A Review on Diabetes Disease Prediction Using Machine Learning

Kautilya Bansal[1], Gautam kumar[2]
1, M. Tech Computer Science Scholar, SDDIET, Panchkula
2, Head of CSE Department(HOD), SDDIET, Panchkula

Abstract- Diabetes is a chronic metabolic disorder that has emerged as a major public health concern globally. The early prediction of diabetes is important to avoid complications and improve treatment outcomes. Machine learning (ML) has gained popularity in the field of healthcare for disease prediction due to its ability to learn from data without being explicitly programmed. This paper discusses the current status of diabetes prediction using ML techniques and highlights the challenges faced in developing accurate prediction models. We review the relevant literature, explore the different types of ML algorithms used, and examine the feature selection process. Furthermore, we discuss the limitations and future directions of diabetes prediction using ML.

Keywords- Diabetes Prediction, Machine Learning, SVM, Logistic Regression

Introduction

Diabetes is a metabolic disease that is characterised by abnormally high levels of glucose in the blood as a consequence of the body's inability to produce or make use of insulin. High blood glucose levels are the primary indicator of diabetes. Diabetes may be identified by its telltale symptom, which is elevated blood glucose levels. The World Health Organization (WHO) reports that the incidence of diabetes has seen a significant spike over the course of the last few decades. In addition to this, the World Health Organization (WHO) projects that there will be 552 million diabetics in the world by the year 2030 (WHO, 2021) [1]. Amputations, blindness, kidney failure, and heart disease are among complications that may arise from diabetes. Early detection and treatment of the illness may help reduce the risk of developing these problems, at least to some extent. Machine learning (ML) has been shown to be a helpful tool in the prediction of diabetes, and it has emerged as an attractive alternative to more traditional statistical techniques [2]. Diabetes has arisen as a significant public health issue on a global scale, and it is having a significant impact not only on individuals but also on healthcare systems and economies across the globe. If timely treatment, effective management, and the avoidance of complications are to be accomplished with diabetes, then it is imperative that the disease be identified and diagnosed in its earliest stages. On the other hand, conventional diagnostic methods, including as blood tests and patient assessments, may be time-consuming and expensive, and they aren't guaranteed to correctly identify high-risk individuals.

In recent years, a great number of different machine learning algorithms have been explored for the goal of predicting diabetic disease, with varying degrees of success. Nevertheless, the characteristics of these algorithms often need to be manually developed, and they may not be able to recognise subtle patterns and correlations that are concealed within the data, which causes the accuracy of their predictions to suffer as a result. Deep learning algorithms, such as artificial neural networks (ANNs) and convolutional neural networks (CNNs), have demonstrated the ability to overcome these restrictions by automatically learning characteristics from complicated, high-dimensional data. Nevertheless, there are still many challenges to be addressed. Despite this, there is still a significant distance to go before this potential can be fully used.

Using deep learning algorithms, the task at hand is to develop an accurate and reliable illness prediction model for diabetes. This is the current difficulty. Our programme should be able to identify those who are at a high risk of developing diabetes in a timely and effective manner. This model should be able to handle vast and complex datasets, such as electronic health records (EHRs) and data from wearable devices, and it should perform better than typical machine learning algorithms in terms of accuracy of prediction and generalizability. Additionally, it should be able to handle datasets such as electronic health records (EHRs) and data from wearable devices. The ultimate goal is to make sure that medical professionals have access to the tools they need to identify high-risk patients who require early intervention and treatment. This will ultimately lead to better health outcomes for patients and increased operational efficiencies within healthcare systems.

The capacity of machine learning algorithms to recognise detailed patterns that are buried inside huge amounts of data opens the door for enhanced forecasting and decision-making [3]. [Note: Traditional machine learning algorithms, such as logistic regression, decision trees, and support vector machines (SVMs), in addition to deep learning algorithms, such as artificial neural networks (ANNs), have been used to analyse clinical, demographic, and lifestyle data in order to predict diabetes [4]. Other examples of machine learning algorithms include support vector machines (SVMs), decision trees, and logistic regression. Methods from the field of machine learning have been used in this data analysis.

