Advances in diabetes prediction: a systematic literature review of Artificial Intelligence based methods

G R Ashisha[1,2], Sai Kiran Oruganti[3]

[1] Postdoctoral Researcher, Lincoln University College, Malaysia; [2] Assistant Professor, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India; [3] Associate Professor, Lincoln University College, Malaysia

Email ID: pdf.ashisha@lincoln.edu.my grashisha27@gmail.com

Abstract: Diabetes mellitus (DM), a common glycemic condition that causes substantial challenges to public health. The growths of Artificial Intelligence (AI) have created notable change in predicting DM, offering novel possibilities to lower its effect. This comprehensive review examined 25 articles concerning machine learning (ML) uses for DM prediction, emphasizing datasets, models, and evaluation techniques. Several datasets, including the Pima Indians Diabetes Database (PIDD), the National Health and Nutrition Examination Survey (NHANES), and REPLACE-BG, have been analyzed, highlighting their typical features and related issues, such as unbalanced data. This study evaluates the efficiency of various ML algorithms, including Support Vector Machines (SVM), Logistic Regression, XGBoost, and Convolutional Neural Networks (CNN), in predicting DM across several datasets. A few validation techniques are discussed, including k-fold cross-validation, and evaluation metrics including area under the curve, accuracy, sensitivity, and specificity. The result shows the importance of ML in handling the issues associated with DM prediction, and the need of maintaining models therapeutic relevance. With the ultimate goal of reducing the prevalence of this common disorder, this review helps current capability to use AI methods for better DM prediction.

Keywords: Predictive algorithms; Machine Learning; Artificial Intelligence; Diabetes Mellitus; Diabetes Dataset.

Introduction

Diabetes mellitus (DM) is a condition of metabolism marked by high levels of sugar in the blood brought on by either inadequate pancreatic production of insulin or incorrect insulin usage by the human body. The pancreas secretes the vital hormone insulin, which helps transfer glucose from the blood into cells so that it is transformed into glucose energy. Diabetes develops when glucose builds up in the blood due to insufficient insulin production or improper cell reaction. Diabetes can cause serious issues like renal failure, cardiovascular disease, and disability of nerves. According to the International Diabetes Federation (IDF), there will be 700 million cases of diabetes worldwide by 2045, emphasizing the critical need for novel therapies and prognostic techniques [1].

Research shows the changes of medical care using Artificial Intelligence (AI), mainly the effect of Machine Learning (ML) in illness prediction and management. These techniques effectively collect big datasets, find correlations, and make predictions. Accuracy of the ML techniques greatly depend on the technique, dataset, and the quantity of data utilized in the prediction model. When Continuous Glucose Monitoring (CGM) data is combined with genomic data and indicators in Electronic Health Record (EHR), the possibility of developing diabetes may be predicted more accurately than when CGM information is used purely. ML based diabetes prediction uses a variety of modeling and improvement methods, typically utilizing models like Support Vector Machine (SVM), Random Forest (RF), and logistic regression [2]. In order to improve the effectiveness of the performance, ensemble approaches are being employed more often, which are assessed utilizing performance like precision, recall, accuracy, sensitivity, F1-score, specificity, and Area under the Receiver Operating Characteristic Curve (AUC-ROC) [3]. This study helps researchers to find gaps and trends to enhance ML techniques. The findings are expected to direct future studies and enhance the performance of ML algorithms. This study addressed the following significant objectives:

- It begins by determining the attributes of datasets used for predicting diabetes and analyzing the way these features affect the performance of the ML techniques.
- Following that, it investigated the variety of ML techniques used to enhance the accuracy of DM prediction.
- Finally, it explored the evaluation metrics and validation techniques utilized, and bringing out the current research gaps and limitations.

This study illustrates the way diabetes is predicted using ML and covers the latest technological techniques used for diabetes care. The analysis emphasizes the application of AI technology to strengthen diabetes therapy by reviewing current research and highlighting important trends and gaps.

Related work

This part includes comprehensive reviews of previous research on predicting diabetes as well as data methodologies, prediction algorithms, and metrics implemented in DM prediction. Numerous studies from these publications are focused on ML models, which becoming the most significant topics in these days. Studies examine into related datasets and came to the conclusion that the quantity of data used in these studies is inconsistent based on the data sets evaluation.

