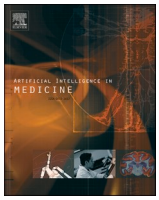
Artificial Intelligence In Medicine 164 (2025) 103132

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review☆

Pir Bakhsh Khokhar [*,1], Carmine Gravino [1], Fabio Palomba [1]

Department of Informatics, University of Salerno, Via Giovanni Paolo II, 132, Fisciano, 84084 Salerno, Italy

ARTICLE INFO

Keywords:
Systematic Literature Review (SLR)
Diabtets Prediction
Diabtetes Management
AI in Healthcare
Artificial Intelligence
Machine Learning
Deep Learning
Predictive Models
Medical Data Analysis

ABSTRACT

Diabetes mellitus (DM), a prevalent metabolic disorder, has significant global health implications. The advent of machine learning (ML) has revolutionized the ability to predict and manage diabetes early, offering new avenues to mitigate its impact. This systematic review examined 53 articles on ML applications for diabetes prediction, focusing on datasets, algorithms, training methods, and evaluation metrics. Various datasets, such as the Singapore National Diabetic Retinopathy Screening Program, REPLACE-BG, National Health and Nutrition Examination Survey (NHANES), and Pima Indians Diabetes Database (PIDD), have been explored, highlighting their unique features and challenges, such as class imbalance. This review assesses the performance of various ML algorithms, such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Logistic Regression, and XGBoost, for the prediction of diabetes outcomes from multiple datasets. In addition, it explores explainable AI (XAI) methods such as Grad-CAM, SHAP, and LIME, which improve the transparency and clinical interpretability of AI models in assessing diabetes risk and detecting diabetic retinopathy. Techniques such as cross-validation, data augmentation, and feature selection are discussed in terms of their influence on the versatility and robustness of the model. Some evaluation techniques involving k-fold cross-validation, external validation, and performance indicators such as accuracy, area under curve, sensitivity, and specificity are presented. The findings highlight the usefulness of ML in addressing the challenges of diabetes prediction, the value of sourcing different data types, the need to make models explainable, and the need to keep models clinically relevant. This study highlights significant implications for healthcare professionals, policymakers, technology developers, patients, and researchers, advocating interdisciplinary collaboration and ethical considerations when implementing ML-based diabetes prediction models. By consolidating existing knowledge, this SLR outlines future research directions aimed at improving diagnostic accuracy, patient care, and healthcare efficiency through advanced ML applications. This comprehensive review contributes to the ongoing efforts to utilize artificial intelligence technology for a better prediction of diabetes, ultimately aiming to reduce the global burden of this widespread disease.

1. Introduction

Diabetes mellitus (DM) is a metabolic disorder characterized by elevated blood glucose levels due to insufficient insulin production by the pancreas or improper insulin utilization by the body. Insulin, a crucial hormone secreted by the pancreas, facilitates the movement of glucose from the blood into cells, where it is converted into energy. It is also essential for the metabolism of proteins and lipids. When the body does not produce sufficient insulin or the cells do not respond to it properly, glucose accumulates in the blood, leading to diabetes.

Diabetes can lead to severe complications, including heart disease, kidney failure, and nerve damage. The International Diabetes Federation projects that global diabetes cases will reach 700 million by 2045[2], highlighting the urgent need for innovative treatments and predictive methods.

Traditional diabetes treatments focus on monitoring blood glucose

☆ This work presents a comprehensive systematic literature review on the application of artificial intelligence for diabetes prediction.

* Corresponding author.
E-mail addresses: pkhokhar@unisa.it (P.B. Khokhar), gravino@unisa.it (C. Gravino), fpalomba@unisa.it (F. Palomba).

[1] This is the first author footnote, also applicable to the second and third author.

[2] The International Diabetes Federation. Accessed July 7, 2024. https://idf.org/about-diabetes/diabetes-facts-figures/

https://doi.org/10.1016/j.artmed.2025.103132
Received 25 September 2024; Received in revised form 23 March 2025; Accepted 10 April 2025
Available online 15 April 2025

and HbA1c levels, which are reactive approaches for detecting the disease at an advanced stage. Thus, the need to develop better models for early prediction to improve the quality of life of patients cannot be overemphasized. Studies highlight the transformative impact of AI in healthcare, particularly through ML and DL in disease prediction and management [1]. These technologies effectively capture large datasets, recognize patterns, and make predictions previously deemed impossible. Since interest in applying ML for predicting diabetes has been on the rise, research in this field has received a boost. The accuracy of the ML models developed to predict diabetes relies greatly on the ML model and the type and amount of data used, such as Electronic Health Records (EHRs), laboratory data, age, gender, and other aspects of lifestyle [2]. The integration of Continuous Glucose Monitoring (CGM) data with EHRs has been more useful than the use of CGM data alone, especially in predicting health outcomes [3]. In addition, the integration of genetic information and biomarkers provides more information regarding the probability of developing diabetes [4].

The training of ML models for the prediction of diabetes includes different procedures and optimizations. Logistic regression, SVM, and random forest are widely used algorithms in supervised learning because of their interpretability and stability [5]. Other deep-learning models, including CNNs and RNNs, have also been used in data analysis to establish complex patterns [6]. Practices such as transfer learning and ensemble approaches have become more popular in efforts to improve the generalization and predictive capabilities [7]. The effectiveness of ML models in diagnosing diabetes, especially the accuracy of diagnosis, must also be determined for the models to be practical. Some of these assessment metrics include the accuracy, precision, recall, F1 measure, and AUC-ROC [8]. Sensitivity, specificity, and MCC are also employed to measure the accuracy of a model in identifying true positive and true negative results [9].

Systematic literature reviews (SLRs) are useful for offering an integrated analysis of the literature for a given subject area in the field of AI-based diabetes prediction. An SLR allows researchers to determine trends, gaps, and patterns in the use of AI for diabetes prediction so that techniques and results can be compared to enhance the creation of better ML models [10].

Through this SLR, specific details are highlighted on the current developments in ML-based diabetes prediction with respect to the datasets, training, and evaluation. These dimensions were scrutinized by the authors to determine the state of the art of research in the field, trends in the field, and possible areas of research that might have not been explored before. It is hoped that the results will help us find future work and improve the precision and applicability of ML models for diabetes prediction. This review addressed the following three main questions:

- First, it discussed the basic data and their properties utilized in the models for the prediction of diabetes and effect of these properties on the predictive models.
- Second, it described conventional training approaches and different ways to improve the accuracy of model and its ability to generalize.
- Lastly, it examined the evaluation criteria for ML models, especially on the commonly used metrics to measure the performance of the model.

Overall, this review provides valuable insights into the current use of ML in predicting diabetes, highlighting the technical aspects of model development and the practical implications for healthcare. By summarizing the current research and identifying key trends and gaps, this review contributes to ongoing efforts to leverage AI technology to improve diabetes care, ultimately aiming to reduce the global impact of this widespread disease and improve patient outcomes.

Structure of the paper: Section 2 provides an in-depth look at the anatomy of diabetes, reviewing previous research and existing systematic literature reviews to set the context for the current study. Section 3 details the stepwise approach of research questions guiding the review, the systematic review methodology, inclusion and exclusion criteria, database search strategies, quality assessment, and data extraction steps. Section 4 presents the findings from the reviewed studies, analyzing datasets used, machine learning algorithms, training strategies, and evaluation metrics. Furthermore, it interprets the results obtained from SLR. Section 5 discusses the limitations, potential biases, and gaps in the literature also provides implications for the researchers and stakeholders. Section 6 discusses the threats to validity of the work conducted in this study and finally Section 7 concludes the key findings, reaffirming the potential of machine learning in diabetes prediction and concluding with thoughts on the future of ML in healthcare and its role in improving diabetes prediction.

2. Background and related work

The main objective of this section is to equip us with background knowledge of the problem, so that we can proceed with our research. This section is devoted to the data methods, prediction models, and metrics used for diabetes prediction, and to systematic literature reviews conducted in the past regarding diabetes prediction.

2.1. Anatomy of diabetes mellitus

Diabetes mellitus (DM) is a metabolic disorder in which the body is unable to regulate the levels of sugar or glucose in the bloodstream, either due to inadequate insulin secretion by the pancreas (type 1) or insulin resistance (type 2). There are two main types of diabetes: Type 1 Diabetes Mellitus (T1DM) and Type 2 Diabetes Mellitus (T2DM), with Type 2 diabetes comprising nearly 90 % of all diabetes cases and with a global prevalence of 537 million [11]. Diabetes is a community health problem that has shown an alarming increase over the last 20 years in many parts of the world. DM is a multiorgan disease with numerous diabetic microvascular complications involving the retina, heart, brain, kidneys, and nerves.

The role of medical personnel in the prevention, treatment, and management of diabetes mellitus and its complications is well-established [12]. Exercise prescriptions and education for rehabilitation management are effective for participation and maintaining physical well-being, improving patient health, and improving health-related quality of life [13]. Diabetes itself is not a high-mortality cause, but it is a significant risk factor for other causes of death and has a high disability burden. Diabetes is also a significant risk factor for cardiovascular diseases, kidney diseases, and blindness [14]. DM is categorized into three types according to its etiology and clinical manifestations: type 1, type 2, and gestational diabetes [15].

Diabetes Mellitus primarily involves the islets of Langerhans in the pancreas, from which glucose is secreted from alpha cells and insulin from beta cells. Glucagon increases blood glucose levels, and insulin reduces glucose levels. T1DM (Insulin-Dependent) is a chronic metabolic disorder that causes 510 % of diabetes mellitus [16]. It is characterized by the autoimmune destruction of insulin-producing beta cells in the islets of the pancreas, and the loss of function of beta cells leads to absolute insulin deficiency. T1DM is most commonly seen in children and adolescents but can affect anyone at any age.

T2DM (Non-insulin dependent) comprises 90 % of all diabetes [15]. Reduction in the effect of insulin on T2DM is called insulin resistance. Under normal conditions, insulin is ineffective and, therefore, is initially countered by an increase in insulin production to maintain glucose homeostasis but later decreases to cause T2DM. T2DM is common in adults aged 45 years or older [17]. It is now more prevalent in children, adolescents, and younger adults as a result of rising obesity, physical inactivity, and energy-dense diet.

Gestational Diabetes Mellitus (GDM) can occur at any stage of pregnancy. Typically, it occurs in pregnant women during the second and third trimesters. The American Diabetes Association (ADA)

estimates that GDM occurs in 7 % of pregnancies. GDM patients and their offspring are at an elevated risk of developing type 2 diabetes mellitus in the future [18].

2.2. Related work

In previous years, as evidence shows, several systematic literature reviews emphasize the diagnosis of predicting type 2 diabetes and studies concerning those predictions. Many articles from these journals and conferences are centered on Machine Learning and Deep Learning techniques, which are among the most relevant topics today. They aimed to investigate similar datasets and concluded through the data sets analysis that the amount of data used in those studies is unstable.

