Diabetes Prediction Based on Machine Learning

Yutong Sui[1,a,*]

[1]Rushan No.1 Middle School
a. 1707862909@qq.com
**corresponding author*

Abstract: Diabetes mellitus is a growing global health issue with an increasing incidence rate. Traditional methods for predicting diabetes, which rely heavily on clinical data and physical examinations, present challenges such as high costs, time-consuming processes, and difficulties in providing timely and personalized risk assessments. However, with the rise of machine learning (ML), new opportunities have emerged in diabetes prediction, utilizing large-scale data and advanced pattern recognition techniques. This study examines the application of ML in diabetes risk assessment by leveraging electronic health records (EHR) and big data, leading to significant improvements in accuracy and efficiency. The results demonstrate that ML methods can more effectively identify high-risk individuals, facilitating early intervention and contributing to the advancement of diabetes prediction.

Keywords: diabetes prediction, machine learning, public health, deep learning.

1. Introduction

Diabetes is an escalating public health concern globally, with its prevalence steadily increasing [1]. Traditional diabetes prediction methods rely on clinical data and physical examinations, which are often costly, time-consuming, and lack the ability to provide personalized, real-time risk assessments. The emergence of machine learning (ML) has revolutionized diabetes prediction by enabling the use of large-scale data and advanced pattern recognition techniques [2, 3]. The growing availability of EHRs and big data has enhanced the capability of machine learning models to significantly improve the accuracy and efficiency of diabetes prediction [4].

ML-based models offer the potential to identify high-risk individuals early, thereby facilitating timely interventions [5, 6]. These models not only aim to enhance predictive precision and reduce costs but also support data-driven healthcare decision-making. By offering personalized risk assessments, they can provide tailored health management strategies, optimizing diabetes prevention and treatment approaches [7, 8].

The development of accurate predictive models plays a pivotal role in improving health outcomes by enabling early detection of diabetes and mitigating the risk of complications [9, 10]. Personalized predictions based on individual health data contribute to customized intervention strategies, advancing the field of personalized medicine [11]. Furthermore, machine learning applications in diabetes prediction can optimize healthcare resource allocation, reduce costs, and provide a scientific basis for public health policy development, ultimately influencing diabetes prevention and management.

2. Traditional Methods vs. Machine Learning for Diabetes Prediction

Traditional diabetes prediction methods predominantly rely on clinicians' expertise and manual statistical analysis. These approaches suffer from low accuracy and are often inefficient for early diagnosis or risk assessment. Additionally, the manual nature of these methods limits their ability to scale and analyze large datasets, making it challenging to provide real-time, data-driven predictions. Machine learning offers computational techniques that mimic human cognitive processes like induction, generalization, and analogy, enabling systems to improve through experience. ML algorithms can automatically refine their performance based on data, making them ideal for medical applications. In diabetes prediction, ML can analyze vast datasets of patient information to create predictive models for early diagnosis and risk assessment. By identifying patterns and correlations within the data, ML systems can deliver more accurate, real-time predictions, supporting healthcare professionals in making informed decisions and improving patient outcomes.

3. Machine Learning Techniques for Diabetes Prediction

3.1. Supervised Learning

Supervised learning is widely used in diabetes prediction models. These models classify whether a patient is diabetic based on clinical attributes like age, BMI, and blood pressure. Numerous supervised learning algorithms, such as artificial neural networks (ANNs), decision trees, and support vector machines (SVM), have been employed for diabetes prediction. These studies demonstrate the effectiveness of supervised learning in handling complex, nonlinear relationships between clinical factors and diabetes outcomes.