Since the ML based prediction model of diabetes reviewed by Usman [4] only considered articles published from the year 2017, it may exclude earlier basic research. Important and relevant research papers may be missed if only 13 studies are used and only a small number of databases are examined. A bias in article selection may occur, that means the results of the review might not completely reflect all of the research that has been conducted in this diabetic prediction. In order to examine advanced approaches for DM prediction that use ensemble techniques, Wadghiri [3] carried out a study of the existing literature, including model types, publishing year, data sets, evaluation metrics, validation techniques, and performance. The findings demonstrated that ensemble techniques have become more widespread in recent years and that they have been a better model than the single ML algorithms.

The research work carried out by Bidwai [5] provided a summary that highlights the knowledge gaps and assisted investigators in examining practical results for ML based diabetic retinopathy prediction. It also

addressed key challenges, and limitations for creating accurate ML prediction techniques. Additionally, they identified 6 AI techniques as significant elements to their research. For the literature survey, publications were collected from PubMed, and Science Direct databases. As the review found several important gaps that should be considered in further research, problems like data consistency and the adding of various populations were found to be ignored. They concentrated on the irregular and unpredictable state of clinical data and addressed the uses of using ML algorithms in DM prediction.

Felizardo [6] conducted a survey of ML techniques that used diabetes data to predict DM. This systematic review includes 63 articles in total. It has not given attention on the dataset but it concentrate more on the performance of the ML techniques. Saxena [7] presented an extensive overview of the most recent studies on ML for DM prediction. The use of ML techniques and databases for DM prediction was the focus of this research. The findings demonstrate the efficiency of the RF technique, which is the popular technique used in research. The study investigated data reliability, sensitivity versus specificity decisions, inaccurate readings, and missing data issues in the DM prediction. The researchers also mentioned that the performance of ML models for DM diagnosis can be improved by expanding the dataset used for training, and focusing on outlier managing techniques. Feature selection techniques should be carefully selected to improve the performance of ML models. The following limitations of ML models for DM prediction have been listed in this study. Although RF algorithm is helpful and widely used in this area, the findings show that it can produce inaccurate results when it uses less reliable data.

Idrissi [8] identified and analyzed the effects of utilizing the diabetes data extraction techniques for DM prediction model. In this investigation, they selected 38 studies to classify and evaluate the research articles on the use of data mining approaches for DM prediction. According to this research, Artificial Neural Network (ANN) is the most widely used prediction model, followed by SVM. The contributors also pointed out some of the problems, such as the extreme complexity of controlling blood sugar levels, the absence of model generalization due to particular patients' data, and differences in measures employed in the evaluation of performance between studies. The researchers found that autoregressive models and ANN methods [9] have significant future potential to improve the prediction of diabetes. The investigation recommended further study on hybrid models, and highlighted their benefits for DM prediction. This review emphasizes the necessity to enhance existing prediction techniques, analyse new AI models, and utilize various types of data sets. Still, there are also disadvantages in the prediction models, including issues with reliability, data consistency, and prediction accuracy, that show the need for innovation in this area. These limitations can be reduced by creating the classification models, and improving the quality of the healthcare data.

Research Methodology

The main aim of this study is to examine the recent article on the use of ML for predicting diabetes. The study finds research gaps, trends and significant results in the expanding field of ML based disease prediction. It focused on existing predictive techniques, covering their methods, advantages, and disadvantages, and validation methods, and recommending the area for future work. To provide an organized process, the study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach [10]. PRISMA concept (Figure 1) provides a systematic method for identifying, assessing, and summarizing articles, allowing large volumes of investigation into meaningful

outcomes. The problem statement of this literature review concentrate on understanding the latest AI techniques utilized for predicting DM. The analysis consolidates the existing techniques and identifies the research gaps that need new algorithms. Its major goal is to investigate all main AI algorithms in DM prediction to acquire a better understanding of their technological constraints. The goals of this research are as follows:

1. To determine the significant features of DM prediction and analyze how these attributes help to improve the performance of ML methods.
2. To examine the variety of ML methods utilized to improve the performance of DM prediction.
3. To examine the performance measures and validation approaches used in this field and to identify the research gaps and shortcomings in existing literature.

This study examines the types of dataset used in DM prediction by estimating the reliability and variety of data reported in existing studies. It compares the AI models utilized to make sure an unbiased analysis of different models. By analyzing the validation models, the research estimates the reliability of current prediction models. Additionally, it highlights the potential directions and research gaps that are often neglected in studies. Finally, these objectives improve the quality of models and provide useful information on the existing prediction model.

