Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/109/20241437

Advances and Challenges in Machine Learning for Diabetes Prediction: A Comprehensive Review

Yueheng Ding

School of Science and Engineering, University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo, 315100, China

scyyd9@nottingham.edu.cn

Abstract. Diabetes mellitus is a prevalent and severe metabolic disorder disease that poses significant health risks globally, leading to substantial healthcare burdens. Recent days, advancements in artificial intelligence (AI) have markedly enhanced the accuracy and efficiency of diabetes outcome predicted by machine learning (ML), offering a promising approach for early intervention and treatment. This paper evaluates several advanced ML models, including Random Forest (RF), Support Vector Machine (SVM), and Neural Networks techniques based on neural networks. Each model's strengths and limitations are discussed, highlighting the improvements in predictive performance and diagnostic precision. Despite these advancements, the field faces ongoing challenges related to ethical considerations and data scale, which impact ML application in healthcare from both technical and moral aspects. Future efforts should focus on these challenges by promoting data sharing and integration while safeguarding privacy. Through these endeavors, we aim to advance the field of diabetes prediction and improve patient care.

Keywords: Machine Learning, Diabetes, AI Medicine, AI, Data Scale.

1. Introduction

Nowadays, diabetes has been one of the most malignant and common chronic diseases of our world. By the IDF Diabetes Atlas, in 2017, 4.0 million people were estimated to have died from diabetes and its complications. By 2019, this number had risen to 4.2 million. In addition, according to the statistics of the past two years, about half (46.2%) of these deaths were over 60 years old. Premature mortality, diabetes-related impairment, and absence from work and school have a detrimental economic impact on nations, even in 2017 alone. With an estimated cost of USD 1.31 trillion, these indirect expenses make up more than one-third of all costs [1, 2]. In response to this phenomenon, some medical institutions have begun to use artificial intelligence to assist in the formulation of specific and detailed detection and prevention strategies to effectively manage and reduce the risks of different patient groups. In instance, in gestational diabetes mellitus (GDM) field, multivariate or machine learning (ML)-based approaches have been proposed to predict result with superior accuracy. In the field of type 2 diabetes, machine learning is used to predict the maximum exercise ability of these diabetics and infer the probability of cardiovascular disease in the future with the proportion of their body's fat [3, 4]. These advancements leverage the vast of data of diabetics to train the artificial intelligence models to develop the more accurate results that can assist clinicians in tailoring personalized treatment plans.

DOI: 10.54254/2755-2721/109/20241437

Every aspect of our lives is changing due to artificial intelligence and machine learning (AI/ML), and the healthcare system is no exception. The use of AI and ML has the potential to significantly expand the scope of diabetes care, increasing its effectiveness [5]. Like random forest (RF) is an effective supervised classification method for making decisions by creating many decision tree models and combining all of the predictions from these trees to obtain a precise diabetic prediction [6]. In addition, due to the need to process a large amount of multiple medical data, using features with Support Vector Machine (SVM) is also a good choice. In this research work, the accuracy of this research may be further increased by using K-Means to eliminate noisy data and evolutionary algorithms to choose the best collection of SVM classifiers [7]. Nevertheless, the techniques are based on linear models and are not suitable for modeling intricate nonlinear data. With accuracy levels below 90%, prediction based on these conventional models occasionally fails to fulfill the performance criteria for clinical applications. It's noteworthy that several academics have dabbled with diabetes prediction research utilizing deep neural networks techniques. Deep neural networks well with complicated nonlinear data because it can automatically learn feature representations, which increases prediction accuracy [8].

This review critically examines the utilization of machine learning technologies in the prognostication of diabetes. It provides a comprehensive overview of contemporary advancements, identifies prevailing challenges, and outlines prospective research trajectories. The review delves into the principal methodologies employed, evaluates relevant datasets, and highlights notable real-world case studies, thereby emphasizing the transformative impact of machine learning techniques on enhancing the prediction and management of diabetes.