For instance, Sapon et al. (2011) conducted a study using artificial neural networks (ANN) to predict diabetes, demonstrating the potential of ANNs in handling complex [12], nonlinear relationships between input features and the outcome variable. Diwani and Sam (2014) applied machine learning techniques such as Naive Bayes and the J48 decision tree algorithm to classify diabetic patients [13], showing how these methods can effectively manage categorical and continuous data in medical diagnosis. Mujumdar and Vaidehi (2019) introduced a diabetes prediction model that incorporated both conventional factors [14], such as glucose levels, BMI, and age, as well as external factors potentially contributing to diabetes, thereby enhancing the model’s classification performance. Muhammad et al. (2020) explored a range of supervised learning algorithms [15], including logistic regression, support vector machines, K-nearest neighbors, random forests, Naive Bayes, and gradient boosting, to develop robust predictive models for diabetes. Additionally, Nnamoko et al. (2018) exploited the diversity of heterogeneous base classifiers and employed feature subset selection to optimize model accuracy, highlighting the importance of feature engineering and model diversity in improving predictive performance [16].

3.2. Unsupervised Learning

In unsupervised learning, the data is unlabeled, and algorithms group the data based on similarities. Unsupervised learning methods, such as clustering, are effective in diabetes prediction by identifying hidden patterns within the dataset. These methods can complement supervised learning models by providing additional insights and improving prediction accuracy.

Harshvardhan et al. (2023) mentioned that using various criteria, we examined and assessed the performance of several unsupervised learning methods [17]. Md. Mehedi Hassan, Swarnali Mollick, and Farhana Yasmin also noted that unsupervised learning algorithms were used to group the diabetes dataset before supervised learning methods were employed to build the model from the diabetes dataset [18]. So, it can be seen that unsupervised learning is also an important form of machine learning. Kim et al. (2022) wrote that there have been outstanding approaches to combine the unsupervised learner and the supervised learner to solve such problems. In most cases, the unsupervised learner (component) detects the cluster relations of the input data and provides such information to the supervised component that learns the desired patterns between clusters and classes. Such a hierarchical combination of the unsupervised and the supervised learner provides a better chance for incremental learning and the overlapping clustering result in the unsupervised learning phase [19].

3.3. Ensemble Methods

Ensemble learning combines multiple models to improve predictive performance. Techniques like bagging, boosting, and stacking have proven effective in diabetes prediction by combining the strengths of various algorithms. Studies have shown that ensemble methods, such as Random Forests and other ensemble frameworks, consistently outperform individual models in predicting diabete.

It can be noted that numerous scientists have dealt with the prediction of diabetes through ensemble approaches. Priyanka and Rajendra et al. (2021) focused on establishing a predictive model for diabetes to ascertain whether a particular patient has diabetes and subsequently explored various techniques to improve accuracy [20]. Singh et al. (2021) proposed an ensemble-based framework named eDiaPredict [21]. Also, Laila et al. (2022) found that the Random Forest Ensemble Method had the highest accuracy (97%) [22]. Abnoosion et al. (2023) proposed an ensemble machine learning model (EMLM) through a combination of MLMs to enhance the prediction accuracy and AUC of diabetes disorders [23].

4. Data Sources and Features

4.1. Description of datasets commonly used for diabetes prediction

Popular datasets for diabetes prediction include EHRs, publicly available datasets like the Pima Indians Diabetes Database, and data from wearable devices. These datasets provide a rich source of information for developing and validating machine learning models. Negi et al. (2016) [24] developed a method employing combined datasets via machine learning techniques. This system is more dependable since it is trained, tested, and validated on a combined dataset. Naz et al. (2020) carried out an article on a deep learning approach for diabetes prediction using the PIMA India dataset [25]. The same subject was also explored by Ganesh et al. (2020), who made use of Pima Indians diabetes datasets to verify the effectiveness of different prediction models [26].

4.2. Important features and attributes used in prediction models

Key features in diabetes prediction models include demographic information (e.g., age, gender), clinical measurements (e.g., blood glucose levels, BMI), and lifestyle factors (e.g., diet, physical activity). Clinical data like blood glucose levels play a critical role in monitoring and predicting diabetes progression, while lifestyle factors, such as sleep duration and exercise, significantly impact diabetes risk. Integrating these features into prediction models enhances their ability to provide personalized health recommendations and improve early intervention strategies.

