Enhancing Diabetes Prediction Through Hybrid Deep Learning: Analysis of ML and DL Techniques

Yuhao Wang

Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Guangdong, China

r130026144@mail.uic.edu.cn

Abstract. This interview investigates how machine learning (ML) and deep learning (DL) techniques are implemented in predicting diabetes, with the goal of identifying the most effective methods for early diagnosis. As diabetes prevalence continues to rise, developing accurate prediction models is essential for enabling timely interventions and reducing related health risks. The research compares traditional ML methods, including Support Vector Machines (SVM), Decision Trees, and Naive Bayes, against advanced DL models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and a hybrid CNN+LSTM model. The hybrid approach leverages the effectiveness of both CNNs and LSTMs, effectively analyzing data patterns in temporal and spatial respective. Extensive experimental results reveal that the CNN+LSTM model reaches to a dominant accuracy of 98%, significantly outperforming the other evaluated methods. This finding highlights the potential of hybrid deep learning approaches in improving the accuracy of diabetes prediction. The study concludes by discussing the implications of these results and suggests future research directions, including the exploration of more diverse datasets and the application of these models in clinical settings to enhance their generalizability and practical utility.

Keywords: Diabetes Prediction, Machine Learning, Deep Learning, Support Vector Machines.

1. Introduction

Diabetes prediction involves the use of various methodologies to forecast the likelihood of developing diabetes before its occurrence. This prediction process is crucial because it helps early intervention, potentially preventing the progression of diabetes and reduces associated complications. The diabetes prediction is significant because its ability to identify individuals at high risk, allowing for timely lifestyle adjustments and medical treatments that can substantially improve quality of life and save amounts of healthcare costs.

The aim of this review is to explore the most advanced advancements in diabetes prediction, which especially focus on the application of machine learning (ML) and deep learning (DL) techniques. Specifically, this study will begin by reviewing and summarizing the foundational concepts and historical development of diabetes prediction technologies. Following that, the study will provide an in-depth analysis of the key technologies, including decision trees, support vector machines (SVM), and naive Bayes, which are most frequently used, as well as more recent deep learning models like Convolutional neural networks (CNN) and long short-term memory networks (LSTM). Additionally, the study will examine each technologies performance in practical applications, the challenges they face, and their future research directions. Compared to other works, the core contribution of this research lies in innovatively studying current ML and DL schemes and comparing and analyzing their applications and insights in healthcare. This study provides an in-depth analysis of the core concepts and future development of this research.

This paper is managed into four chapters: First, it introduces the popular datasets that current research in diabetes prediction. Then it provides a detailed analysis to the practical usage of traditional ML Method in diabetes prediction. Third thesis discuss how DL method implemented in. Finally, thesis end this paper by a summary with findings and a prediction on future research directions.

2. Literature Review

Currently, a variety of methods is implemented in diabetes prediction, like traditional ML and more advanced DL. Traditional ML methods, such as decision trees, SVM, and naive Bayes, have been utilized extensively. These techniques often involve constructing models relied on databanks of the past, which are then used to classify or predict the likelihood of diabetes. For instance, decision trees work by splitting data into branches based on feature values, while SVM aim to find the optimal hyperplane that separates different classes. Naive Bayes, whose idea comes from Bayes' theorem, presume the independence between each feature, simplifying the modelling process and enabling efficient computations. In recent years, DL have become a significant topic in diabetes prediction because they are able to catch complex patterns and relationships between figures in large datasets. CNN and its developed Method---LSTM, are among the most advanced DL methods applied in diabetes prediction. CNN plays important roles in processing spatial hierarchies in data, making them suitable for image-based features, while LSTM are designed to handle temporal sequences, which is beneficial for analyzing longitudinal health data.

These methods have achieved notable results in predicting diabetes, with accuracy rates varying from 68% to 95.1% depending on the algorithms and data used. Aakansha Rathoreet’s team achieved remarkable success with the SVM method, attaining an impressive accuracy of approximately 82% in their diabetes prediction model [1]. Such a high accuracy highlights the effectiveness of SVM in handling the complexities and nuances associated with medical data. What is more, combination of CNN and LSTM networks has yielded a record-high accuracy of 95.1% [2]. This significant leap demonstrates the capability of leveraging different DL architectures in tandem.