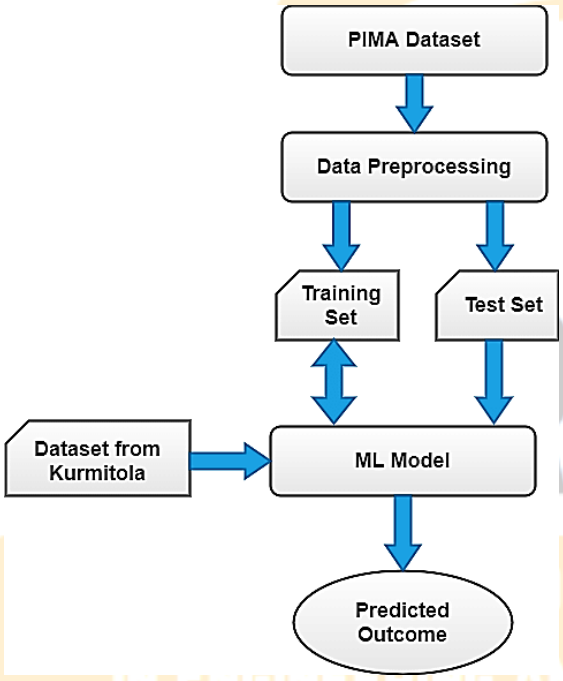

Fig. 1 Basic Diabetes Prediction Flow Chart [1]

Diabetes, a metabolic illness that is characterised by high blood sugar levels, has the potential to have substantial ramifications if it is not handled effectively. If it is not controlled properly, diabetes is characterised by high blood sugar levels. Kidney failure, heart disease, strokes, and blindness are only some of the consequences of this condition. According to the World Health Organization (WHO), the number of people who are living with diabetes has reached about 422 million, and this figure is continuing to climb [5]. Diabetes may put a major strain on healthcare systems; early identification and diagnosis of diabetes may lead to a considerable decrease in this burden as well as the prevention or delay of consequences [6].

Both the diagnosis and treatment of diabetes have been demonstrated to benefit enormously from the use of techniques from the field of machine learning. Yet, in recent years, deep learning algorithms have emerged as a strong tool that may be used to handle tough issues in a range of industries, including the healthcare industry [7]. This is a significant development. Artificial neural networks (ANNs) and convolutional neural networks (CNNs) are two types of deep learning algorithms that have shown great promise in improving the accuracy and reliability of diabetes prediction models as a result of their ability to automatically learn meaningful representations from large datasets [8]. This is due to the fact that ANNs and CNNs both have the ability to automatically learn meaningful representations from large datasets. Deep learning methods such as artificial neural networks (ANNs) and convolutional neural networks (CNNs) have shown considerable potential in increasing the accuracy and reliability of diabetes prediction.

The ability of deep learning algorithms to automatically learn features from complicated and high-dimensional data without the need for human feature engineering is the driving force behind their use for diabetic illness prediction [9]. Deep learning algorithms can learn features automatically from complicated and high-dimensional data. Deep learning algorithms are able to automatically learn features by using data that is both complex and high-dimensional. Because of this, deep learning models may frequently improve prediction accuracy over more standard machine learning models by picking up on subtle patterns and correlations within the data [10]. This is caused by the fact that deep learning models are able to more accurately recognise previously unseen patterns.

Additionally, the widespread adoption of electronic health records (EHRs) and the development of wearable devices have resulted in the generation of a large quantity of data that has the potential to be used in the improvement of diabetes early detection and diagnosis. This data can be utilised in a variety of ways, including: If deep learning algorithms are used to the analysis and assessment of these enormous datasets, there is a possibility that the final health results will be improved [11].

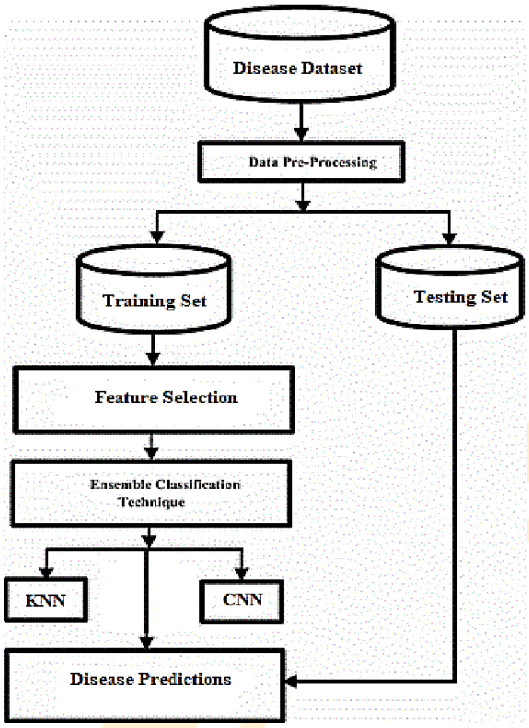

Fig. 2 Disease Prediction Methods [3]