In any event, lifestyle factors are a factor that must not be disregarded. Gottlieb et al. have studied that sleep duration can influence diabetes [27]. Dietary habits are also crucial as they have an impact on diabetes [28]. Lifestyle is an extremely significant factor in the emergence and progression of diabetes and is one of the highly essential controllable elements. The poor lifestyle that gives rise to diabetes mainly comprises smoking, a sedentary lifestyle, lack of physical activity, and a diet high in salt, sugar, and fat. Such a lifestyle plays a very prominent role in the development of insulin resistance. If insulin resistance continues to deteriorate, it will lead to insulin deficiency and ultimately result in the occurrence of diabetes, which represents such a pathophysiological alteration. For the majority of individuals, it is extremely important to manage lifestyle in a timely manner and improve bad lifestyle habits. This is also the core aspect of primary prevention of diabetes. Since for all people, innate genetic factors, including the susceptibility genes for diabetes, cannot be altered. The only thing that can be changed is the acquired lifestyle. Cultivating a healthy lifestyle, quitting smoking, adhering to a low-salt and low-fat diet, a low-calorie diet, and engaging in more exercise can prevent the onset of diabetes to the greatest extent possible.

5. Conclusion

In conclusion, machine learning has proven to be a highly effective tool in diabetes prediction, offering significant improvements over traditional methods in terms of accuracy, efficiency, and scalability. By leveraging large datasets such as electronic health records, supervised, unsupervised, and ensemble learning techniques can accurately identify high-risk individuals, providing timely and personalized risk assessments. The integration of key clinical measurements, demographic information, and lifestyle factors enhances the predictive capabilities of these models. Ultimately, machine learning-based approaches have the potential to revolutionize diabetes prevention and management, improving public health outcomes through early detection and targeted interventions.

References

[1] *Ahmad, J. (2015). Management of diabetic nephropathy: Recent progress and future perspective. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 9(4), 343-358.*

[2] *Andersen, J. D., Stoltenberg, C. W., Jensen, M. H., Vestergaard, P., Hejlesen, O., & Hangaard, S. (2024). Machine learning-driven prediction of comorbidities and mortality in adults with type 1 diabetes. Journal of Diabetes Science and Technology.*

[3] *Metwally, A. A., Perelman, D., Park, H., Wu, Y., Jha, A., Sharp, S., & Snyder, M. (2024). Predicting type 2 diabetes metabolic phenotypes using continuous glucose monitoring and a machine learning framework. medRxiv: The Preprint Server for Health Sciences.*

[4] *Ge, X., Zhang, A., Li, L., Sun, Q., He, J., Wu, Y., ... & Gao, Y. (2022). Application of machine learning tools: potential and useful approach for the prediction of type 2 diabetes mellitus based on the gut microbiome profile. Experimental and Therapeutic Medicine, 23(4), 1-10.*

[5] *Mukkesh, K., Li Ting, A., Png, H., Ng, M., Tan, K., Loy, S. L., & Karnani, N. (2022). Automated machine learning (AutoML)-derived preconception predictive risk model to guide early intervention for gestational diabetes mellitus. International Journal of Environmental Research and Public Health, 19(11), 6792.*

[6] *Foppiani, A., De Amicis, R., Leone, A., Bertoli, S., & Battezzati, A. (2022). 25-OR: Machine learning for early detection of nonresponders to lifestyle intervention for prediabetes. Diabetes, 71(Supplement 1).*

[7] *Matboli, M., Al Amodi, H. S., Khaled, A., Khaled, R., Roushdy, M. M. S., Ali, M., & Aboughaleb, I. H. (2024). Comprehensive machine learning models for predicting therapeutic targets in type 2 diabetes utilizing molecular and biochemical features in rats. Frontiers in Endocrinology, 15, 1384984.*