2. An Overview of Diabetes

Diabetes is a chronic illness characterized by persistently elevated blood sugar levels, leading to various health complications if not properly managed. It is primarily divided into three main categories:

Type 1 Diabetes: In this form of diabetes, the body's immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas. As a result, individuals with Type 1 diabetes are unable to produce insulin, a hormone essential for regulating blood glucose levels. Without insulin, sugar cannot enter the cells to be used for energy, leading to high blood glucose levels. Patients with Type 1 diabetes must administer insulin daily, either through injections or insulin pumps, and closely monitor their blood sugar levels to manage the disease effectively. This form of diabetes often develops in childhood or adolescence but can occur at any age.

Type 2 Diabetes: The most common form of diabetes, Type 2, occurs when the body either doesn't produce enough insulin or becomes resistant to it. In this case, the pancreas may still produce some insulin, but the body's cells are unable to use it effectively, leading to elevated blood sugar levels. Type 2 diabetes is strongly linked to lifestyle factors such as poor diet, lack of physical activity, and obesity, although genetic predisposition also plays a role. Management typically involves lifestyle modifications such as a healthy diet, regular exercise, weight loss, and medications that help improve insulin sensitivity or stimulate insulin production. In more advanced cases, insulin therapy may be required.

Gestational Diabetes: This type occurs during pregnancy when the body becomes less sensitive to insulin. While gestational diabetes usually resolves after the baby is born, it can pose risks to both the mother and the baby. Women with gestational diabetes are at a higher risk of complications during pregnancy, such as preeclampsia and high birth weight in the infant. Additionally, they are more likely to develop Type 2 diabetes later in life. Close monitoring of blood sugar levels and adopting a healthy lifestyle during pregnancy are crucial for managing gestational diabetes and reducing future health risks.

Advances in technology, particularly machine learning, are revolutionizing diabetes management. By analyzing vast amounts of health data from patients, machine learning algorithms can identify patterns and trends that may not be immediately apparent to healthcare providers. These insights can help tailor individualized treatment plans, optimize insulin dosing, and predict potential complications before they occur. Machine learning can also aid in the development of better diagnostic tools, improve disease monitoring, and enhance patient outcomes across all types of diabetes. As machine learning

DOI: 10.54254/2755-2721/109/20241437

continues to evolve, it has the potential to simplify diabetes management, reduce the burden on patients, and ultimately improve the quality of life for those living with the disease.

3. Practical Models Used for Diabetes Prediction

3.1. Random Forest

Due to the ensemble nature and capacity for handling the missing information of the RF method [9], it has become a relatively common model for predicting diabetes. For the same purpose of obtaining more accurate prediction results, scholars from different regions have trained RF models with diverse data in recent years.

RF possesses an inherent feature selection mechanism, allowing it to accommodate a substantial number of input variables without necessitating dimensionality reduction. Furthermore, the potential for overfitting can be effectively mitigated through the application of out-of-bag validation [10]. Since those features, in a study conducted by Alazwari et al. [10], this methodology was employed to predict Type 1 diabetes in children within Saudi Arabia. The results yielded an R-squared value ranging from 0.85 to 0.89, thereby indicating a high degree of accuracy for the Random Forest model.

Yau et al. [11] developed a liver cancer risk scoring system using random survival forest for variable selection and Cox regression for weight assignment. Among 1,995 liver cancer patients and 1,969 cancer-free individuals, the cancer incidence was 0.92 per 1,000 person-years with a median follow-up of 6.2 years. Key predictors included chronic hepatitis B/C, alanine aminotransferase (ALT), age, cirrhosis, and sex. Yielding a concordance index of 0.706, which represents a high level of accuracy. The system provides an efficient tool for liver cancer risk prediction in diabetes patients.

Ruby et al. [12] used a novel and complex method, namely Random Forest Fuzzy Entropy (RFFE), to predict diabetes. It uses Fuzzy Entropy random vectors to enhance Random Forest tree development. The prototype incorporates data imputation, sampling, feature selection, and various predictive techniques, including Fuzzy Entropy, Synthetic Minority Oversampling Technique (SMOTE), Convolutional Neural Network (CNN) with Stochastic Gradient Descent with Momentum (SGDM), Support Vector Machines (SVM), Classification and Regression Tree (CART), K-Nearest Neighbor (KNN), and Naïve Bayes. Utilizing the Pima Indian Diabetes dataset, the study evaluates predictions through a confusion matrix and receiver operating characteristic area under the curve (ROCAUC) metrics. Results indicate that the proposed RFFE method achieves an impressive accuracy of 98%, demonstrating its effectiveness for diabetes prediction.