The research conducted by Bidwai, P. et al. [19] suggested a new review that aimed to eliminate the gaps left by current reviews and help other researchers in selecting the current results from the studies that they can use in predicting ML-based risk of Diabetic Retinopathy progression and related diseases by synthesizing the current results from these studies and putting in place the research challenges, limitations, and gaps for the selection of efficient machine learning techniques in the establishment of my model of prediction. Furthermore, they pointed out six AI-related technical discussions and approaches as these two crucial points for the adopted strategy. For the SLR, data collection was used to obtain suitable studies. They searched the IEEE Xplore, PubMed, Springer Link, Google Scholar, and Science Direct electronic databases for literature reviews published between January 2017 and April 30, 2023. Thirteen (13) studies appearing in the broad discussion were subsequently shortlisted based on their relevance to the reviewing questions and the filters applied. While the literature review revealed some significant research gaps to be considered in future research that will improve the performance of Diabetic Retinopathy (DR) progression risk prediction models, issues such as comparability and inclusion of diverse DR populations are inattentive.

They also discussed different approaches to the problem of diabetes prediction in general and the problem of selecting and integrating necessary research articles for ML-based diabetic prediction models. They discussed how the medical data are nonlinear, non-normal, and correlation structured, and how beneficial machine learning is in healthcare, especially in medical imaging. While their review was not comprehensive in some of the areas of interest, especially in early diagnosis and risk stratification, it provided researchers with a source of reference. However, the current systematic literature review (SLR) follows the PRISMA guidelines much more closely to ensure a more exhaustive and objective approach to analysis and provides a discussion of the practical recommendations for further research that would consider the intricacies of medical data for diabetes prediction.

This may preclude older basic studies because ML-based risk prediction of DR progression as shown in the study by Usman et al. [20] is limited to papers published between January 2017 and April 2023. Using only 13 studies and a few databases may not have identified all relevant materials, which can lead to selection bias. The authors did not extensively discuss the comparability and inclusion of different DR populations, which would influence the generalizability of the findings. Our SLR alleviates these limitations by focusing on a more extended period (from 2014 to 2023), covering more first-hand papers (53), and incorporating more criteria such as algorithms, datasets, and validation methods. This methodological approach increased the likelihood of identifying relevant and inclusive studies. Thus, this study provides a more comprehensive synthesis of the literature as a foundation for future research on blood glucose prediction and DR progression.

A systematic literature review performed by Wadghiri et al. [18] aimed to review state-of-the-art methods for predicting blood glucose using ensemble methods based on eight criteria: types of algorithms, year of publication, journal, database, types of ensembles, learners, combination methods, performance measures, validation methods, overall performance, and accuracy. This systematic literature review was performed to compare the primary studies on digital libraries from to 2000–2020. Among the 32 primary papers reviewed, eight review questions were chosen for this study. The results indicated an increase in the use of ensembles in recent years; overall, they were better than the other single models. However, the process of formation of the groups and performance criteria is not entirely flawless. Here, some suggestions are provided regarding the design of compelling ensembles for blood glucose level prediction.

Digital libraries may have missed some crucial studies. The exclusion of some research and a small number of evaluated primary papers may affect the comprehension and bias of the selection process. However, the study approach for measuring ensemble formation and performance has limitations. This discovery is particularly relevant to blood glucose prediction and may not be applicable to other contexts.

Datasets and validation methods also affect the dependability. Finally, the pace of technological progress may render certain conclusions outdated and irrelevant. This study addresses these concerns by analyzing various databases.

The review by Eijoseno, M.R et al. [21] was designed to present diabetes in general, its prevalence, complications, and opportunities for artificial intelligence in early diagnosis and classification of diabetic retinopathy. The research also focused on ML-based methods such as machine learning and deep learning. New research areas including transfer learning using generative adversarial networks, domain adaptation, multitask learning, and explainable artificial intelligence in diabetic retinopathy were also considered. A list of methods already in use, screening systems, performance measurement, biomarkers in diabetic retinopathy, potential issues, and challenges in ophthalmology. The future scope is elaborated upon in the conclusion section. The review may lack systemic rigor because it focuses on diabetes and ML methods without using the Preferred Reporting Items for Systematic reviews and meta-analyses (PRISMA). Only a few powerful ML algorithms can be included, whereas others are omitted. In addition, the assessment may not provide immediate practical suggestions while planning future work. Our SLR is more rigorous and systematic, because it conforms to the PRISMA framework. It also includes more criteria and approaches and provides a more comprehensive analysis and application recommendations for future research on AI-based prediction.

A comprehensive review by Saxena et al. [22] presented the current literature on machine learning for diagnosing DM. This research dealt with the use of machine learning models and datasets for the diagnosis of diabetes. The results show how Random Forest can be used successfully and how it is prevalent in this area of research. Prompt diagnosis of diabetes is essential because it helps to control the disease and avoid complications. Nevertheless, the fact that people have no access to care and that there are cases that go undiagnosed are also challenging. The analysis presented problem areas such as data quality, sensitivity-specificity trade-offs, incorrect readings, and missing data. The authors further stated that future research must be expanded by enlarging the training dataset, including additional parameters, and addressing outlier handling methods to overcome these challenges. Moreover, feature selection methods, the issues of which are more critical, sensitivity, or specificity, should be considered. Although this process has some problems, machine learning can make diabetes detection easier and improve medical care. Therefore, the present research gives future researchers a chance to learn more about implementing ML algorithms for diabetes diagnosis.

This review has listed the following drawbacks of ML for DM diagnosis: Random Forest is practical and widely used in this field, but the assessment identifies data quality issues, the sensitivity-specificity curve, false readings, and incomplete records. The study also suggests more significant training datasets, parameters, and better outlier handling. This also implies improved feature selection and a better understanding of the relationship between sensitivity and specificity. However, our SLR aims to overcome these constraints by becoming more inclusive. It enforces the PRISMA framework for systematic data

gathering and analysis, encompasses several machine-learning techniques and considerations, and provides actionable research recommendations. This comprehensive strategy improves ML for diabetes diagnosis and resolves the emphasized issues.

A systematic review of the literature on data-driven algorithms and models was performed by Felizardo, V. et al. [23] using accurate diabetic data to predict hypoglycemia. The review process was intense and spanned five electronic databases: ScienceDirect, IEEE Xplore, ACM Digital Library, SCOPUS, and PubMed, covering publications from January 2014 to June 2020. This search yielded 63 studies that were included in the analysis owing to their relevance. This review showed that data models developed for predicting blood glucose and hypoglycemia might have to balance applicability and performance. This has resulted in the integration of other data sources or the use of different modeling approaches. The study outcomes proved the current trends and prompted further research on hypoglycemia prediction. This systematic analysis of data-driven hypoglycemia prediction comprised 63 articles from five databases from 2014 to 2020. Although comprehensive, the brief timespan may omit recent developments. It may not pay much attention to data variety and the combination of its distinct techniques to focus on the applicability and performance of the model.

El Idrissi et al. [24] mapped and reviewed existing literature that explored the use of data mining (DM) predictive techniques in diabetes self-management (DSM). In their review, they preferred 38 papers published between the years 2000 and April 2017 to categorize and review the literature on the application of DM techniques for DSM tasks, including blood glucose level prediction, hypoglycemia detection, and insulin dose estimation. The review established that artificial neural networks were the most popular type of predictive technique, followed by the auto-regressive type of models, and support vector machines. Interestingly, in the majority of investigations concerning T1DM, the most frequent clinical issue was blood glucose prediction, which was the target of more than 57 % of the selected investigations.

The authors also highlighted some of the issues, including the lack of model generalization as a result of patient-specific data, high complexity involved in regulating blood glucose levels, and variations in metrics used in the assessment of results across studies. Nevertheless, the review pointed out that DM techniques, such as ANNs and autoregressive models, could hold a significant future capacity for enhancing DSM prediction accuracy and decision-making. Nevertheless, the study called for more research on the use of hybrid models and the extension of these techniques in T2DM and gestational diabetes, as well as the need for a more standardized experimental design in future research.

This study employs PRISMA for further rigor and coverage. It includes research conducted from the foundational to the contemporary period, and covers 20,142,023. Compared with evaluating the model performance, our evaluation involves algorithms, datasets, validation, and challenges for blood glucose prediction, which provides a more comprehensive and applicable perspective for this research.

These studies aim to provide an in-depth discussion on diabetes mellitus and include discussions on the various types of diabetes, the number of people affected by diabetes, and the different health complications associated with diabetes. They emphasized the importance of systematic reviews and ML-based strategies when studying the use of ML and deep learning (DL) technologies for the effective prediction and management of diabetes through the analysis of the application of these technologies.

However, several limitations remain, including dataquality issues, system interoperability challenges, and disease classification. These limitations underline the fact that continuous innovation in this discipline is necessary. This research emphasizes the importance of developing current predictive models, exploring novel approaches to artificial intelligence, and utilizing various data sources to enhance the efficiency and accuracy of diabetes prediction tools. To overcome these constraints, there is a need to improve the quality of data, establish better approaches for system integration, and improve classification algorithms to create more effective and applicable artificial intelligence models for diabetes prediction.

3. Research approach

The predominant goal of this study is to achieve critical systematic integration and provide a summary of the latest published scientific literature on the application of machine learning in predicting and managing diabetes. This review analyzes emerging trends, identifies gaps, and summarizes key takeaways in the rapidly evolving field of AI for diabetes prediction. This review aims to examine the predictive models; additionally, it outlines the approaches used, both the strengths and weaknesses, used datasets, training and validation strategies, categorizes the effectiveness of the current hypothesis, gives a critique, and considers the areas to advance further research. To achieve this, the review process must be arranged thoroughly according to the PRISMA framework [25]. With the solemnity and complexity of the procedure, PRISMA is considered a reference for conducting systemic reviews, supporting hearings, and ensuring clarity in the appraisal of scientific literature. It provides a systematic technique that is evaluative regarding literature selection, assessments, and syntheses. This makes it an appropriate analytical tool that condenses vast research findings into coherent conclusions.

3.1. Research objectives and research questions

The research questions designed for the systematic literature review aim to answer how machine learning and artificial intelligence are used for diabetes prediction and establish a framework for the current state-of-the-art in the field. This review aims to summarize and analyze previous studies while identifying gaps where technological innovations and new methodologies are needed. The main objective is to conduct a systematic review of all possible areas of the application of machine learning and artificial intelligence technologies in diabetes prediction to create a framework for understanding the limitations of what is technically feasible and clinically applicable. The objectives of this study are as follows:

Objective 1: To identify and synthesize the findings on datasets with their characteristics utilized in diabetes prediction.