Research Question 1 discusses the types of databases and their features used in DM prediction. It focuses on understanding the demographic information, diversity, and data size, which are crucial for creating reliable system. Prediction accuracy can be enhanced by analyzing these datasets in future study. By examining Research Question 1, Objective 1 can be achieved by examining the quality and standards of the datasets. Research Question 2 concentrate on examining the structure of ML models for DM prediction. It focuses the independent parameters utilized, such as demographics information and other healthcare variables, in addition with the classification methods to predict DM. This also investigates the variety of AI models that are utilized to enhance the performance of the models. Addressing this achieves Objective 2 by providing information on existing approaches and highlighting the areas for improvement. Research Question 3 focuses on the evaluation techniques used in predicting diabetes, especially the evaluation measures and validation techniques employed to estimate the effectiveness of the model. It examines common validation methods along with evaluation measures such as AUC-ROC, accuracy, sensitivity, and specificity. By investigating these evaluation settings, the study finds the directions for improvement in evaluating ML models. This question supports Objective 3 by enhancing the understanding of reliability of the model. Finally, this review helps highlight research gaps pertaining to data accuracy, computational difficulty, and model comprehension, thus, objective 3 is fulfilled and directs further research paths.

Search keywords and databases

When executing a systematic literature survey, especially in areas involving innovative technologies, like AI in medical care. The original sources of this review were ScienceDirect, and IEEE. In addition to that, a search in google was also performed to find AI based DM prediction. By skillfully using these sources and keywords, a search technique was followed for finding relevant publication. This phase is significant because it ensure that the procedure is systematic and has a direct effect on the quality of the review.

Criteria for inclusion and elimination

Inclusion and exclusion rules are required for choosing relevant literatures in this study. In this systematic literature survey, specific factors were applied to select the suitable resources specially

addressing all three research questions were included. The exclusion condition removed literature that satisfied the conditions: publications not in English language, conference proceeding publications, duplicate articles, and articles with very less pages. In order to further examine and analyze the data for addressing the research problem and accomplishing the defined objectives, all literatures that utilized AI techniques to predict DM were selected in our review.

Search query execution

After establishing the structure of systematic review, a detailed search method was developed to find related articles from Science Direct, and IEEE. A first search selected 500 literatures, which were decreased to 320 after eliminating duplications. After following the criteria of Inclusion and exclusion, 25 highly relevant articles were selected in this review.

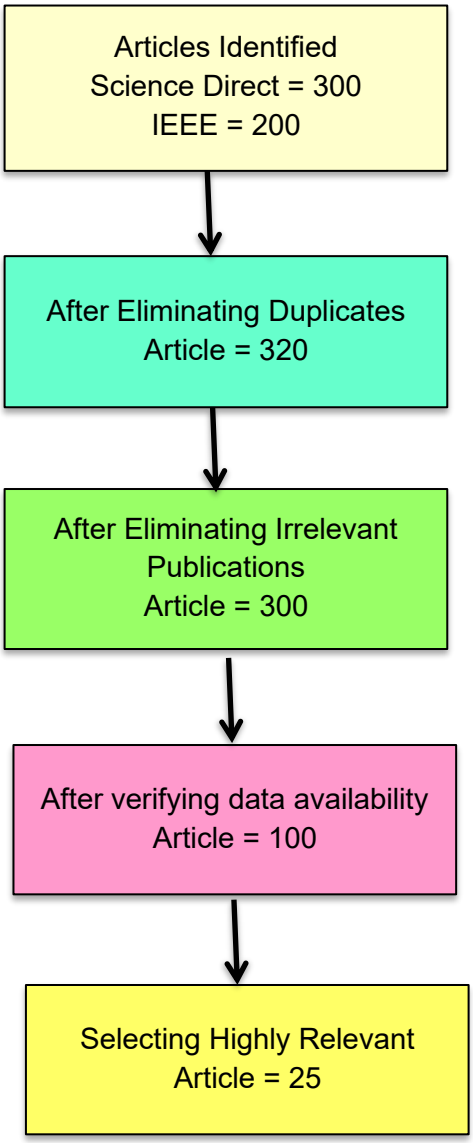

Figure 1. PRISMA structure for identifying relevant studies.