3.2. Support Vector Machine

Apart from the RF model previously discussed, another popular machine learning technique in use today to forecast and support diabetes is the SVM models. These models are centered around the classification of diabetes illness from high-dimensional medical datasets [13]. SVM can therefore handle bigger data sets more effectively.

Kumari & Chitra [13] used diabetic database from the machine learning laboratory at University of California, Irvine as sample data and performed prediction using SVM model. After multiple experiments, an accuracy of about 78% was finally obtained, and the conclusion was drawn using the cross-validation accuracy that the prediction accuracy will increase with the increase of sample size. In this period, they chose Radial Basis Function (RBF) to be the kernel, and adjusting specific parameters for a particular kernel is crucial in the prediction process, so there is still a lot of room for optimization in the accuracy of SVM models.

Yilmaz et al. [14] used a new modified K-Means Algorithm for clustering-based data preparation system for the elimination of noisy and inconsistent data and used Support Vector Machines for classification. Compared to the 93% accuracy obtained by using the traditional K-Means SVM algorithm, this algorithm achieved higher accuracies of 97.87%, 98.18%, and 96.71% using data of the Statlog (Heart), the SPECT images and the Pima Indians Diabetes obtained from the UCI database, respectively.

DOI: 10.54254/2755-2721/109/20241437

Using data from Pima Indians Diabetes from the UCI repository, Santhanam, & Padmavathi [7] proposed a better approach, which involves not only using K-Means and SVM to denoise and classify the data, but also using Genetic Algorithms (GA) for feature selection. As demonstrated by the experimental results, the proposed model has an average accuracy of 98.79%, which is 2.08% higher than the Modified K-Means and SVM published in the literature above.

Edehet al. [15] compared several common machine learning algorithms using the database mentioned multiple times earlier (Pima Indians Diabetes), namely supervised learning algorithms (RF, SVM and Naïve Bayes, Decision Tree) and unsupervised learning algorithm (K-Means). The results indicate that SVM achieved the highest accuracy among several algorithms with 83.1%. However, it's still unclear if SVM works well on unstructured information [8].

These studies collectively demonstrate the evolving capabilities of SVM models in diabetes prediction. The progression from basic SVM implementations to hybrid approaches incorporating clustering, genetic algorithms, and other techniques showcases the potential for continued improvement in predictive accuracy. However, it's important to note that the variability in reported accuracies across studies may be attributed to differences in dataset characteristics, preprocessing methods, and specific implementation details. Future research may focus on standardizing these aspects to provide more directly comparable results and further optimize SVM performance in diabetes prediction.

3.3. Neural Network

The quality and amount of data are critical components of machine learning techniques like RF and SVM, which makes handling noisy and missing data difficult. Issues such as overfitting may happen when the model is extremely complicated or the training data is inadequate, leading to inferior performance on fresh data [8]. Therefore, Qi et al. [8] believes that using neural network to predict diabetes will be a new development direction in the future.

Sharma et al. [16] employed an Extreme Learning Machine (ELM), a specific type of Artificial Neural Network (ANN), renowned for its efficacy in addressing classification problems, to predict diabetes in the Pima Indians dataset. The methodology included the application of Principal Component Analysis (PCA) for feature selection, coupled with ELM for the classification task. The experiments were conducted in a cloud computing environment utilizing three distinct virtual machine configurations (vCPU-4, vCPU-8, and vCPU-16). The model demonstrated an impressive performance, achieving an accuracy of 90.57%, sensitivity of 82.24%, specificity of 73.23%, and an F-1 score of 75.03% when utilizing the vCPU-16 virtual machine. While achieving high accuracy, ELM also has the characteristics of avoiding local minima, fast convergence speed, and simplicity compared to other models.