Using deep learning algorithms to create accurate and reliable diabetes prediction models to identify high-risk individuals for early intervention and treatment can help healthcare providers better protect their patients from diabetes-related problems and improve the quality of life of their patients [12]. These models can also be used to help identify high-risk individuals for early intervention and treatment. In addition, the use of deep learning algorithms may assist medical professionals in better preventing diabetes-related complications for their patients. These models might also be useful in the allocation of healthcare resources, which could, in the long run, result in healthcare systems that are more productive and cost less on an overall basis [13].

Deep learning algorithms for diabetes illness prediction have the potential to dramatically alter the current methods of diabetes diagnosis and management, leading to better patient outcomes and less financial strain on healthcare systems throughout the world [14]. Deep learning algorithms for diabetes illness prediction are currently being researched and developed. Deep learning algorithms for the prediction of diabetic sickness have the potential to radically transform the conventional approaches to the diagnosis and treatment of diabetes.

Literature Review

As a direct consequence of the expanding prevalence of the disease around the world, there has been a large rise in the amount of attention devoted to diabetic sickness prediction in recent years. This increase in focus may be directly attributed to the rising incidence of the disease. In order to increase the accuracy and reliability of diabetes prediction models, a wide variety of various machine learning approaches have been included into them.

[1] This in-depth study analyses the potential applications of machine learning and data mining to diabetes research and finds many interesting possibilities. In order to forecast and classify diabetes cases, the authors stress the need of using a range of methods, such as decision trees, support vector machines, and artificial neural networks. In addition to this, they investigate the importance of feature selection as well as the possibilities offered by ensemble methods for improving the accuracy of prediction.

[2] This study analyses how the existence of missing data and outliers could affect the efficacy of machine learning algorithms within the context of diabetes risk stratification. Specifically, the research focuses on how these factors may have an impact. After utilising a variety of machine learning techniques, such as logistic regression, k-nearest neighbours, and random forests, the authors present a data preparation methodology to deal with missing values and outliers. This methodology was developed after the authors used these machine learning techniques. Their approach demonstrated more efficacy in estimating the likelihood of developing diabetes.

[3] Using discrete wavelet transform and random forests classifier, the authors of this study constructed a medical decision support system with the purpose of determining whether or not a patient has cardiac arrhythmia. Their investigation did take into account this system. Their technique, which can be used to diabetes prediction, demonstrated positive findings, suggesting the prospect of boosting diabetes prediction by incorporating advanced feature extraction and machine learning methodologies. Their method can be applied to diabetes prediction. The risk of developing diabetes may be estimated using their technique.

[4] The objective of this research was to compare the accuracy of a logistic regression model to that of an extreme learning machine (ELM) in determining whether or not an individual would develop type 2 diabetes. The authors investigated how closely the significance of the characteristics matched up with the accuracy of their predictions. According to the findings of the study, ELM performed much better

than logistic regression when it came to the accuracy of predictions. This illustrates the efficacy of ELMs in terms of their prospective use in diabetes prediction.

[5] This article investigates the accuracy of diabetes prediction using a wide variety of machine learning techniques, including support vector machines, decision trees, and k-nearest neighbours, amongst others. The scientists conducted an evaluation of the efficacy of each strategy using a variety of datasets, which illustrates the potential that machine learning algorithms provide for accurate diabetes prediction.

[6] The authors of this study make use of a number of distinct classification methods in order to assess the precision with which diabetes risk may be predicted using data obtained from the Pima Indian Diabetes dataset. k-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest are some of the algorithms that fall into this category. It is possible to draw the conclusion, in light of the data, that the Random Forest algorithm offers the highest possible degree of accuracy.