Nevertheless, difficulties persist. For example, the data availability is in limit, and the need for effective model deployment, and the risk of overfitting. Due to the different lifestyles caused by varied regions, generally, researchers take 80% of their time to clean and modify data for model training. Some researchers use one or few parameters for classification, which will critically decrease the accuracy of models. Future advancements may focus on improving data quality, exploring novel algorithms, and addressing deployment issues. Additionally, emerging approaches like computer vision and iris pattern analysis are being investigated, offering potential approaches for expanding predictive capabilities. For instance, A facial image analysis system, called computer-assisted non-invasive DM detection system, quickly assesses skin health by examining specific areas. It extracts features using a local binary pattern and performs classification using k-nearest neighbor (KNN) clustering and a SVM Real-time results are displayed through connected software [3].

3. Methodology

3.1. Transitional ML-based methods

In diabetes prediction, traditional ML methods have been widely applied because they can effectively manage structured data and deliver reliable predictions. Among those algorithms, Decision Tree, SVM and Naïve Bayes are believed that the most practical ones in numbers of experiments.

Decision Trees are particularly favored for their simplicity and interpretability. They work by splitting the dataset into branches based on specific feature values, creating a flowchart-like model that is easy to understand. Permana's team use decision tree to find the interrelationship between each figure they chosen and diabetes' occurrence [4]. This method is powerful in classification tasks, including distinguishing between patients who is positive in diabetes or not. The pipeline is shown in the Figure 1.

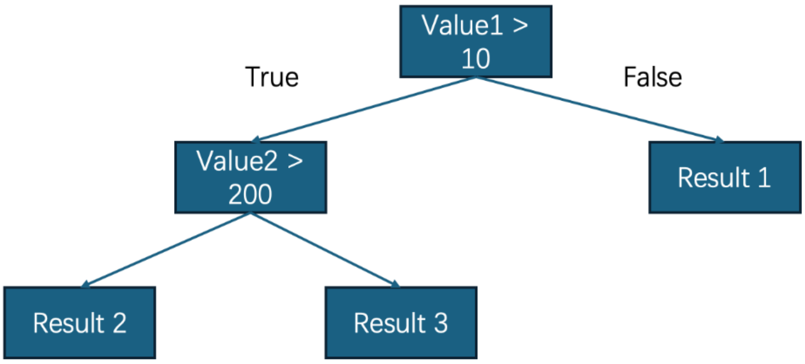

Figure 1. Basic example of Decision Tree.

SVM are another robust tool in diabetes prediction. SVM operate by searching the optimal hyperplane that best distinguishes the different classes (e.g., diabetic vs. non-diabetic) [5]. This method performs well in dealing with high-dimensional data since it is designed for maximizing the margin between data points of different classes, thereby improving prediction accuracy. Naive Bayes, based on Bayes' Theorem, set a hypothesis that the presence of a particular feature in a class is independent of the presence of any other feature [6]. Despite its simplicity, Naive Bayes performs well in many real-world applications, including diabetes prediction, because it is computationally efficient and shows great performance with large datasets. It is particularly useful when the assumption of feature independence holds, allowing for quick predictions with minimal computational cost. These traditional ML methods have been proven effective in predicting diabetes by leveraging patient data. Studies show that these methods, especially when combined with feature selection techniques, can achieve high accuracy in predicting diabetes, making them valuable tools in healthcare.

3.2. DL-based methods

Deep learning methods have been increasingly utilized in diabetes prediction nearly as they have strong capability to uncover complex patterns within huge datasets [7]. Among these methods, CNN and LSTM have shown significant promise and the combination application of those two methods wildly used as well. CNN are particularly effective in handling spatial hierarchies, making them ideal for processing image-based features. Figure 2 shows the basic conception of CNN. In the context of diabetes prediction, CNN can be used to analyze medical imaging data or other spatially organized data to identify patterns indicative of diabetes. For example, Jaloli's team develops a high-accuracy 4D CNN model for early detection of Type2 Diabetes in Oman [8].

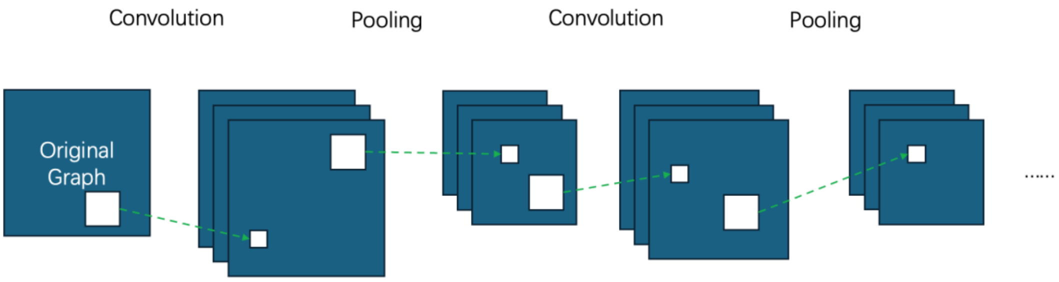

Figure 2. CNN basic concept.

