internal decision-making process of neural network models is difficult to explain, so their credibility in disease diagnosis still needs improvement.

Jiang Aijuan et al. used the neural network algorithm to construct a prediction model for the prediction of diabetic symmetrical polyneuropathy, collected complete cases of 4,107 hospitalized diabetic patients from Anhui University of Traditional Chinese Medicine's First Affiliated Hospital from 2017 to 2022, collected multiple indicators, and used neural networks to establish a prediction model, ranking the weights of variable characteristics, analyzing the potential factors of lesions, and experimental results prove that it has high accuracy in early prediction of lesions and has high clinical value [2].

2.2.2 Supervised Learning

Supervised learning is a machine learning paradigm that learns the mapping relationship between input and output from labeled training data. In supervised learning, algorithms build models by learning the relationship between input features and corresponding output labels, thereby being able to predict or classify new unlabeled data after training. Supervised learning, as it uses labeled data for training, can make more accurate predictions and classifications. Moreover, supervised learning can choose different models according to task requirements, being more flexible. Some models have good interpretability, able to explain the final decisions, suitable for the field of diabetes. However, supervised learning is not entirely autonomous and requires manual labeling and preliminary classification of data, which poses a significant challenge, and has high requirements for computation and storage.

2.2.3 Semi-Supervised Learning

Semi-supervised learning is a learning paradigm between supervised learning and unsupervised learning. In semi-supervised learning, the training dataset usually contains a large amount of unlabeled data and a small amount of labeled data. Unlike supervised learning, semi-supervised learning utilizes information from unlabeled data to enhance the performance and generalization ability of models. Compared to supervised learning methods, semi-supervised learning can use more easily and widely obtained unlabeled data for training, making data acquisition simpler and closer to real-world usage scenarios, as unlabeled data are often obtained in actual usage and training scenarios. However, the quality of unlabeled data varies, which may affect the quality of the model, and when using labeled and unlabeled data simultaneously, if there is a large difference in data distribution, the model may tend to be biased towards labeled data, leading to a decrease in generalization ability.

Since there are more than 50,000 unlabeled data in the Kaggle competition, how to build a semi-supervised learning model and make full use of unlabeled data is also an important direction for future research [3].

2.2.4 Deep Learning

Deep learning is a learning method in machine learning based on artificial neural network models, aiming to abstract and learn data at a high level through multi-level nonlinear transformations. Deep learning relies on deep neural networks, typically consisting of multiple hidden layers, each hidden layer composed of a large number of neurons. These neurons transmit and process information through weighted connections, enabling deep learning models to learn complex features and patterns.

Deep learning models have strong adaptability, being able to autonomously learn features and patterns in input data without the need for humans to design extractors. However, labeled data are still required, and acquiring these data is time-consuming. For high performance of deep learning models, many parameters need to be adjusted, requiring a lot of time for trial and error.

Pang Hao et al. By constructing a two-level deep convolutional neural network, the feature extraction, feature combination and result classification of the original photos are completed, and the screening results are finally obtained [3].

2.2.5 Ensemble Learning

Ensemble learning is an emerging machine learning method that combines multiple weak learners to build a strong learner, improving the accuracy and robustness of the model. Many weak models in the ensemble can reduce the bias of the final result by combining the results of multiple models, thereby improving accuracy. This decision logic is relatively easy to explain, and multiple learners can be flexibly integrated to adapt to different tasks. However, the practice of introducing multiple learners also has disadvantages. It increases the complexity of the model, and ensemble learning is sensitive to the quality and distribution of training data. If the training data contain noise or have significant class imbalance issues, it may affect the performance of the ensemble model.

Jiaqi Wang et al. used the ensemble learning method to construct a prediction model for the prediction of diabetes risk, using data form the UCI machine learning database, 520 samples, including 200 healthy and 320 affected, operating with random forest, GBDT, XGBoost algorithm. Five indicators were used to evaluate the effect of the model, including accuracy, precision, recall, F1 score and AUC, and the importance of variables was ranked based on the best performance model. These diabetes risk prediction model has shown good performance, which can more accurately identify early high-risk patients than single classifiers, which is helpful for clinicians to make more accurate medical decisions [4].

3. Recommendations for Current and Future Research

This paper reviews and analyzes the research progress of machine learning in the field of diabetes, exploring specific applications such as disease prediction and diagnosis through a review of existing research literature. It is found that machine learning has made significant achievements in the field of diabetes, helping doctors diagnose diseases more accurately and predict patient risks. However, we also note some shortcomings and limitations in current research, such as insufficient data quality, poor model interpretability, high computational requirements, and difficulties in clinical practical application. Therefore, it is hoped that future research can focus on addressing these issues, further improving the effectiveness and operability of machine learning in the field of diabetes. We believe that with the continuous development of technology and in-depth research, this technology will further advance the application of machine learning in the field of diabetes.

The use of artificial intelligence (AI) for medical imaging-assisted diagnosis is a rapidly evolving field that has recently yielded several impressive results: in late 2016, Google’s DR detection algorithms were on par with ophthalmologists’ performance in JAMA, a top medical journal [5].

In early 2017, a paper published by Stanford University in Nature showed that the accuracy and sensitivity of the skin cancer detection model constructed by Stanford University can reach or even exceed the level of professional physicians [6].

With the rapid development of deep learning technology and the continuous opening of medical imaging data, artificial intelligence is gradually maturing in early disease screening and doctor-assisted diagnosis.

4. Conclusion

The paper provides a comprehensive examination of the application of machine learning (ML) in the field of diabetes research. Highlighting the escalating global prevalence of diabetes and its impact on healthcare systems, the study underscores ML as a transformative technology offering solutions for prediction and diagnosis. Through an extensive review of current research, the paper delineates the utilization of various ML techniques, including neural networks, supervised and semi-supervised learning, deep learning, and ensemble learning, in addressing different facets of diabetes management. While acknowledging the progress made in enhancing diagnostic accuracy and efficiency, the paper also identifies persistent challenges such as data quality, model interpretability, and computational constraints. To overcome these hurdles, the study offers recommendations for future research, aiming to propel the practical application of ML in diabetes care. By consolidating existing knowledge and delineating avenues for advancement, the paper serves as a valuable resource for researchers and clinicians alike, fostering the continued integration of ML technology in diabetes management strategies.

References

[1] Liu Ruiyi, Qu Yimin, Liu Xuan, Jiang Yu. Application of ensemble learning and decision tree in prospective risk assessment of type 2 diabetes mellitus. Chinese Journal of Chronic Disease Prevention and Control, 2023, 31(04): 278-283.

[2] Jiang Aijuan, Wang Lujie, Li Jiagao, et al. Construction and Validation of a Prediction Model for Diabetic Distal Symmetric Polyneuropathy Based on Neural Networks. China XunZheng Medical Magzine, 2024, 24(03): 265-271.

[3] Pang Hao, Wang Cong. Deep learning model for diabetic retinopathy detection. Journal of Software, 2017, 28(11): 3018−3029.

[4] Wang Qiqi, Dai Jiajia, Cui Xiongwei. Diabetes risk prediction based on ensemble learning model. Software Guide, 2022, 21(04): 62-66.

[5] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 2016, 316(22): 2402-2410.

[6] Andre E, Brett K, et al. Dermatologist-Level classification of skin cancer with deep neural networks. Nature, 2017,542(7639): 115−118.