Objective 2: To examine the configurations and the range of ML techniques used in diabetes prediction.

Objective 3: To analyze evaluation setups and performance metrics used in ML models to predict diabetes.

Objective 4: To identify the limitations of current research in diabetes prediction.

This study considers the type of data used in AI-based diabetes prediction by analyzing the datasets highlighted in the reviewed studies. This adds comprehensiveness to the evaluation of AI models by considering the nature, quality, and representativeness of data. We compared the methodologies of the primary consideration and the configuration of the ML and DL algorithms recommended in this study. This eliminates assumptions about the models and ensures that a wide range of approaches are considered, thereby strengthening the methodological robustness of the study. When assessing how different studies ensure the validity of their models, this review provides insights into the reliability and generalizability of AI-based diabetes prediction systems. This objective enhances the rigor of the review by addressing reproducibility and benchmarking. One of the integral steps in constructing an SLR is identifying its limitations and future directions, which are often overlooked. Evaluating the current gaps in AI-based diabetes prediction studies allows this review to not only synthesize past research but also highlight opportunities for future advancements. Collectively, these objectives methodically establish the procedural foundation of the SLR, ensuring that the review remains comprehensive and methodologically sound while offering valuable insights into AI-based diabetes prediction.

RQ_1. What datasets, including their characteristics, have been utilized in research studies focused on diabetes prediction?

This research question aims to discover and explain the datasets that have been used in studies that have focused on diabetes prediction. Through this process, we can estimate the size of the data, diversity, and representativeness of the population, which are crucial for developing a robust and applicable model for different populations. Additionally, analyzing these datasets will also reveal any deficits in data utilization that could be corrected in future studies, thereby contributing to an improvement in the accuracy and generalizability of diabetes diagnostic tools. Through this study, we set data standards for diabetes prediction research and thus provide a basis for other studies to build upon the foundations of such data.

By addressing RQ_1, we can meet **Objective 1**. This study aimed to explore these datasets and their characteristics for predicting diabetes. Understanding these elements helps us assess current data standards and identify potential gaps in data utilization.

RQ_2. What are the configurations of ML approaches used in diabetes prediction, including the independent variables, classification types, ML algorithms, and training strategies?

This extensive research topic aimed to investigate the peculiarities of the application of artificial intelligence tools in diabetes prediction. It seeks to understand the various components that contribute to the development and optimization of ML models in this context. This includes identifying independent variables considered influential in predicting diabetes, such as patient demographics, health metrics, and genetic information. Additionally, it explores the classification types used to differentiate between outcomes, such as distinguishing between diabetes types or predicting disease progression stages. The question also investigates the range of AI algorithms, from traditional machine learning to advanced deep learning techniques, harnessed to analyze and interpret complex datasets effectively. Finally, it examines the training strategies implemented to enhance the model performance, including methods for training data selection, model validation, and techniques to prevent overfitting. Understanding these aspects provides a clear picture of the current state of AI applications in diabetes prediction, and identifies areas for potential improvement and innovation. By examining the configurations of ML approaches, using independent variables, and finding classification types, we will be able to meet **Objective 2**.

RQ_3. What are the various evaluation setups employed in the context of diabetes prediction, explicitly focusing on the types of validation methods used and the metrics applied to assess the effectiveness of these models?

This research question seeks to explore and characterize the evaluation frameworks used in the context of diabetes prediction, emphasizing the methodologies applied for validating predictive models and the metrics used to measure their effectiveness. It aims to understand the diversity and robustness of validation techniques, such as cross-validation, bootstrapping, and external validation, to ensure the reliability and generalizability of ML-driven prediction models. In addition, this study analyzes the specific performance metrics used to evaluate these models. These metrics include accuracy, sensitivity, specificity, and AUC-ROC. By analyzing these aspects, this research can identify best practices and potential areas for enhancement in the assessment of ML models, contributing to improved outcomes in diabetes prediction. RQ_3 plays a crucial role in fulfilling **Objective 3** by providing detailed insights into the setup of ML models. This includes independent variables, classification types, ML algorithms, and training strategies, offering a comprehensive view of how ML tools are tailored to enhance the predictive accuracy in diabetes care.

Therefore, this research examines datasets, ML methods, and training schemes to understand the shortcomings of existing diabetes prediction research, where shortcomings in data quality and characteristics are highlighted. Others include overfitting, computational burden, and interpretability of the model. The investigation delineates the blind spots of the current research and outlines further study directions by fulfilling **Objective-4**, which will enhance the credibility and generalizability of the diabetes prediction models.

3.2. Search databases and search queries

When conducting a systematic literature review, particularly in fields involving advanced technologies, such as ML in healthcare, selecting suitable databases for search is crucial. The primary sources of literature were three major databases: IEEE, PubMed, and ScienceDirect. Additionally, a Google search was conducted to identify AI and medical databases with substantial coverage of diabetes prediction. The selection of IEEE, PubMed, and ScienceDirect as core databases aligns with methodologies employed in previous studies on AI and diabetes prediction. For instance, Gargeya and Leng (2017) conducted a systematic review on AI for diabetic retinopathy screening using PubMed and IEEE [26], highlighting their relevance for technical and medical research. Similarly, Sneha and Gangil (2019) utilized PubMed and ScienceDirect to explore AI in diabetes management, demonstrating their validity in capturing high-quality studies within the medical domain [27]. These studies affirm the appropriateness and acceptance of the selected databases for review in this field. While this review primarily utilizes IEEE, PubMed, and ScienceDirect, the exclusion of databases such as Web of Science and Scopus may limit the comprehensiveness of the literature search.

In the context of systematic literature reviews, a research query definition specifies the exact terms, scope, and parameters of a search strategy used to gather relevant literature on a given topic. This definition is crucial, as it directly influences the quality, relevance, and comprehensiveness of the collected literature. Defining a research query helps to ensure that the review is systematic, reproducible, and closely aligned with the research objectives. In line with this approach, the study period begins in 2014, marking the emergence of deep learning in healthcare [28–30]. During this time, large-scale medical datasets, such as the National Health and NHANES, Optumő EHR, and EyePACS became increasingly available, provide essential data to support the development of AI-driven healthcare models. Additionally, advancements in computational power, particularly GPU-based AI training, have facilitated the practical application of deep-learning techniques in diabetes prediction. The notable increase in AI research publications since 2014 further validates this choice, ensuring that our review covers the most relevant and impactful studies in this field and we set the following strategy:

- Specific words and phrases were used in the database search. These are usually derived from the main topics of the research questions and are critical for retrieving relevant literature. Keywords were carefully selected to capture the various aspects of the investigated topic.
- For all those keywords, we searched for synonyms, alternative spellings, and other names for the disease.
- We incorporated Boolean operators (“AND”, “OR”) to formulate search queries.

This section provides a detailed description of the critical databases with their search queries that were used for such research, highlighting their specific relevance and benefits.

PubMed is the premier database for anyone researching medical and healthcare topics. Managed by the National Institutes of Health, PubMed provides access to more than 30 million citations of biomedical literature from MEDLINE, life science journals, and online books. It is especially useful for identifying peer-reviewed articles on medical studies, clinical trials, and epidemiology.

Search Query for PubMed

(((((data*) OR (variable*)) AND ((diabetes*) OR (diabetes insipidus) OR (diabetes mellitus) OR (polygenic disease) OR (polygenic disorder)) AND ((AI) OR (artificial intelligence) OR (ML) OR (machine learning) OR (deep learning)) AND ((predict*) OR (detect*) OR (identify*) OR (discover*) OR (find*) OR (recogniz*) OR (determin*) OR (anticipat*) OR (project*) OR (estimat*)) AND ((train*) OR (validat*) OR (metric*) OR (evaluat*))))

By effectively using these databases, we can access the most relevant and comprehensive information for systematic reviews in ML applications for diabetes prediction. Each database offers unique tools and collections that can significantly enhance the depth and breadth of literature reviews.

Search Query for IEEE Xplore

((((("data" OR "dataset" OR "variable*") AND ("diabetes" OR "diabetes insipidus" OR "diabetes mellitus" OR "polygenic disease" OR "polygenic disorder") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "ML" OR "DL") AND ("predict*" OR "detect*" OR "identif*" OR "discover*" OR "find*" OR "recogni*" OR "anticipat*") AND ("training" OR "validating" OR "validation" OR "matric*" OR "evaluate" OR "evaluating" OR "evaluation" OR "examine" OR "examining" OR "examination"))))

The PubMed search query focuses on artificial intelligence and machine learning in diabetes research, specifically on data forms and variables. It includes terms for deep learning, prediction, detection, identification, and estimation of diabetes-related aspects. The query also included terms on the methodologies used, ensuring comprehensive discussions on the effectiveness of ML models in diabetes prediction and management.

A critical resource for technology-focused research, **IEEE Xplore**, provides access to content from the Institute of Electrical and Electronics Engineers (IEEE) and the Institute of Engineering and Technology (IET). It includes over four million documents, including articles, conference papers, and standards, essential for research involving technological applications in healthcare, such as AI and ML algorithms, software, and system implementation.

The search query for IEEE Xplore includes various terms related to diabetes from abstracts of papers, such as AI, ML, machine learning, and deep learning, with the aim of predicting, detecting, discovering, finding, recognizing, determining, anticipating, projecting, evaluating, and training. Elsevier owns the **ScienceDirect** and offers various scientific and technical research covering the physical sciences, engineering, life sciences, health sciences, social sciences, and humanities. This database is valuable for comprehensive searches in interdisciplinary fields that combine technology and healthcare, and provides access to a vast library of scientific articles, book chapters, and other resources.

Search Query for Science Direct

((data*) OR (variable*)) AND ((diabetes*) OR (diabetes insipidus) OR (diabetes mellitus) OR (polygenic disease) OR (polygenic disorder)) AND ((AI) OR (artificial intelligence) OR (ML) OR (machine learning) OR (deep learning)) AND ((predict*) OR (detect*) OR (identif*) OR (discover*) OR (find*) OR (recogniz*) OR (detremin*) OR (anticipat*) OR (project*) OR (estimat*)) AND ((train*) OR (validat*) OR (metric*) OR (evaluat*)))))))

The Science Direct search query focuses on machine learning (ML) in diabetes research, specifically predictive and diagnostic models. It included keywords related to data handling and variables, ensuring relevance to diabetes and its genetic interactions. This query highlights innovative methods, functional objectives, and methodological aspects, providing insights into current trends, challenges, and advancements in the field.

Owing to the restriction of using only eight Boolean operators in the Science Direct database, we split the search query into five subqueries to search all relevant articles. Subsequently, we combined all the articles searched from other databases as well, filtered out unique articles and removed duplicates.