LSTM are designed to process sequential data, making them suitable for analyzing time-series data such as blood glucose levels over time. Figure 3 illustrate the concept of LSTM. LSTM can capture long-term dependencies and trends in the data, which are crucial for accurately predicting future glucose levels or the onset of diabetes based on historical data [9]. For instance, LSTM have been employed to predict blood glucose levels in patients with Type 1 diabetes, providing real-time forecasts that can help in managing the condition more effectively [10].

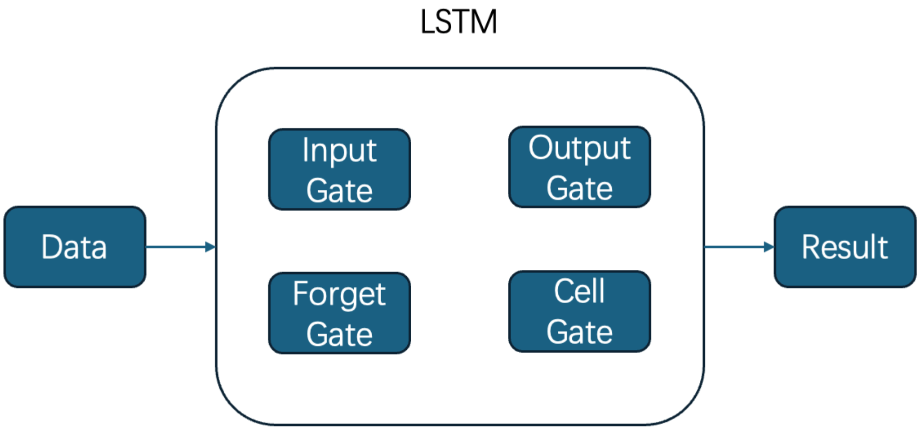

Figure 3. LSTM structure.

Moreover, hybrid models that combine CNN and LSTM have been developed to leverage the strengths of both architectures [8]. These models can simultaneously process spatial and temporal data, resulting in more accurate and comprehensive diabetes prediction. Such a model might first use a CNN to extract features from medical images and then feed these features into an LSTM to predict special disease by both information in images and temporal trends. The integration of CNN and LSTM in diabetes prediction has yielded encouraging results, with some studies reporting that these methods performs better than conventional methods. However, challenges remain. For instance, the large datasets are still in need to feed these complex models for model’s accuracy. And the lack of the computational resources is extremely in high demand if need to deploy them in real-time applications.

4. Result and Discussion

4.1. Effectiveness of each ML/DL methods in diabetes prediction

The results in the table 1 indicate that the CNN+LSTM hybrid model outperforms all other methods, whose accuracy reaches to 98%. The high accuracy suggests that the combination of CNN and LSTM models is highly effective for diabetes prediction, likely because it leverages both spatial (from CNN) and temporal (from LSTM) features in the data. The individual LSTM model also performs well with an accuracy of 91%, demonstrating its ability to handle time-series data, which may be crucial in predicting diabetes based on patterns over time. On the other hand, the CNN model, with an accuracy of 89.47%, shows strong performance as well, highlighting the effectiveness of convolutional operations in feature extraction, even without the temporal dynamics captured by LSTM. However, its performance slightly lags behind LSTM, indicating that while CNN is powerful, incorporating temporal analysis through LSTM adds additional predictive power.

Table 1. Accuracy of each method in diabetes' prediction.

Method	Accuracy
SVM	90%
Decision Tree	85.5%
Naive Bayes	77%
CNN	89.47%
LSTM	91%
CNN+LSTM	98%

This study presents a comprehensive evaluation of different ML models, both traditional and DL-based, in the context of diabetes prediction. The high accuracy of 90% achieved by SVM underscores its strong capability in handling classification tasks, particularly when there are clear boundaries between classes. This suggests that diabetes prediction data has distinguishable patterns that SVM can effectively leverage. The moderate performance of Decision Trees at 85.5% reflects their simplicity and interpretability, though it also highlights their limitation in capturing complex patterns. Naive Bayes, with the lowest accuracy of 77%, reveals the drawback of assuming feature independence in datasets where interdependencies are crucial. The quantitative and qualitative analysis in this research emphasizes the need for more sophisticated models, as traditional methods struggle with the complexities of the data, reinforcing the importance of model selection in achieving higher predictive accuracy.