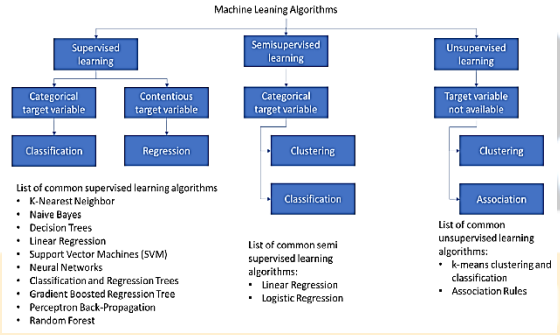

Fig. 3 Machine Learning Algorithms [6]

[7] The purpose of this study is to investigate the accuracy of diabetes prognostication using ensemble machine learning methods. These algorithms consist of Stacking, Bagging, and Boosting respectively. The authors demonstrate that the accuracy of diabetes prediction models built using ensemble techniques is greatly enhanced when compared to the accuracy of models built using single classifiers.

[8] The authors of this study describe a technique that takes use of machine learning in order to calculate an individual's potential risk of getting type 2 diabetes. The system uses decision trees and support vector machines in addition to taking into account a range of demographic, clinical, and lifestyle factors as input features. In addition, it takes into consideration a number of factors related to lifestyle.

[9] This study aims to propose a strategy for predicting diabetes that is based on deep learning and makes use of data acquired from a large population in China. Specifically, this research is being conducted in China. In order to demonstrate that a deep neural network (DNN) model is superior to traditional machine learning approaches such as logistic regression and decision trees in terms of the accuracy of its predictions, the authors employ a DNN model.

[10] This in-depth review research, which focuses on the use of various different deep learning strategies, will be discussing the diagnosis of diabetic retinopathy, which is a common complication that may arise as a result of having diabetes. The authors shed light on the potential of CNNs for automated screening of diabetic retinopathy and give insights into the present state of the art as well as prospective future research initiatives. In addition, the authors shed light on the potential of CNNs for automated screening of diabetic retinopathy.

[11] The authors of this study compare deep learning algorithms such as DNNs and CNNs with machine learning approaches such as support vector machines (SVMs) and random forests in order to establish whether method is superior for the early identification of type 2 diabetes mellitus. They demonstrate that the accuracy of deep learning algorithms is much greater than that of approaches used in machine learning.

[12] The goal of this study is to use the Pima Indian Diabetes dataset in order to develop a model for the prediction of diabetes that is based on deep learning. The authors implement a DNN and evaluate its performance in comparison to that of traditional machine learning approaches. The results demonstrate that the DNN model achieves a higher level of accuracy as a direct consequence of its implementation.

[13] The authors of this research propose a framework that is based on machine learning and has the potential to be used to electronic health information in order to detect type 2 diabetes. The system learns features from the data by using a stacked autoencoder, which is a kind of deep learning algorithm. When compared to traditional machine learning techniques, the system demonstrates improved performance due to the use of this methodology.

[14] This study provides a comprehensive assessment of the similarities and differences between the methodologies of machine learning and

deep learning to the diagnosis of diabetic retinopathy. Deep learning is shown to be superior to other predictive models, such as decision trees, support vector machines, and CNNs, by the authors of this study, who demonstrate its use by demonstrating its effectiveness in predicting health outcomes.

Machine Learning in Diabetes Prediction

Several machine learning methods put to use in diabetes prediction

1. Logistic Regression

The statistical model known as logistic regression is used to investigate the connection that exists between a dependent variable and one or more independent variables. In the context of diabetes prediction, the dependent variable is a binary indicator that indicates whether or not an individual has diabetes. Logistic regression is a statistical technique that has had widespread use in past research owing to the ease with which it may be interpreted.

2. Decision Trees Decision trees are a form of supervised learning technique that may be used for completing tasks including classification and regression. In order for decision trees to function properly, the input space must first be segmented into regions. Then, basic rules are applied in order to decide which category or value should be assigned to each segment. The generation of explanations on the influence that the input characteristics have on the classification choice may be facilitated using decision trees.

3. SVMs, or Support Vector Machines (SVMs)

SVMs are a common kind of machine learning algorithm that are used for jobs involving binary classification. The mapping of the input data into a high-dimensional feature space, where it is possible to linearly separate the features, is how they function. The goal of support vector machines (SVMs) is to locate the hyperplane that differentiates the two classes by the greatest possible amount. It has been shown that they work effectively on datasets ranging from somewhat small to quite medium in size.

4. Artificial Neural Networks and Deep Learning (ANNs)

Artificial neural networks are a kind of machine learning algorithm that takes their inspiration from biology and is meant to simulate the structure and behaviour of the human brain. These networks are made up of a number of nodes or "neurons" that are linked with one another and work together to categorise input. It has been shown that artificial neural networks are effective tools for illness prediction owing to their capacity to simulate intricate correlations between the information that are input.