Search Query 1 for Science Direct

((diabet) OR (diabetes insipidus) OR (diabetes mellitus) OR (polygenic disease) OR (polygenic disorder)) AND ((data) OR (variable))

Search Query 2 for Science Direct

((diabet) OR (diabetes insipidus) OR (diabetes mellitus) OR (polygenic disease) OR (polygenic disorder)) AND ((AI) OR (artificial intelligence) OR (ML) OR (machine learning) OR (deep learning))

Search Query 3 for Science Direct

((diabet) OR (diabetes insipidus) OR (diabetes mellitus) OR (polygenic disease) OR (polygenic disorder)) AND ((predict) OR (detect) OR (identify) OR (discover) OR (find))

Search Query 4 for Science Direct

((diabet) OR (diabetes insipidus) OR (diabetes mellitus) OR (polygenic disease) OR (polygenic disorder)) AND ((recognize) OR (determine) OR (anticipate) OR (project) OR (estimate))

Search Query 5 for Science Direct

((diabet) OR (diabetes insipidus) OR (diabetes mellitus) OR (polygenic disease) OR (polygenic disorder)) AND ((train) OR (validate) OR (metric) OR (evaluate))

By effectively using these databases and their search queries, a comprehensive search strategy was crafted to retrieve relevant literature, and we were able to access the most relevant and comprehensive information for systematic reviews of ML applications for diabetes prediction from 2014 to 2023.

3.3. Inclusion and exclusion criteria

The exclusion and inclusion criteria can facilitate the selection of resources that address the research questions in a systematic literature review. Within the framework of our investigation, we determined and implemented the "Inclusion/Exclusion" criteria that should be followed. For the **Exclusion Criteria** during our research, we eliminated the resources that satisfied the following constraints:

- Articles written in languages other than English.
- Short papers are defined as papers that consist of fewer than seven pages.
- Workshop papers
- Papers that are duplicated.
- The full text of the papers that were not available for reading.
- In subsequent years, conference papers were published in journals.

For the **Inclusion Criteria**, all articles that applied machine learning methods to predict diabetes were included in our study to further analyze and extract data for answering the research questions and achieving the defined objectives.

3.4. Execution of search queries

Once we had the general framework for the SLR in hand, we designed a thorough search strategy to cover all databases and widen the scope of our search.

Step A. The search yielded many relevant articles from three major databases. Consequently, the research data were collected from 321 articles from IEEE Xplore, 728 papers from PubMed, and 807 articles from Science Direct. These diverse sources were useful in building a good pool for the review, which will be useful in the development of the database. The first search yielded 1856 articles in all the databases searched. For the records that were collected, the process of deduplication was also performed to avoid having the same record entered twice. Finally, after excluding duplicate articles 336, there were 1520 articles underwent the screening process.

Step B. Each manuscript was then subjected to exclusion criteria in a stepwise manner. In this phase, all records identified by the search process amounted to 1520, and all records were screened by title and abstract to determine their suitability for the study. Of these, 1468 records were screened because they were not relevant to the research questions or did not meet the inclusion criteria. Of these, 53 were potentially relevant and were retrieved for a full-text review based on the title and abstract.

Step C. The first author of the study systematically reviewed 53 manuscripts and strictly obeyed the inclusion criteria. Thus, 37 studies were included in the analysis after the full-text review of the articles and according to the data quality, relevance of the studies, and purpose of this study. Out of the total 37 studies, a total of 16 studies were removed at this step; 11 studies were not related to the research questions, and for five studies, the required information was missing. The remaining 37 studies were considered to be of high quality and more closely related to the systematic review.

Step D. To make the process of identifying relevant articles even more rigorous, the snowballing technique was used. The citation searching method refers to the use of references or citations from previous studies that have been included in the current study and helps in identifying more related studies that may have been retrieved from the database search. There are two types of snowballing: forward snowballing, which entails identifying papers that have cited the included papers; and backward snowballing, which involves identifying papers cited in the included papers. Backward snowballing was applied to ensure a systematic review of the reference lists of the 37 studies. Using this snowballing method, other 16 related studies were found, and all of them were included in the final review. These studies greatly expanded the range of the literature and greatly reduced the likelihood of missing pertinent research.

Step E. The last step is the process of incorporating the studies. By the end of the study, 53 papers were analyzed in the systematic review after the eligibility step, and the snowballing technique was performed after obtaining 37 papers from the eligibility step and an additional 16 papers from the snowballing technique. This approach ensured a consistent and systematic method of selecting studies, as shown in Fig. 1, which provides a solid basis for answering the research questions and yielding useful knowledge. A detailed summary of the 53 reviewed studies, including dataset sources, machine learning algorithms, training strategies, evaluation metrics, and key findings, is provided in Supplementary Material (Table 1).

Step F. We progressed to the data extraction stage, which is crucial for answering our study questions, by identifying the exact datasets and their characteristics, training strategies, evaluation approaches, metrics, and ML algorithms used in the studies. The data collection process was simple, enabling the primary author to handle the extraction independently. However, assessing the possible constraints of these investigations is more difficult. This analysis required a thorough and concentrated discussion, collaboratively carried out by all authors of our work. They carefully analyzed the sections of publications that addressed potential constraints and threats to validity. They examined the features and qualities of each ML technique employed to pinpoint further constraints. All authors have experience in artificial intelligence and machine learning, with years of expertise and engagement in teaching academic courses. This significantly improved the thoroughness and depth of the analysis in this phase of the systematic literature evaluation.

Information obtained from the selected publications answered the research questions. This section offers a concise review of the most important findings of our investigation.

3.5. Quality assessment

Before moving on to the process of extracting the material necessary to answer our research questions, we evaluated the quality and comprehensiveness of the collected resources. Papers that did not provide sufficient details to be utilized in our investigation were discarded. We devised a checklist containing the following queries:

Q-1: Are there datasets used related to diabetes prediction?

Q-2: Are Machine Learning techniques for diabetes prediction clearly defined?

Q-3: Are there training and validation strategies used for the models?

Q-4: Are there any metrics used to evaluate models for diabetes prediction?

There are three possible responses to each question: **"Yes," "Partially,"** and **"No."** We assigned a numerical value to each label in order to evaluate the quality and comprehensiveness of each source. For example, the label **"Yes"** was assigned the value **"1," "Partially"** was assigned the value **"0.5,"** and **"No"** was assigned the value **"0"** The overall quality score was determined by adding the scores of the responses to the two questions, and the articles with a quality score of at least one was accepted for publication.

Therefore, all 53 papers that underwent previous rounds of evaluation also passed the quality assessment test. No paper was omitted in this stage because all papers were found to have reached the minimum quality score for the next stage of analysis. This phase confirmed that the last set of studies was both complete and of quality for the systematic review, thus guaranteeing that the included studies would provide

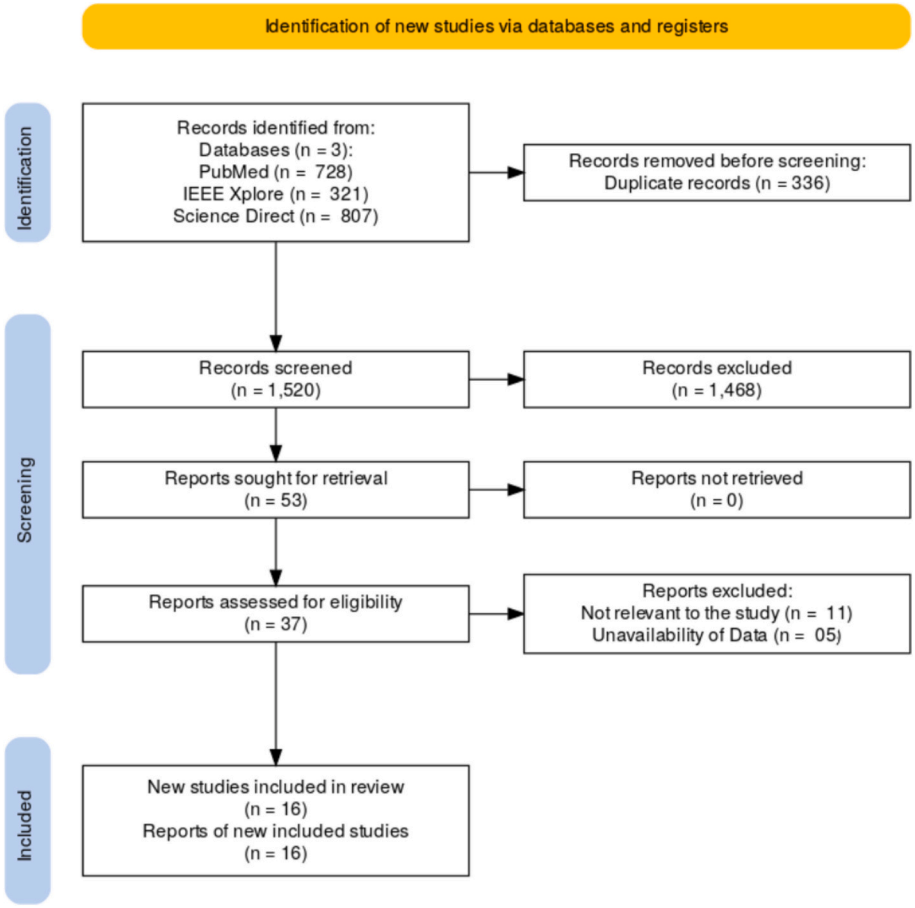


Fig. 1. PRISMA flowchart of study identification, screening, and inclusion process.

Table 1
Summary of extracted articles from various databases.

Database	Period	Document type	Publication stage	Language	Media format	Subject of interest	No. of papers
IEEE Xplore	2014–23	Conference Proceedings and Journals	Final and published	ENG	PDF	Computer Science/Engineering	321
PubMed	2014–23	Conference Proceedings and Journals	Final and published	ENG	PDF	Computer Science/Engineering	728
ScienceDirect	2014–23	Conference Proceedings and Journals	Final and published	ENG	PDF	Computer Science/Engineering	807
Total							1856

meaningful insights and reliable information to the research objectives.

3.6. Data extraction

As a part of the research on ML models for diabetes prediction, researchers carefully choose the datasets suitable for training the models, emphasizing their relevance and representativeness [31]. Classifiers usually selected to deal with medical data interpretation issues have been developed to cope with the specific difficulties of medical data classification. The training process of ML models often consists of a detailed scheme that can include cross-validation to ensure that the results are accurate and not over fitted.