4.2. Discussion

The overall trend in the results indicates that deep learning methods, particularly those combining different architectures like CNN and LSTM, significantly outperform traditional ML models in diabetes prediction. The superior accuracy of CNN+LSTM suggests that capturing both spatial and temporal patterns is crucial for accurate prediction. The discovery emphases the significant of model selection in diabetes predictions and highlights the potential of hybrid models in achieving higher accuracy. The lower performance of traditional models like Naive Bayes points to complexity of diabetes prediction, where simple models may not suffice. These results suggest that research and applications in the future should focus on advanced deep learning methods, particularly hybrid models, to improve prediction accuracy in diabetes and potentially other medical conditions. This study underscores the innovation and significance of hybrid deep learning models, specifically the combination of CNN and LSTM, in diabetes prediction. By demonstrating that these models capture both spatial and temporal patterns more effectively than traditional machine learning models, the research highlights the critical role of model architecture in improving predictive accuracy. The findings advocate for the use of advanced deep learning approaches in medical prediction tasks, suggesting that hybrid models have the potential to transform not only diabetes prediction but also broader applications in healthcare analytics.

5. Conclusion

This study addresses the critical issue of diabetes prediction, focusing on the effectiveness of a variety of ML and DL algorithm in predicting the likelihood of diabetes onset. The primary objective is to identify the most accurate techniques for early detection, which can great effect on timely intervention and potentially slow diabetes' progression. To this end, a hybrid model who is the combination of CNN and LSTM is proposed to analyze the intricate features of huge datasets. The methodology involves leveraging CNN to capture spatial hierarchies in the data, followed by the use of LSTM to analyze sequential dependencies, thereby enhancing the overall predictive accuracy for diabetes. Extensive experiments were conducted to test the effectiveness of this hybrid algorithm. The results shows that the CNN+LSTM model is in significantly higher performance than that of traditional ML and other DL methods, reaching an accuracy of 98%. The finding means that the potential of integrating different deep learning architectures to improve prediction accuracy in complex medical conditions such as diabetes. Looking ahead, future research will aim to expand the scope by incorporating more diverse datasets and exploring the clinical application of these models in real-world environments. The focus will be on assessing the robustness and generalizability of the hybrid model across various populations and medical conditions, with the goal of further improving the accuracy and applicability of diabetes prediction models.

References

[1] Sharma T and Shah M 2021 A comprehensive review of machine learning techniques on diabetes detection Journal of Medical Artificial Intelligence vol 5 no 2 pp 123-138

[2] Larabi-Marie-Sainte S Aburahmah L Almohaini R and Saba T 2019 Current Techniques for Diabetes Prediction: Review and Case Study Computer Science Department vol 9 no 21 p 4604

[3] Rajeswari M and Prabhu P 2019 A Review of Diabetic Prediction Using Machine Learning International Journal of Engineering and Techniques vol 5 no 4 pp 2395-1303

[4] Permana B A C Ahmad R Bahtiar H Sudianto A and Gunawan I 2021 Classification of diabetes disease using decision tree algorithm Journal of Physics: Conference Series vol 1869 no 1 p 012082

[5] Butt U M Letchmunan S Ali M et al. 2021 Machine learning based diabetes classification and prediction for healthcare applications Journal of healthcare engineering vol 2021 p 9930985

[6] Priya K L Kypa M S C R Reddy M M S and Reddy G R M 2020 A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier International Conference on Trends in Electronics and Informatics pp 1-6

[7] Mustafa M 2024 Diabetes Prediction Dataset Retrieved on 2024 Retrieved from: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[8] Al-Sadi K and Balachandran W 2023 Revolutionizing Early Disease Detection: A High-Accuracy 4D CNN Model for Type 2 Diabetes Screening in Oman Bioengineering vol 10 no 12 p 1420

[9] Chowdary P B K Kumar R U 2021 An effective approach for detecting diabetes using deep learning techniques based on convolutional LSTM networks International Journal of Advanced Computer Science and Applications vol 12 no 4 pp 519-525

[10] Jaloli M Cescon M 2023 Long-term prediction of blood glucose levels in type 1 diabetes using a cnn-lstm-based deep neural network Journal of diabetes science and technology vol 17 no 6 pp 1590-1601