[8] *Noaro, G., Cappon, G., Sparacino, G., Del Favero, S., & Facchinetti, A. (2020). Nonlinear machine learning models for insulin bolus estimation in type 1 diabetes therapy. 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 5502-5505.*

[9] *Lee, J., Choi, Y., Ko, T., Lee, K., Shin, J., & Kim, H. (2023). Prediction of cardiovascular complication in patients with newly diagnosed type 2 diabetes using an XGBoost/GRU-ODE-Bayes-based machine-learning algorithm. Endocrinology and Metabolism (Seoul, Korea), 38(3).*

[10] *Mora, T., Roche, D., & Rodríguez-Sánchez, B. (2023). Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms. Diabetes Research and Clinical Practice, 201, 110910.*

[11] *Zarch, M. E., & Masoud, S. (2024). Application of machine learning in affordable and accessible insulin management for type 1 and 2 diabetes: A comprehensive review. Artificial Intelligence in Medicine, 141, 102868.*

[12] *Sapon, M. A., Ismail, K., Zainudin, S., & Ping, C. S. (2011). Diabetes predictionwith supervised learning algorithms of artificial neural network. In International Conference on Software and Computer Applications, Kathmandu, Nepal (Vol. 9).*

[13] Diwani, S. A., & Sam, A. (2014). Diabetes forecasting using supervised learning techniques. Adv Comput Sci an Int J, 3, 10-18.

[14] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.

[15] Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive supervised machine learning models for diabetes mellitus. SN Computer Science, 1(5), 240.

[16] Nnamoko, N., Hussain, A., & England, D. (2018, July). Predicting diabetes onset: an ensemble supervised learning approach. In 2018 IEEE Congress on evolutionary computation (CEC) (pp. 1-7). IEEE.

[17] Joshi, S. (2023). Diabetes Prediction Using Unsupervised Learning. NEU Journal for Artificial Intelligence and Internet of Things, 1(1), 19-24.

*[18] Hassan, M. M., Mollick, S., & Yasmin, F. (2022). An unsupervised cluster-based feature grou** model for early diabetes detection. Healthcare Analytics, 2, 100112.*

[19] Kim, K. B., Park, H. J., & Song, D. H. (2022). Combining Supervised and Unsupervised Fuzzy Learning Algorithms for Robust Diabetes Diagnosis. Applied Sciences, 13(1), 351.

[20] Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.

[21] Singh, A., Dhillon, A., Kumar, N., Hossain, M. S., Muhammad, G., & Kumar, M. (2021). eDiaPredict: an ensemble-based framework for diabetes prediction. ACM Transactions on Multimidia Computing Communications and Applications, 17(2s), 1-26.

[22] Laila, U. E., Mahboob, K., Khan, A. W., Khan, F., & Taekeun, W. (2022). An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. Sensors, 22(14), 5247.

[23] Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC bioinformatics, 24(1), 337.

[24] Negi, A., & Jaiswal, V. (2016, December). A first attempt to develop a diabetes prediction method based on different global datasets. In 2016 fourth international conference on parallel, distributed and grid computing (PDGC) (pp. 237-241). IEEE.

[25] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19, 391-403.

[26] Sankar Ganesh, P. V., & Sripriya, P. (2020). A comparative review of prediction methods for pima indians diabetes dataset. Computational Vision and Bio-Inspired Computing: ICCVBIC 2019, 735-750.

[27] Gottlieb, D. J., Punjabi, N. M., Newman, A. B., Resnick, H. E., Redline, S., Baldwin, C. M., & Nieto, F. J. (2005). Association of sleep time with diabetes mellitus and impaired glucose tolerance. Archives of internal medicine, 165(8), 863-867.

[28] Sierra-Johnson, J., Undén, A. L., Linestrand, M., Rosell, M., Sjogren, P., Kolak, M., ... & Hellénius, M. L. (2008). Eating meals irregularly: a novel environmental risk factor for the metabolic syndrome. Obesity, 16(6), 1302-1307.