Validations were performed using test sets of specific data to estimate the generalization of the model. Essential metrics, such as accuracy, sensitivity, specificity, and AUC, measure the performance of the model [32]. The choice of independent variables used in training is of paramount importance, and could be demographic, biochemical, or clinical factors related to the risk of diabetes. Nevertheless, the research is confronted with restrictions, such as biased datasets, variability of data quality, and generalization of the results for other communities. These constraints highlight the need for continuous research and overall improvement of ML models in the medical domain. Once we determined the specific group of sources to be considered, we retrieved the information to answer our research questions. The first step was to specify the data extraction displayed in Table 2. We also sought to extract data from the datasets used and the training and validations used to develop the technique. These facts could help to enhance the picture of the chosen features of the papers. In addition to the fundamental information on

Table 2
Data extraction form.

Dimension	Attribute description		
Datasets	Which datasets were chosen to train the models of ML? Classification Types	Which classification algorithms were selected for diabetes prediction? Training Strategy	What strategy was followed by the ML models for training?
Independent variables	Which independent variables were selected during the model training? Validation	What type of validations were performed for the evaluation of the ML model? Evaluation Metrics	Which metrics were considered for the evaluation of the ML model?

diabetes prediction topics discussed in the article or the prediction techniques of ML, we also sought to extract data from the datasets. In addition, the data extraction sheet allowed us to extract "*Limitation(s)*" identified for the reviewed research methodologies.

4. Results and discussion

Before embarking on the analysis of findings from our systematic literature review, it is prudent to anticipate a systematic approach that allows for the accurate interpretation and synthesis of the collected data. This preparatory step requires careful classification of all collected articles based on criteria, such as study design, methodological approaches, ML applications, and effect measures. It is an efficient method of organizing information, which not only helps in the analysis process, but also increases the accuracy and reliability of the results. Second, we explain the specific methods applied in the qualitative and quantitative analyses, enabling a meaningful comparison of results across multiple studies. This will help us provide a comprehensive analysis of the issues discussed in the review and address the research questions and objectives established in the first stages of the review. Fig. 2 shows the fluctuating interest in diabetes prediction research from 2014 to 2023. From 2014, there was a surge in publications, peaking at 8 in 2017. The highest number was in 2021, likely due to technological advancements and the COVID-19 pandemic situation. However, the decline in 2022 and 2023 suggests a need for further research.

Journals play a vital role in disseminating research done by people. In the context of diabetes prediction research from 2014 to 2023, the top five journals contributing significantly to diabetes prediction research were Diabetes Care, IEEE Access and IEEE Transactions on Biomedical Engineering. Scientific Reports and the Journal of Diabetes Science and Technology closely followed, as shown in Fig. 3, accounting for 4 % and 4 %, respectively. This interdisciplinary approach highlights the growing use of advanced computational methods and data analytics for the prediction of diabetes.

From our study, it is clear that research on diabetes prediction focuses on critical components such as diabetes, model, data, and machine learning, as shown in Fig. 4. The highest frequency of these keywords highlights the importance of data-driven models and machine-learning techniques for enhancing prediction. This study emphasizes the need for accurate predictions to mitigate the risks associated with diabetes, utilizing clinical insights and algorithmic advancements.

4.1. RQ_1: on the datasets and their characteristics

For diabetes prediction research, dataset selection is crucial [33].

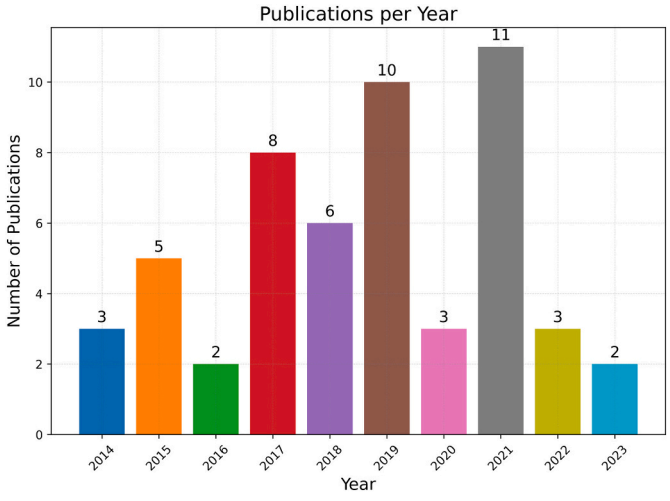


Fig. 2. Distribution of publication by year.

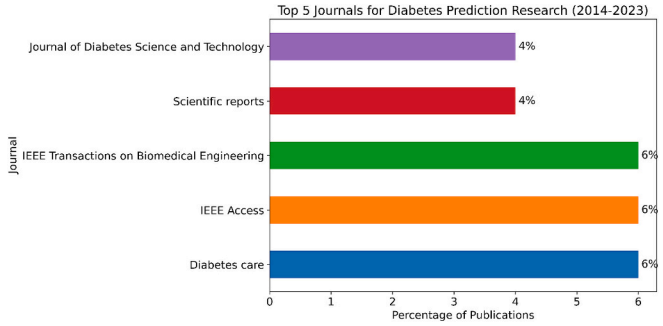


Fig. 3. Top 5 % journals for diabetes prediction research (2014–23).

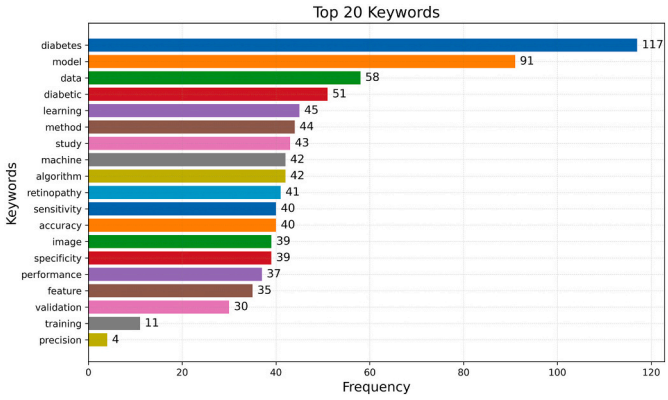


Fig. 4. Top 20 keywords used in the studies.

Data or datasets are the key inputs to the development of any predictive model, and the quality of the data defines the efficacy of the resulting model. Advanced datasets include people with different demographics, diseases, and geography, which provide extensive information for diabetes control and prognosis. From the different datasets, one can see that the way diabetes is researched and combined varies in terms of methods and strategies. In response to the $\mathbf{RQ}_1$, the current studies revealed that the employed data embrace different populations and geographic areas, which offers a comprehensive view of diabetes research from different perspectives. Both datasets require different features that represent the richness and complexity of diabetes management and prediction. The different distributions of the datasets used in the studies are shown in Table 3.

4.1.1. Multiethnic and population-based datasets

These datasets provide diverse demographic and clinical data, enabling models to generalize across different populations.

4.1.1.1. Singapore National Diabetic Retinopathy Screening Program dataset. The retinal images and health records of a large multi-ethnic population in Singapore constitute an excellent resource for diabetic retinopathy (DR) detection since this dataset includes retina images. This dataset contains graded DR severity levels and can be used to develop deep learning models for the automated screening of DR. Modeling should be performed on a wide range of DR presentations via dataset diversity, which ensures that the models are more applicable to real-world settings [34].

4.1.1.2. AusDiab Dataset. AusDiab is a large study involving over 11,000 people in Australia conducted as a cross-sectional and longitudinal study; back of the newspaper ad. This dataset contained both diabetes risk factors (fasting glucose, HbA1c, obesity, and lifestyle) and lifestyle factors (work status, exercise, alcohol consumption, smoking,

Table 3
Distribution of datasets used in the literature.

Dataset	Sample size	Class ratio (diabetic: non-diabetic)	Missing data %	Geographic population	Challenges faced
NHANES [28]	10,000 per cycle	1:9	10 % (Lifestyle data)	United States	Class imbalance, inconsistent data collection
PIDD [38]	768	1:1.5	8 % (Insulin)	Indian Population	Small sample size, ethnic/gender bias, limited generalizability
Optum® EHR [20]	95 M+ records	1:7	20 % (Lab results)	United States	Missing data, variations in medical coding, lack of standardization
EyePACS [39]	88,702 images	1:3 (DR severity)	5 % (Labels)	Global dataset with ethnic variations	Ethnic representation gaps, dataset skewed toward certain populations
REPLACE-BG [40]	226 participants	N/A (Type 1 diabetes only)	12 % (BG data)	United States	Small dataset size, limited to Type 1 diabetes patients
Aizawa Hospital Study [41]	11,247	1:3.5	15 % (Lab data)	Japan	Single-center data, limited external validity
AusDiab [42]	11,247	1:8	25 % (Self-reported)	Australia	Self-reported data bias, missing lifestyle metrics
Singapore DR Screening [34]	100,000+	1:3.5	7 % (Severity labels)	Singapore	Variability in screening
criteria, class imbalance MESSIDOR [43]	1200 images	1:2 (DR severity)	5 % (Labels)	France	Small dataset size, limited generalizability
Botnia Prospective Study [44]	4389	1:2.8	10 % (Follow-up data)	Finland/Sweden	Longitudinal costs, attrition bias
KNHANES [36]	20,000	1:7	12 % (Blood sugar levels)	South Korea	Variability in diagnostic criteria, potential class imbalance
Humedica [45]	32 M+ records	1:6	25 % (Biomarkers)	United States	EHR inconsistencies, missing values in key biomarkers
Practice Fusion EHR [46]	1.2 M records	1:5	30 % (Clinical notes)	United States	Inconsistent documentation, unstructured clinical notes
DPDS [47]	1000+ cases	1:1.2	15 % (Metabolic data)	Multi-country dataset	Limited to specific clinical environments, potential selection bias
Itabuna Diabetes Campaign [48]	5000+	1:1.8	30 % (Self-reported)	Brazil	Self-reported diabetes status, lack of medical validation
UWF-SLO Retinal 250,000+ images Dataset [49]	1:1.5 (DR severity)	8 % (Image quality)	Global dataset	Image quality variations, need for advanced preprocessing	Data Quality Issues
BARICAN Cohort [50]	10,000+	1:4	18 % (Progression data)	Barbados/Caribbean	Limited regional applicability, sample diversity constraints

diabetes medications, and supplements). For example, it is important to understand the impact of lifestyle factors on the development of diabetes and its complications. Nevertheless, this dataset has a limitation in the use of self-researched lifestyle data that can lead to bias and inaccuracy in the analysis [35].

4.1.1.3. National Health and Nutrition Examination Survey. It was collected by Centers for Disease Control and Prevention Search in the U. S. is a nationally representative dataset that includes demographics, laboratory results (including HbA1c and fasting glucose), health history (including BMI, blood pressure, cholesterol), etc. It is most commonly used to predict diabetes risk and to separately stratify patients according to their metabolic health indicators [36,37].

4.1.1.4. Korean National Health and Nutrition Examination Survey. The Korean National Health and Nutrition Examination Survey (KNHANES) is similar to NHANES, but the available data in it includes biometrics, lifestyle, and clinical data including HbA1c, fasting glucose, and insulin. In particular, it is a useful predictor of diabetes risk in East Asian populations, where genetic and lifestyle factors may differ from those in the West [48].