Compared to ANN mentioned earlier, Deep Neural Network (DNN) is a model that places more emphasis on accuracy and can handle more complex and variable data than ANN. Ayon & Islam [17] used DNN approach to identify diabetes on many medical parameters to predict diabetes in the Pima Indians dataset. They discovered that five-fold cross-validation has an accuracy of 98.35%, which is higher than other techniques that are currently in use to predict diabetes mellitus at that time.

Qi et al. [8] developed a deep neural network named Kendall's correlation coefficient and an attention mechanism within a deep neural network (KCCAM_DNN), which utilizes the self-attention mechanism to prioritize essential features influential in diabetes prediction. This approach enhances the model's predictive capabilities. To further improve interpretability, they utilized the SHapley Additive exPlanations (SHAP) model to assess the contribution of each feature to the diabetes predictions. Experimental findings indicate that KCCAM_DNN demonstrates exceptional performance on the Pima Indian diabetes dataset, achieving test accuracy of 99.09%. These results indicate that KCCAM_DNN is highly effective for diabetes prediction, laying the groundwork for better-informed decisions in the diagnosis and prevention of diabetes.

The evolution of neural networks in diabetes prediction over recent years consistently demonstrates their efficiency and accuracy in handling large and complex datasets. This progression highlights the potential of neural network-based approaches to significantly improve diabetes prediction and management in clinical settings.

DOI: 10.54254/2755-2721/109/20241437

4. Challenges and Future Directions

Despite these developments, the area continues to confront obstacles because of data volume and ethical issues, which have an influence on ML use in healthcare from both a technological and moral standpoint.

Lack of Data Scale: Khan et al. [18] noticed that patient confidentiality is a major obstacle for the healthcare industry when it comes to data accessibility for ML models. Healthcare institutions are often reluctant to share health data, which makes it difficult for ML systems to continuously develop as fresh data is difficult to include.

Ethical Issues of AI: Alanazi, A. [19] emphasized that the advent of novel technologies, such as ML algorithms, has given rise to noteworthy ethical quandaries. These may be attributed to several factors such as inadequate legislative safeguards, innate prejudices against minority groups, and restricted technological capabilities. Experts stress that to develop ethical standards that address these concerns, thorough study is required. Furthermore, ML algorithms frequently make already-existing disparities in underlying data worse, which can disproportionately impact racial, ethnic, and gender groups. Therefore, creating thorough ethical frameworks is crucial to addressing these issues.

Even though the relevant application of machine learning in the field of diabetes prediction has reached a relatively mature stage, as mentioned above, it still has some problems in terms of morality and data scale. Therefore, it is believed that the future research direction should start from these aspects, acquire more data while protecting privacy, and develop relevant moral frameworks to solve the ethical problems of AI. In addition, the application of deep neural network in this field is also a new direction for the development of machine learning in the prediction of diabetes in the future.

5. Conclusion

In conclusion, the integration of machine learning (ML) techniques in diabetes prediction has demonstrated significant advancements in accuracy and efficiency, particularly through models such as Random Forest, Support Vector Machine and Neural Networks approaches. These methodologies show promise for early intervention and improved patient care. However, the field faces critical challenges related to data accessibility and ethical considerations. The reluctance of healthcare institutions to share patient data hampers the scalability and effectiveness of ML systems, while ethical dilemmas arising from biases and a lack of regulatory frameworks further complicate implementation. To address these issues, future research must prioritize data sharing while ensuring privacy protection and develop robust ethical guidelines to navigate the moral complexities of AI in healthcare. By tackling these challenges, the potential of ML in enhancing diabetes prediction and overall patient outcomes can be fully realized, paving the way for more equitable and effective healthcare solutions.

References

[1] Saeedi, P., Salpea, P., Karuranga, S., Petersohn, I., Malanda, B., Gregg, E. W., … Williams, R. (2020). Mortality attributable to diabetes in 20–79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Research and Clinical Practice, 162, 108086.

[2] Aminian, A., Zajichek, A., Arterburn, D. E., Wolski, K. E., Brethauer, S. A., Schauer, P. R., … Kattan, M. W. (2020). Erratum. Predicting 10-Year Risk of End-Organ Complications of Type 2 Diabetes With and Without Metabolic Surgery: A Machine Learning Approach. Diabetes Care 2020;43:852-859. Diabetes Care, 43(6), 1367.