Measures like sensitivity, accuracy, AUC, and specificity are utilized to determine the efficiency of ML algorithms for DM prediction. These prediction methods depend on carefully selecting databases, suitable ML models, and validation techniques.

Results and Discussion

To obtain a systematic and credible analysis, all selected studies were divided based on condition including the research design, ML model, and evaluation measures. The objectives were then analyzed and outcomes from several studies were compared using qualitative as well as numerical methods. Journals are important for promoting scientific results. Our research focuses on diabetes, dataset, and ML algorithms. Common utilization of these terms emphasizes the significance of ML algorithms and analytical approaches in enhancing the performance of prediction model. Thus, the review highlights the importance of precise predictions to decrease the complications of diabetes.

Diabetes Dataset

The final selection of diabetes dataset is crucial for DM prediction research. Any ML model is developed on data, and the reliability of the dataset has a major impact on the performance of the model. These dataset many times include multiple populations, lifestyle factors, and countries, providing valuable information for DM prediction. Ongoing research confirms that several datasets use a variety of methodologies, which shows the extensive nature of research and offers a complete, multi-dimensional view of prediction.

Multicultural Diabetes Dataset

These multicultural dataset provide various kinds of demographic data allowing ML algorithms to be used to various populations. Table 1 shows different kind of diabetes dataset that can be utilized for diabetes research.

Table 1. Dataset for Diabetes Research

Diabetes Dataset	Name of the Population	Number of Data Records	Limitation
AusDiab [11]	Australia	11,247	Lifestyle features are missing
REPLACE-BG [12]	United State of America	226	Size is small
NHANES [13]	United State of America	10,000	Issues in Class Balancing
Optum EHR [14]	United State of America	95 Million	Missing Information
KNHANES [15]	Korea	20,000	Issues in Class Balancing
PIDD [16]	India	768	Size is small
Practice Fusion EHR	United State of America	1.2 Million	Inconsistent

[17]			Information
Aizawa [18]	Japan	11,247	Not generic to other population
Humedica [19]	United State of America	32 Million	Missing Information

AusDiab Dataset [11] is a study conducted in Australia which includes 11,000 subjects. This study contains several features like lifestyle factor, smoking habits, physical activity, employment details, and glucose level. This dataset can cause accuracy in the analysis due to its self-testing research. Diabetes Prediction Dataset (DPDS) [20] consists of significant attributes like Body Mass Index, demographic information, and glucose level. This data was collected in a single medical center and so it is not generic for other regions. Singapore Diabetic Retinopathy (DR) Screening Dataset [21] is a collection of retinal images which helps the researches to create an AI algorithms to detect DR. Randomized Trial Comparing Continuous Glucose Monitoring With and Without Blood Glucose (REPLACE-BG) [12] contains features like therapy information, glucose level, and glycemic indicators. It is a helpful dataset for developing ML based DM prediction model. National Health and Nutrition Examination Survey [13] was conducted in United States of America. It includes features like Blood pressure, Body mass index, glucose level, and cholesterol. This dataset is most popularly used dataset to predict DM. Optumo EHR dataset [14] is collected by hospital from United States of America. It includes features like medication data, demographic information, and laboratory values. This data will be helpful to predict future complications of diabetes. Korean National Health and Nutrition Examination Survey (KNHANES) [15] is a best dataset which will be helpful to identify the diabetes risk of Asia. It contains features like data from various glucose test, lifestyle factors, and biometric data. Humedica dataset [19] have 24,331 data records of diabetes patient. This dataset can be used to create a prediction model that can predict the future risk of diabetes.

Although various databases contribute in the DM prediction, they usually have shortcomings like low quality, demographic issues, and issue with privacy. Numerous datasets lack monitoring information, and reliable diabetic indicators.

AI Model for DM Prediction

AI based model has enhanced DM prediction by examining huge datasets and identifying complex designs. To provide accurate predictions and support better diabetes detection, algorithms like Logistic Regression (LR), Decision Tree (DT), SVM, Naive Bayes (NB), XGBoost, Convolutional Neural Network (CNN), and ANN use complex datasets [22]. Glucose level, gender, and age are the significant features that can be utilized for AI based prediction model. These features have an enormous impact on the performance of AI model. Choosing the right features are essential to achieve efficient predictive model. To eliminate overfitting, and to improve the prediction, feature selection, tuning parameters, and validation are essential. Specially, feature selection reduces irrelevant features from the dataset and enhances the efficiency of the prediction model, which is important in healthcare setting. Additionally validation technique can be included to validate the predictive system and to enhance or to prove the reliability of it.