4.1.2. Advanced monitoring and glycemic control datasets

Continuous glucose monitoring (CGM) and glycemic control are the datasets of focus, which allow real-time management of diabetes.

4.1.2.1. REPLACE-BG Dataset. This dataset, which is designed to compare CGM with blood glucose testing (BGT) provides glycemic indices, insulin therapy data as well as glucose time in range (TIR). Additionally, it is utilized to optimize blood sugar control and hypoglycemic blood event predictors, making it a valuable asset for AI-based diabetes management systems [51].

4.1.2.2. The Diabetes Prediction Data Set (DPDS). It contains over 1000 points and is used very often for diabetes classification using machine learning. The dataset possesses critical features, such as age, BMI, blood sugar level, and demographic indicators, which makes it an important dataset for diabetes risk prediction. Nevertheless, it does not generalize well to the general population because the data are collected in specific clinical settings [52].

4.1.3. Imaging-based datasets for Diabetic Retinopathy (DR)

The datasets in this work consist of retinal images and are typically used for training deep learning models to detect and classify DR.

4.1.3.1. EyePACS database. Diabetic patients color fundus images: over 22,000 images with DR severity marked. It is extensively used in deep-learning-based screening models, particularly for training CNNs [26,53].

4.1.3.2. Messidor dataset. A retinal fundus imaging dataset with 1200 images collected from three ophthalmology clinics in France. Used for benchmarking DR classification models, MESSIDOR provides gold-standard labeled images for training and validating AI models in ophthalmology [29,54].

4.1.3.3. Kaggle Diabetic Retinopathy Dataset. A dataset containing 35,000+ retinal images sourced from the Kaggle DR Challenge. It provided high-quality DR annotations across multiple severity grades. It is widely used in convolutional neural network (CNN) training for DR detection [55,56].

4.1.3.4. Ultra-wide field Scanning Laser Ophthalmoscopy (UWF-SLO) dataset. Retinal images dataset of 9392 retinal images with retinal longitudinal tracking data for 10 years globally. It has been applied to the evaluation of AI-based DR progression models and to increase diagnostic accuracy [48,49].

4.1.4. Electronic Health Record (EHR) datasets

Longitudinal studies and predictive modeling are possible with EHR datasets due to data from comprehensive patient records.

4.1.4.1. Optumő EHR dataset. It is a longitudinal dataset that contains millions of de-identified patient records from U.S. hospitals. It covers demographics, laboratory results, medications, and clinical visit data, and is optimal for determining when diabetes will begin, how well treatment works, and what future complications are expected [20].

4.1.4.2. Humedica database. This dataset contains 24,331 patient records published from 2007 to 2012. It monitors the transition from normoglycemia to prediabetes to diabetes for use in AI-informed risk prediction and management models [36].

4.1.4.3. Practice Fusion EHR dataset. This dataset also makes use of a de-identified EHR dataset of 9948 patients from 20,092,011 for longitudinal studies of diabetes onset prediction and comorbidity analysis [46,57].

4.1.5. Community-based and low-cost screening initiatives

These datasets consider community health and low-cost screening solutions for such communities.

4.1.5.1. Itabuna diabetes campaign dataset. This dataset of 824 diabetic patient records and their associated fundus images is derived from a community-based DR screening initiative in Brazil. Thus, it is employed to design low-cost DR detection solutions that are mobile-friendly [47].

4.1.6. Surgical and longitudinal outcome studies

These are datasets that follow the long-term consequences of interventions like bariatric surgery on diabetes outcomes.

4.1.6.1. BARICAN Cohort. The post-bariatric surgery type 2 diabetes patient dataset is longitudinal and underwent follow-up for 18 months. This can provide insight into glucose metabolism, weight loss trajectories, and long-term diabetes remission [50].

4.1.6.2. Aizawa Hospital dataset. Data for this study included 2105 adults with prediabetes recruited at Matsumoto, Japan who were followed for over 2 years in the Aizawa Hospital Study. This dataset studies changes in medical history, laboratory results, and the course of diabetes to understand early indicators and causes of diabetes onset. Nevertheless, the limited generalizability of the results is because they rely on a single hospital study population [41].

4.1.6.3. Botina Prospective Study dataset. In Botnia Prospective Study (Finland, Sweden), insulin response and glucose duration are focused on in this for extended follow-up. Studying the genetic and metabolic factors that determine diabetes onset using this particular dataset is of great value. Nevertheless, the longitudinal nature of this dataset makes it challenging for follow-up resources and increases study costs, which may not be feasible for large-scale subsequent studies [49].

4.1.7. Widely Used Benchmark datasets

These datasets are widely used for benchmarking and ML model evaluation.

4.1.7.1. Pima Indians Diabetes Database (PIDD). PIDD is one of most frequently used datasets for diabetes prediction with 768 records of female patients: Glucose levels, Insulin, BMI and Diabetes Pedigree Function although some may argue that the labels are not correct. This is a benchmark for classification algorithms [58–60].

Summary of RQ_1

Several data enhance diabetes prediction but have drawbacks, such as quality, data acquisition, demography, and privacy. A substantial number of datasets fail to contain follow-up data and do not use uniform parameters for diabetes. It is necessary to improve the mechanism for creating prediction models using integrated data and more refined algorithms.

4.2. RQ_2: on the configurations of ML algorithms in diabetes prediction

The application of ML techniques has enhanced the prediction of diabetes and reliability of predictions. Owing to their exposure to different datasets, ML systems can handle large amounts of data, discover complex and sophisticated patterns, and enhance the accuracy of the outcomes. These algorithms involve independent variables and training strategies. Training processes allow for tweaking and verifying the ML models, while other factors outside the training process, such as the demographic data of the patients and the medical statistics, provide input data. Diabetes can also be predicted with the help of machine learning algorithms to analyze large datasets and patterns that are not statistically significant. The common ML algorithms used for prediction/classification purposes are Decision Tree Classifier, Naive Bayes, Linear Regression, Logistic Regression, K-Nearest Neighbor, CNN, SVM, and XGBoost are suitable for organizing various sorts of data and providing accurate predictions or classifications [61]. These algorithms employ complex data inputs, such as medical images or physiological readings, to enhance diabetes diagnosis and care.

Diabetes prediction models require independent variables as the inputs for the algorithm training, some of these variables include age, gender, blood glucose levels, and retinal images. The accuracy of diabetes prediction is influenced by independent variables [62]. Choosing appropriate and full variables is useful for creating predictions using ML algorithms. Training methods are essential for enhancing ML models. Cross-validation, data augmentation, hyperparameter tuning, and feature selection improve models and avoid overfitting. Cross-validation was used to validate the model on different data subsets to increase its reliability. The augmentation of the data enhances the generalization of the model because the training data contain variations in the new dataset. Feature selection helps in removing irrelevant features, thereby reducing noise in the data and improving overall performance of the models. It simplifies the model by eliminating redundant variables, which also enhances computational efficiency. Additionally, it contributes to better generalization and interpretability, especially in clinical applications where understanding feature importance is more important.

4.2.1. The role of independent variables and training techniques in diabetic prediction using ML algorithms

Some studies aimed at the diagnosis of diseases such as diabetic retinopathy, possibly glaucoma, and AMD, in which retinal imaging was the main independent factor. The type of ML algorithm that has been mainly employed for these tasks is the CNN, which is used for image analysis and classification. The training plan relies on providing numerous retinal images to the deep-learning system. For instance, research based on the Singapore National Diabetic Retinopathy Screening Program and other multi-ethnic population-based studies

further optimized their classifiers with large databases of retinal images to achieve high levels of classification accuracy for such diseases [34,35,41,63]. In the REPLACE-BG studies, the principal independent variables were various glycemic indices, such as mean blood glucose level and time in range. Classification is often performed using SVM because it is efficient for small sample sizes and avoids the problem of over-learning. The training strategy included data partitioning into the training and test sets and feature selection by recursive feature elimination.

This method allows for the appropriate utilization of features that enhance the model performance [51].

A wide range of demographic and health-related predictors were included in the present study, and data were obtained from the NHANES dataset. Most of the statistical tests applied were logistic regressions. In the classification, emphasis was placed on biomarkers that could help to distinguish between prediabetes and DM. The training strategies employed comprised five-fold cross-validation to prevent overtraining of the models; large demographic and health data could then be used to accurately predict diabetes risk [26,37].

The dataset concerning the Itabuna Diabetes Campaign involved determining the DR severity from fundus images using deep CNNs, such as PhelcomNet. The training process also involved some augmentation, where the images were rotated and brightness was changed to obtain the best results. These studies were expected to enhance the diagnostic abilities of CNNs on a large dataset of fundus images [14,64].

The features that were detected when working with studies that employed the Optumő EHR Dataset; XGBoost, were commonly utilized because of its capacity to accommodate big data. The training activities included feature selection and hyperparameter tuning using five-fold cross-validation. This approach helped deal with the large EHR data that resulted in the prediction of diabetes and related diseases [30,45,52,60,65,66].

The EyePACS mainly consisted of fundus images, and the CNN was used for diagnosing and categorizing DR. Training practices included data augmentation and cross validation, which allowed the model to recognize different DR stages across the population [53].

Studies based on data from ELSA-Brasil used random forest algorithms because they are designed to work with high dimensions and also give the probability of variable importance. The training included parameter optimization and selection of the most appropriate variables with the help of wrapper methods and demographic and clinical factors to predict diabetes risk [67].

In the Botnia Prospective Study, regularized least-squares regression was used to predict the risk of type 2 diabetes. The training strategy used in this study increased the model generalization and predictive accuracy, including multivariate logistic regression and repeated nested cross-validation [50]. The models were validated with the VA Puget Sound Health Care System dataset in terms of sensitivity and specificity to DR detection with FDA (Food and Drug Administration (FDA)-approved models. The training strategy was to use models trained on other datasets and tested on this dataset without retraining to prove the transferability and robustness of the ML algorithms [63,68].

Based on the NHANES data of Korea, prediabetes was predicted based on fasting plasma glucose levels. The machine learning approaches, ANN and SVM, first applied the grid search and then used 10-fold cross-validation to fit the models and correctly recognize the prediabetes within the population [48].