[3] Mennickent, D., Rodríguez, A., Farías-Jofré, M., Araya, J., & Guzmán-Gutiérrez, E. (2022). Machine learning-based models for gestational diabetes mellitus prediction before 24–28 weeks of pregnancy: A review. Artificial Intelligence in Medicine, 132, 102378.

[4] Tanmay, N., Rexford, S. A., & Prasanna, S. (2021). Body fat predicts exercise capacity in persons with Type 2 Diabetes Mellitus: A machine learning approach. United States: Public Library of Science PloS one, 2021-03, Vol.16 (3), p.e0248039-e0248039, Article e0248039

[5] Singla, R., Singla, A., Gupta, Y., & Kalra, S. (2019). Artificial intelligence/machine learning in diabetes care. Indian Journal of Endocrinology and Metabolism, 23(4), 495–497.

DOI: 10.54254/2755-2721/109/20241437

[6] Jiang, J.-J., Sham, T.-T., Gu, X.-F., Chan, C.-O., Dong, N.-P., Lim, W.-H., Song, G.-F., Li, S.-M., Mok, D. K.-W., & Ge, N. (2024). Insights into serum metabolic biomarkers for early detection of incident diabetic kidney disease in Chinese patients with type 2 diabetes by random forest. Aging (Albany, NY.), 16(4), 3420–3530.

[7] Santhanam, T., & Padmavathi, M. S. (2015). Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. Procedia Computer Science, 47, 76–83.

[8] Qi, X., Lu, Y., Shi, Y., Qi, H., & Ren, L. (2024). A deep neural network prediction method for diabetes based on Kendall's correlation coefficient and attention mechanism. PloS One, 19(7), e0306090.

[9] López, B., Torrent-Fontbona, F., Viñas, R., & Fernández-Real, J. M. (2018). Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. Artificial Intelligence in Medicine, 85, 43–49.

[10] Alazwari, A., Abdollahian, M., Tafakori, L., Johnstone, A., Alshumrani, R. A., Alhelal, M. T., Alsaheel, A. Y., Almoosa, E. S., & Alkhaldi, A. R. (2022). Predicting age at onset of type 1 diabetes in children using regression, artificial neural network and Random Forest: A case study in Saudi Arabia. PloS One, 17(2), e0264118–e0264118.

[11] Yau, S. T.-Y., Leung, E. Y.-M., Hung, C.-T., Wong, M. C.-S., Chong, K.-C., Lee, A., & Yeoh, E.-K. (2024). Scoring System for Predicting the Risk of Liver Cancer among Diabetes Patients: A Random Survival Forest-Guided Approach. Cancers, 16(13), 2310.

[12] Usha Ruby, A., George Chellin Chandran, J., Swasthika Jain, T. J., Chaithanya, B. N., & Patil, R. (2023). RFFE - Random Forest Fuzzy Entropy for the classification of Diabetes Mellitus. AIMS Public Health, 10(2), 422–442.

[13] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.

[14] Yilmaz, N., Inan, O., & Uzer, M. S. (2014). A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases. Journal of medical systems, 38(5), 48.

[15] Edeh, M. O., Khalaf, O. I., Tavera, C. A., Tayeb, S., Ghouali, S., Abdulsahib, G. M., ... & Louni, A. (2022). A classification algorithm-based hybrid diabetes prediction model. Frontiers in Public Health, 10, 829519.

[16] Sharma, S. K., Zamani, A. T., Abdelsalam, A., Muduli, D., Alabrah, A. A., Parveen, N., & Alanazi, S. M. (2023). A diabetes monitoring system and health-medical service composition model in cloud environment. IEEE Access, 11, 32804-32819.

[17] Ayon, S. I., & Islam, M. M. (2019). Diabetes prediction: a deep learning approach. International Journal of Information Engineering and Electronic Business, 13(2), 21.

[18] Khan, B., Fatima, H., Qureshi, A., Kumar, S., Hanan, A., Hussain, J., & Abdullah, S. (2023). Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomedical Materials & Devices, 1(2), 731-738.

[19] Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. Informatics in Medicine Unlocked, 30, 100924.