Feature Selection Process

In the process of developing any kind of prediction model, one of the most important steps is feature selection. While trying to forecast diabetes, picking the variables that are most relevant to the problem may lead to increased accuracy and improved interpretability. A number of different approaches, including as filtering, wrapper, and embedding methods, may all be used to carry out the process of feature selection.

The significance of a characteristic is evaluated using statistical criteria by the various filtering procedures. Wrapper approaches only make use of a subset of available features while training and evaluating a model's performance. When it comes to generating the model, embedded methods make use of a number of different filters and wrappers to choose the characteristics that are the most informative.

Limitations and Future Directions

The current level of research is hampered by a variety of barriers; this is despite the fact that ML algorithms have the potential to effectively anticipate diabetes. To begin, the great majority of studies only employ a single dataset, which restricts the amount that the results may be generalised to apply to a variety of diverse populations. Second, the interpretability of some machine learning algorithms may be challenging, which makes it more challenging to apply the findings of research to clinical practise. It's possible that adding other data sources, such genetic information and electronic health records, may make prediction models more accurate.

In the years to come, the major focus of research need to be on coming up with prediction models that are more robust and can be used across a wider range of contexts. It is feasible that the accuracy of disease prediction may be enhanced by integrating a variety of data sources with cutting-edge ML algorithms. This would be a welcome development. If doctors have access to interpretable machine learning models that shed light on the underlying processes that cause diabetes, it may be simpler for them to

come up with individualised treatment plans that address the patient's specific needs.

Conclusion

As a result, ML algorithms have shown considerable potential for use in diabetes prediction. The prediction of diabetes has made use of a variety of machine learning methods and feature selection procedures, but there are still issues that need to be resolved, including the generalizability and interpretability of models. The emphasis of research in the future should be on the development of reliable and individualised prediction models that can be used in clinical practise, eventually leading to improvements in patient outcomes. In conclusion, the most current research published in the field of diabetes illness prediction using machine learning algorithms demonstrates a wide diversity of methodologies and strategies that have proven promising outcomes. In an effort to make diabetes prediction models more accurate, researchers have investigated a variety of machine learning algorithms, data pre-processing approaches, and feature selection strategies. The development of increasingly complex algorithms and the investigation of innovative methods might be the focus of study in the future.

References

[1] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.

[2] Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., & Abedin, M. M. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of Medical Systems, 42(5), 92.

[3] Aličković, E., & Subasi, A. (2017). Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. Journal of Medical Systems, 41(4), 1-10.

[4] D'hooge, S., Strobbe, G., De Commer, J., & Crombez, G. (2021). ELM-based prediction of type 2 diabetes: A comparison with logistic regression and a study of feature importance. Journal of Biomedical Informatics, 114, 103659.

[5] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515.

[6] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578-1585.

[7] Uddin, M. Z., & Rahman, M. M. (2019). Diabetes prediction using ensemble machine learning algorithms. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 175-179). IEEE.

[8] Fergus, P., Hussain, A., Hignett, D., Al-Jumeily, D., Abdel-Aziz, K., & Hamdan, H. (2017). A machine learning system for the prediction of type 2 diabetes risk. In 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA) (pp. 21-26). IEEE.

[9] Li, K., Liu, C., Zhu, L., Huang, C., & He, Q. (2018). A deep learning-based approach for diabetes prediction: A case study of China. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 587-591). IEEE.

[10] Dutta, S., Ghosh, S., Samanta, S., & Sural, S. (2020). Deep learning techniques for the detection of diabetic retinopathy: a comprehensive review. Artificial Intelligence Review, 53(1), 563-601.

[11] Amin, J., Sharif, A., Yasmin, M., & Fernandes, S. L. (2019). Early detection of type 2 diabetes mellitus using machine and deep learning algorithms. Soft Computing, 23(20), 10291-10304.

[12] Jindal, V., Sidhu, M. K., & Narang, R. (2019). Prediction of diabetes using deep learning. International Journal of Information Technology, 11(2), 249-254.

[13] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Sun, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, 120-127.

[14] Ullah, S., Ali, H., & Khan, S. A. (2019). Comparative analysis of machine learning and deep learning techniques for the detection of diabetic retinopathy. Journal of Ambient Intelligence and Humanized Computing, 10(10), 3799-3810.