Consequently, based on the visual analysis depicted in Fig. 5, we are provided with a rather obvious realization that the age variable is used as the independent variable in the vast majority of studies to a significant extent. In the process of diabetes forecasting, there are some factors that include, but are not limited to, body mass index, blood pressure, glucose level, cholesterol level, insulin level, family history of diabetes, physical activity, and diet patterns. Moreover, some researchers have pointed out that in the training procedure, researchers must pay attention to and choose independent variables to improve the accuracy of the

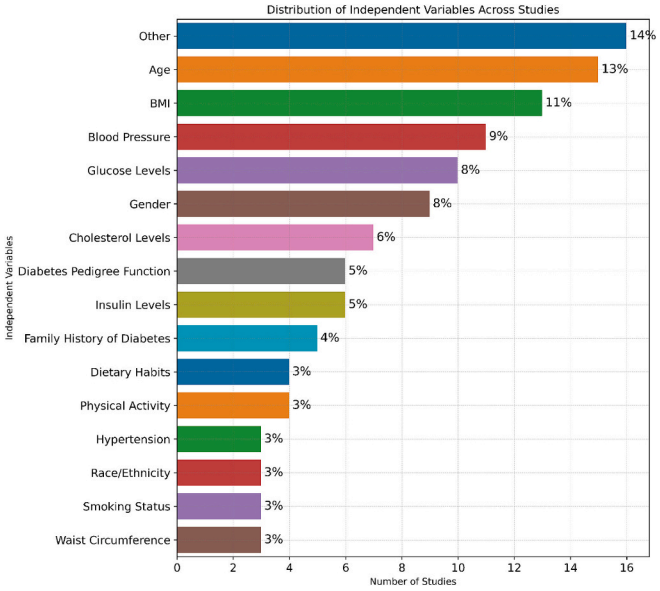


Fig. 5. Distribution of independent variables across studies.

forecast. Some of the sub-indices that are grouped under the other category depicted in Fig. 5 include: environmental factors, social factors, pharmaceutical use and the likes.

4.2.2. Classification types and corresponding ML algorithms for diabetes prediction

CNNs are mainly applied for image analysis in experiments, and the most common types of images used are retinal images for diagnosing diabetic retinopathy, suspected glaucoma, and AMD. Training was performed using a large number of retinal images, which provided high-accuracy models [34,35,41,63].

In analyzing glycemic metrics, Support Vector Machines (SVMs) were used to analyze small data points with the added advantage of avoiding overfitting. Filtering is the most commonly used technique for adjusting the feature list and extracting the best set of features to be applied in model [51]. The motivation for applying logistic regression to the NHANES dataset is based on the previous application of the method in research on factors such as BMI percentiles and family history of diabetes and hypertension. The five-fold cross-validation process is a common training methodology to obtain high model stability [26,37].

Deep CNNs were applied in the Itabuna Diabetes Campaign dataset to classify DR severity, including data augmentation to increase the transferability of the model [47,53].

Specifically, XGBoost has often been adopted in research with the Optumő EHR dataset, which is characterized by high performance and data compatibility. The training methods used were feature selection pre-processing and fivefold cross-validation [30,45,52,60,65,66,69].

In studies that employed ELSA-Brasil data, random forest. algorithms were incorporated. These algorithms were selected based on their capacity to work with a large number of predictors and provide a quantitative evaluation of the importance of the variables [67].

In the Botnia Prospective Study, least-squares regression analysis was used to determine the risk of type 2 diabetes related to the metabolomic profiles. The training methodologies employed were multivariate logistic regression and repeated nested cross-validation [48,50,52].

For the HRV signals obtained from ECG, CNN-LSTMSVM Hybrid Models were utilized in the research, as component algorithms complement each other, improving the predictive potential. [27,46,54,55,59,64,67,68,70–75].

The deep learning model Inception-V3, which is ideal for image data, was applied in a study that analyzed OCT measurements of diabetic patients [56].

Demographic, clinical, and lifestyle factors were used and employed with Bayesian networks because this approach is suitable for representing probabilistic dependencies between variables [29,48,49,76–84].

Traditional ML and deep learning algorithms are applied in half of the studies on diabetes prediction, thus proving that they play equal and significant roles. Logistic regression and decision trees are basic approaches to machine learning, whereas CNN is suitable for big data. Ensemble Learning, which utilizes more models to enhance performance, constitutes 6 % of the approaches. Evolutionary Computing is less than 2 %, and Bayesian Inference is less than 2 %, indicating that several methods are used to enhance the accuracy and complexity of diabetes prediction. Fig. 6 shows how ML algorithms were used in diabetes prediction, that is, the proportion of studies that used each of these methods in a systematic literature review (SLR). This shows the most important approaches adopted in the analyzed studies, namely basic ML algorithms (such as Logistic Regression, SVM) and deep learning models (such as CNNs, LSTMs, and XGBoost). It mostly shows commonly used methods, but conveys insights into the evolution of AI in diabetes prediction. The future is to develop more complicated AI models that take advantage of both deep-learning techniques and hybrid AI models. Ensemble learning, evolutionary algorithms, and Bayesian inference methods are less commonly used in literature. However, their presence suggests ongoing interest in diversifying the minds of AI. It also supports the focus of this study on the advancement of AI by showing methodological trends in diabetes prediction and the growing role of deep learning, hybrid models, and explainable AI (XAI) in predictive healthcare.

Owing to clinical interpretability and trust in AI-driven decisions, Explainable AI (XAI) methods are increasingly being integrated into diabetes predictions [32]. In CNN-based diabetic retinopathy image detection, gradient-weighted Class Activation Mapping (Grad-CAM) is commonly used to identify image regions that affect model predictions, thereby assisting their clinical validation [23]. SHAP and LIME were also used to identify the most potent biomarkers for diabetes risk prediction using tree-based models, including XGBoost and Random Forest [26]. Random Forest models with feature importance scores have been used in studies such as the ELSA-Brasil cohort to determine the association between clinical or demographic factors and the onset of diabetes [28]. CNN-based diabetic retinopathy models augmented with saliency maps and occlusion sensitivity analysis also serve as the gold standard, corroborating the reliability of the model in detecting early stage diabetic retinopathy [79]. Despite these advances have been achieved, the utilization of XAI has been minimal, particularly beyond CNN architectures in deep learning.

Future research should focus on the development of hybrid models that augment the readability and availability of AI powered healthcare systems.

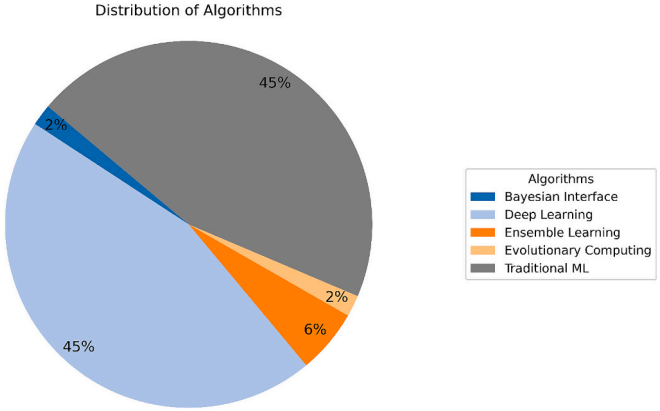


Fig. 6. Percentage of ML Algorithms used for diabetes prediction.

4.2.3. *Comparative analysis of machine learning and deep learning algorithms in diabetes prediction*

Studies of diabetes prediction applications using ML algorithms and deep learning have revealed both strengths and limitations. Models such as Logistic Regression and Support Vector Machines (SVM) demonstrate interpretability features, which make them optimal choices for structured tabular data [62,85]. Logistic Regression achieves moderate performance with an accuracy of 80 %–85 % and an AUC of 0.80–0.85, whereas SVM performs slightly better with an accuracy of 85 %–90 % and an AUC of 0.85–0.90. However, these models experience difficulties in identifying intricate patterns in medical imaging and in analyzing high dimensional datasets.

Deep learning algorithms such as Convolutional Neural Networks (CNN) are used for pattern detection and feature extraction, making them ideal for medical image diagnosis including diabetic retinopathy screening [26,55,70]. As summarized in Table 4, CNNs achieve very high accuracy (95 %–98 %) and AUC (0.95–0.98), outperforming traditional ML models in tasks involving medical imaging. Despite their superior performance, deep learning models require large datasets and robust computing resources, which reduce their operational feasibility in healthcare delivery environments with minimal resources [27,75]. For example, Recurrent Neural Networks (RNN/LSTM), which are effective for sequential data, such as glucose monitoring, are computationally expensive and difficult to interpret, as noted in Table 4.

Future studies should focus on hybrid systems that integrate ML approaches with deep learning methods, leveraging transfer learning features to enhance explainable AI capabilities and improve clinical practice adoption [46,68,74]. Hybrid systems combine the interpretability of ML with the pattern recognition capabilities of DL, achieving high accuracy (90 %–95 %) and AUC (0.90–0.95). However, they face challenges, such as complex implementation and high development costs. The comparative analysis in Table 4 provides a detailed overview of the strengths, limitations, and suitability of these algorithms, offering valuable insights for future research.

Summary of RQ_2

Considering the data type and their appropriateness for certain tasks, the algorithms used for diabetes prediction included CNN, SVM, and XGBoost. Cross-validation, data augmentation, and feature selection help to increase the convergence of predictive models, which can also demonstrate the versatility of the different machine learning frameworks in altering the variables, algorithms, and training paradigms for diabetes prediction research.

4.3. *RQ_3: on the evaluation techniques and metrics*

The evaluation setups are important, especially in the diagnosis of diabetes using ML algorithms. It is important to have evaluation sets to ensure that the models are accurate and practical in real-life situations. These provide a robust approach for comparing ML models; thus, it is easier to identify the most appropriate techniques and increase the reliability and efficiency of the models. The primary finding of this study is that the configurations of ML model evaluations affect the reliability and robustness of the models. Overfitting checks whether the model runs properly on other data and describes its advantages and limitations. Scholars can determine which diabetes prediction models can be selected based on available evaluation methods.

As seen in the literature on diabetes prediction, there are different approaches for measuring performance. One such method is cross-validation, which is helpful in enhancing the assessment of the model

Table 4
Comparative analysis of machine learning and deep learning algorithms in diabetes prediction.