In the analysis of DR, CNN models are most commonly used, particularly for the processing of retinal images to detect DR. SVM based prediction model can be used for any small size dataset to improve the efficiency and to avoid over fitting [23]. LR model can be utilized for any statistical analysis and to remove over learning 5 fold cross validation strategy is used. Demographic information and clinical data could be the main factors which influence the efficiency.

Table 2. Comparison of various ML models

Models	Advantages	Disadvantages
SVM	Better for structured information	Difficult to address complex data
LR	Extremely Interpretable	Not suitable for large dataset
RF	Avoids Overfitting	Costly
DT	Faster	Overfitting Issue

Validation Techniques

To enhance the reliability, k-fold cross validation technique is used to split the information into k parts in which it utilized one division for testing and remaining 9 divisions for training. By doing this overfitting got reduced and accuracy got increased. In most of the predictive system, 10 or 5 fold cross validation techniques have been used to validate the reliability of the model [24]. Additionally, external validation technique can be utilized to check how the model is suitable for medical devices. Some articles utilize bootstrap sampling, which can estimate the variability of the model performance. This can lead to ensure the stability, and accuracy of the ML model.

Evaluation Techniques

Several performance measures are utilized to estimate the prediction methods. Accuracy is the popularly used metric which will show the correct predictions. AUC is a main measure for any binary classification as it helps to separate positive and negative class. To detect positive and negative classes in the model, specificity and sensitivity metrics have been used [25]. To estimate the accurate positive class F1-score and precision measures are used. In clinical work, estimation of multiple measures helps to identify the accuracy of the DM prediction.

Discussion

This study finds they key elements involved in creating accurate prediction systems, which includes data identification, ML models, feature selection, and evaluation techniques. The study highlights the differences in population, size and the quality of the dataset. Data imbalance, missing data, and generalization are the major issues in this research. To eliminate these challenges, over sampling, ensemble model, and validation techniques can be incorporated. The results indicate that models like SVM, XGBoost, LR, and CNN are efficient for all data, but to achieve reliable model suitable feature selection algorithms and proper training methods should be included. This research still has various

limitations. Inconsistency, Cost, ethical problem and lack of collaboration between physicians and researchers are the major limitations. These challenges can be resolved by standardizing the data which will lead to a reliable DM prediction model.

Conclusions

This study demonstrates the growing significance of ML model by analyzing the publications. It reports the DM dataset, ML algorithms, feature extraction, and evaluation metrics used in DM detection, describing the models like XGBoost, LR, SVM, and CNN are useful. Explainable AI models are very important to increase the quality of the model. Efficient methods need proper feature selection model, and evaluation metrics like sensitivity, specificity, AUC, and accuracy. Despite existing limitations, future researches should focuses on standardizing the dataset, inclusion of various demographic regions, and ethical interaction to further improve the prediction model.

References

1. V. Ehrenstein, H. Kharrazi, H. Lehmann, and C. O. Taylor, "Obtaining data from electronic health records," in *Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide*, 3rd ed., Addendum 2. Agency for Healthcare Research and Quality, 2019.
2. Y. Qin, J. Wu, W. Xiao, K. Wang, A. Huang, B. Liu, et al., "Machine learning models for data-driven prediction of diabetes by lifestyle type," *Int. J. Environ. Res. Public Health*, vol. 19, p. 15027, 2022.
3. M. Z. Wadghiri, A. Idri, T. El Idrissi, and H. Hakkoum, "Ensemble blood glucose prediction in diabetes mellitus: A review," *Comput. Biol. Med.*, vol. 147, p. 105674, 2022.
4. T. M. Usman, Y. K. Saheed, A. Nsang, A. Ajibesin, and S. Rakshit, "A systematic literature review of machine learning-based risk prediction models for diabetic retinopathy progression," *Artif. Intell. Med.*, vol. 143, p. 102617, 2023.
5. P. Bidwai, S. Gite, K. Pahuja, and K. Kotecha, "A systematic literature review on diabetic retinopathy using an artificial intelligence approach," *Big Data Cogn. Comput.*, vol. 6, p. 152, 2022.
6. V. Felizardo, N. Garcia, N. Pombo, and I. Megdiche, "Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction: A systematic literature review," *Artif. Intell. Med.*, vol. 118, p. 102120, 2021.
7. R. Saxena, S. K. Sharma, M. Gupta, and G. Sampada, "A comprehensive review of various diabetic prediction models: A literature survey," *J. Healthcare Eng.*, vol. 2022, p. 8100697, 2022 (retracted).
8. T. E. Idrissi, A. Idri, and Z. Bakkoury, "Systematic map and review of predictive techniques in diabetes self-management," *Int. J. Inf. Manage.*, vol. 46, pp. 263–277, 2019.
9. P. Bidwai, S. Gite, K. Pahuja, and K. Kotecha, "A systematic literature review on diabetic retinopathy using an artificial intelligence approach," *Big Data Cogn. Comput.*, vol. 6, p. 152, 2022.

10. M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021.
11. M. Sangi, K. T. Win, F. Shirvani, M. R. Namazi-Rad, and N. Shukla, "Applying a novel combination of techniques to develop a predictive model for diabetes complications," *PLOS ONE*, vol. 10, p. e0121569, 2015.
12. P. Herrero, M. Reddy, P. Georgiou, and N. S. Oliver, "Identifying continuous glucose monitoring data using machine learning," *Diabetes Technol. Ther.*, vol. 24, pp. 403–408, 2022.
13. J. Kim, J. Kim, M. Kwak, and M. Bajaj, "Genetic prediction of type 2 diabetes using deep neural network," *Clin. Genet.*, vol. 93, pp. 822–829, 2018.
14. T. M. Usman, Y. K. Saheed, A. Nsang, A. Ajibesin, and S. Rakshit, "A systematic literature review of machine learning-based risk prediction models for diabetic retinopathy progression," *Artif. Intell. Med.*, vol. 143, p. 102617, 2023.
15. F. Tang, P. Luenam, A. R. Ran, A. A. Quadeer, R. Raman, and P. Sen, et al., "Detection of diabetic retinopathy from ultra-widefield scanning laser ophthalmoscope images: A multicenter deep learning analysis," *Ophthalmology Retina*, vol. 5, pp. 1097–1106, 2021.
16. Pima Indians Diabetes Database, "Pima Indians Diabetes Dataset (PIDD) documentation," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
17. B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, T. T. Do, et al., "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Comput. Methods Programs Biomed.*, vol. 182, p. 105055, 2019.
18. N. Yokota, T. Miyakoshi, Y. Sato, Y. Nakasone, K. Yamashita, T. Imai, et al., "Predictive models for conversion of prediabetes to diabetes," *J. Diabetes Complications*, vol. 31, pp. 1266–1271, 2017.
19. L. Li, C. C. Lee, F. L. Zhou, C. Molony, Z. Doder, E. Zalmover, et al., "Performance assessment of machine learning approaches in predicting diabetic ketoacidosis using electronic health records," *Pharmacoepidemiol. Drug Saf.*, vol. 30, pp. 610–618, 2021.
20. Z. Anggraeni and H. A. Wibawa, "Detection of retinal exudates using extreme learning machine," in *Proc. 3rd Int. Conf. Informatics and Computational Sciences (ICICoS)*, IEEE, 2019, pp. 1–6.
21. D. S. W. Ting, C. Y. L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, et al., "Development and validation of a deep learning system for diabetic retinopathy using multiethnic retinal images," *JAMA*, vol. 318, pp. 2211–2223, 2017.
22. S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: Review and case study," *Appl. Sci.*, vol. 9, p. 4604, 2019.
23. R. Casanova, S. Saldana, S. L. Simpson, M. E. Lacy, A. R. Subauste, C. Blackshear, et al., "Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning," *PLOS ONE*, vol. 11, p. e0163942, 2016.
24. S. B. Choi, W. J. Kim, T. K. Yoo, J. S. Park, J. W. Chung, Y. H. Lee, E. S. Kang, and D. W. Kim, "Screening for prediabetes using machine learning models," *Comput. Math. Methods Med.*, vol. 2014, p. 618976, 2014.

25. P. B. K. Chowdary and R. U. Kumar, “An effective approach for detecting diabetes using deep learning based on convolutional LSTM networks,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, pp. 519–526, 2021.