Algorithm	Performance metrics	Strengths	Limitations	Suitability	Challenges faced
Logistic Regression	Accuracy: 80 %85 %, AUC: 0.800.85, Sensitivity/ Specificity: Moderate	Highly Interpretable, Efficient for structured data	Struggles with complex Patterns in medical imaging, Limited in high-dimensional datasets	Ideal for Structured data (e.g., patient demographics, lab results)	Limited to simple Patterns, requires feature engineering
Support Vector Machines (SVM)	Accuracy: 85 %90 %, AUC: 0.850.90, Sensitivity/ Specificity: High	Effective for structured data, Interpretable	Computationally intensive, Limited for complex patterns	Small to medium datasets, Structured data	Scalability issues, Requires careful tuning
Random Forest	Accuracy: 85 %90 %, AUC: 0.850.90, Sensitivity/ Specificity: High	Handles high-dimensional data, Robust to overfitting	Less interpretable, Computationally expensive	Structured data with many features (e.g., EHR data)	Memory-intensive, Limited interpretability
Convolutional Neural Networks (CNN)	Accuracy: 95 %98 %, AUC: 0.950.98, Sensitivity/ Specificity: Very High	Superior in pattern detection, Ideal for medical imaging	Requires large datasets, Demands robust computing	Medical imaging tasks (e. g., retinal scans, OCT images)	Resource-intensive, Limited interpretability
Recurrent Neural Networks (RNN/ LSTM)	Accuracy: 90 %95 %, AUC: 0.900.95, Sensitivity/ Specificity: High	Effective for sequential data (e.g., time-series data)	Computationally expensive, Requires large datasets	Time-series data (e.g., glucose monitoring)	Difficult to interpret, High training costs
Hybrid Systems	Accuracy: 90 %95 %, AUC: 0.900.95, Sensitivity/ Specificity: High	Combines ML interpretability and DL pattern recognition	Complex implementation, Requires careful tuning	Tasks requiring structured data and complex patterns	Integration challenges, High development costs

by checking its performance on different partitions of data. This approach makes the model more reliable and accurate when tested on different datasets, thereby providing more reliable evaluations [37,51].

Validation tests the model on data other than training data. In the validation process, it is important to observe the generalization of the model to new data. In the Optumő EHR dataset, the model was externally validated by comparing the prediction with the scores of new images from the screening program, as well as ten other datasets with different populations [36]. Bootstrap sampling involves taking a random number of samples from the dataset, using this sample many times to train the model, and obtaining an empirical distribution of the performance measures to analyze model variability.

It is essential to know the types of assessment criteria because they help to quantify the performance of ML models. These include accuracy, Area Under the Curve (AUC), sensitivity, specificity, precision, and F1 score, which provide a detailed performance view of the models. These metrics assist in determining models that not only forecast diabetes with high accuracy but also approach the issue of false positive and false negative cases, which are costly in practice [37,51].

The methods used in the evaluation setup in diabetes prediction research are elaborated to ensure that the performance of the models is tested comprehensively using different approaches and measures.

4.3.1. *Validation methods*

At high reliability, the data sets are divided into portions and the most common technique used is the k-fold cross validation. This technique divides the dataset into k portions and constructs a model k times. The validation dataset was one of the k portions, whereas the training dataset contained the remaining portions of the dataset. For instance, the authors of the research conducted on the REPLACE-BG dataset applied 10-fold cross validation to check the efficiency of the SVM model and to avoid obtaining performance indicators that would not reflect the performance of the model [51]. Similarly, research that used NHANES data used five-fold cross-validation to verify the overall accuracy of the logistic regression model; however, this method requires more training data [37]. This method is vital in preventing overfitting, whereby the model provides excellent results on the training data but poor results on the test data, providing a better evaluation of the model.

External validation was performed using a different dataset than that used to train the models to infer the generality of the models. It is helpful to evaluate the model in conditions closer to the real world, because cross-validation does not show all aspects of performance. For instance, in a study that relied on the Optumő EHR dataset, external validation was conducted by comparing the predictions with the scores assigned to new images from the screening program and 10 datasets comprising other populations [36]. This type of validation means that the model is applicable for making predictions under different populations and conditions, or in other words, in different clinical settings.

Bootstrap sampling has been adopted in some studies to assess the extent of fluctuations in model performance markers. In this method, a dataset is applied such that it samples randomly and successively with replacement, and feeds these samples to the model until the empirical distribution of the performance measure is obtained. Similarly, a study that was conducted on the Optumő EHR dataset to identify genetic variants associated with diabetic ketoacidosis (DKA) also used 1000 bootstrap samples to calculate the 95 % confidence interval for all the aforementioned performance parameters, thus confirming the authenticity of the statistical measures used in the study [36]. This technique enables analysts to identify various levels of variability in the model and the stability of the model in the sampling distribution.

4.3.2. *Evaluation metrics*

Precision is one of the most broad and the most often used measures in the analysis of investigations to describe the proportion of the actual positives and actual negatives regarding the total number of the investigated cases. For instance, while validating logistic regression models, the measure used was the accuracy rate obtained in NHANES-based studies [37]. Other studies also considered accuracy as the criterion for the performance of their models, with specific distribution percentages for each [30,36,49,51,54,63,67,68,76,77].

The AUC is relevant when comparing binary classifiers, as it offers information on the ability of the classifier to distinguish between the two classes. It measures the percentage or rate at which it is appropriate to categorize the positive and negative samples. The AUCs obtained were high, indicating the adequate diagnostic capability of the models for DR using the EyePACS dataset. This metric allows a comparison of the true positive rate (sensitivity) with the false positive rate and obtains a general performance fig. [37,48,50,53,56,60,66,83].

The sensitivity, true positive rate, specificity, and true negative rate indicate the capability of the model to identify positive and negative cases, respectively. Sensitivity measures the true positive rate, and specificity provides the ability to identify actual negative cases. For example, a study conducted on data collected from the Itabuna Diabetes Campaign indicated that the sensitivity of the screening model was 97 %. To detect more than mild DR, the model had a sensitivity of 8 % and a specificity of 61 %. 4 % in detecting severe cases, suggesting that the proposed model is capable of raising awareness of severe cases of DR, while simultaneously pointing out the features that require improvement [34,35,41,45–47,58,59,70,71,81,85].

In the case of dealing with data mining in imbalanced datasets, some

of the measures that are considered to be crucial include precision, which is the total number of correct predictions of the positive observation over the total number of positive observations in the dataset, and F1 score, which is a weighted average of both precision and recall. The evaluation of the SVM model in the study with the REPLACE-BG dataset incorporated these measures, where not only true positives were correctly identified, but the proper precision and recall of the model was also achieved [26,51,57,64,69,75,79,80,82,83,86]. These are good metrics, particularly when it is necessary to avoid the position where the model provides both high false positive and false negative values.

The diagnostic accuracy measures are the Positive Predictive Value (PPV) and Negative Predictive Value (NPV), which reveal the proportion of actual positives and actual negatives out of all the cases predicted to be positive or negative. The study that used the data from the VA Puget Sound Health Care System with the purpose of comparing the effectiveness of the screening algorithms developed with the help of AI and PPV and NPV results in a better understanding of the efficiency of the models in actual health care centers [63]. These metrics are useful when a model is applied in a clinical context in which false-positive and false-negative results can have consequences.

In Fig. 7, the evaluation metrics in the diabetes prediction models are distributed, which makes methodological choices. The most common metric is the accuracy, which is a general metric. However, it does not address the problem of class imbalance, which is an important feature in diabetes prediction. Sensitivity and specificity are prominent because balanced classification is the focus, and AUC is used to assess discriminative ability.

Precision and F1-score, which focus on true-positive detection, are generally less frequently reported as metrics. While PPV and NPV are important from a clinical point of view, their utility is almost less used, favoring broad accuracy over case reliability. These trends highlight the necessity of considering several performance indicators for the holistic validation of AI-driven diabetes prediction.

Summary of RQ3

This systematic review validated the methods and metrics used to predict diabetes using machine learning across the spectrum. To enhance the model reliability, cross-validation, external validation, and bootstrap were used, whereas to check the model effectiveness, performance evaluation metrics such as accuracy, AUC, sensitivity, specificity, precision, and F1 measure were employed.

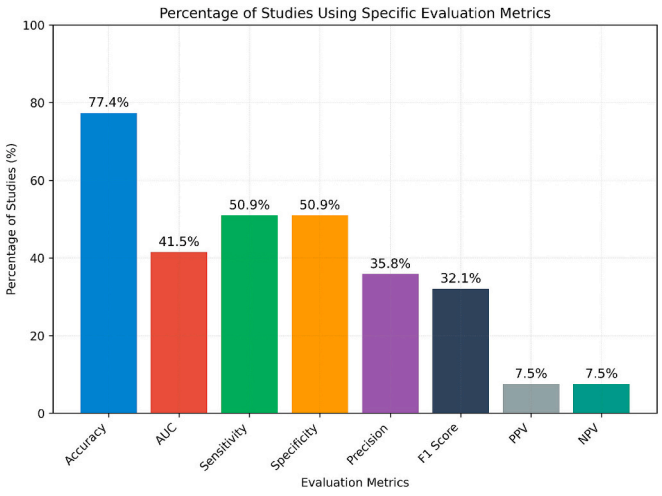


Fig. 7. Percentage of studies using specific evaluation metrics.

4.4. Discussion

Drawing from the studies analyzed with the SLR, this study identifies the important elements that define the process of developing accurate predictive models for diabetes. The discussion covers the choice of datasets and their quality, the chosen machine-learning algorithms and training paradigms, and the evaluation scenarios and measures used in the assessment of the performance of the model. In this review, the strengths and limitations of different approaches are discussed based on many publications, focusing on the issues of diversity and standardization of datasets, feature selection and preprocessing, and evaluation methods. Thus, this study seeks to provide an understanding of the status of diabetes prediction research and define the directions for its future enhancement to create more precise, accurate, and universally applicable predictive models.

4.4.1. Dataset utilization and Insights

While addressing **RQ$_1$**, it was necessary to initially analyze different datasets that are commonly employed in the studies aiming at diabetes prediction. As indicated, the reviewed studies used different datasets, from a longitudinal study, such as Aizawa Hospital, to large population-based studies, such as AusDiab. These datasets are valuable sources of information on diabetes and its prediction, as the nature and variety of this disease indicate. Nevertheless, problems such as variable quality of data and irregular approaches to its collection point to issues of weak data standardization. Combining datasets obtained from different regions or populations can help avoid creating models that do not work in different populations and under different clinical conditions.

4.4.2. Dataset selection and quality

Our SLR shows the great variety of datasets used in diabetes prediction studies proving the value of diversification and dataset samples. The characteristics and challenges of each dataset were different. For example, the NHANES has a large and racially/ethnically diverse sample, with extensive demographic, clinical, and lifestyle data. However, problems such as class imbalance, where one class has many samples and the other has few, become a problem for the model performance and its ability to generalize. However, diabetes prediction remains a challenging task in the presence of a class imbalance, especially when the number of non-diabetic cases is much greater than that of diabetic cases. To address this issue, various data and algorithm-level solutions have been proposed to generalize the model and reduce model bias.

In most cases, the synthetic minority oversampling technique (SMOTE) is employed to produce synthetic diabetic cases, thereby enhancing the sensitivity while reducing the downside impact on specificity. With this technique seemingly learning from a more balanced distribution, the likelihood of bias toward the majority class was reduced.

At the algorithm level, class weighting in loss functions, such as weighted cross-entropy, is used to penalize the misclassification of underrepresented cases with diabetes. The models achieved more balanced predictions with higher penalties for the misclassification of diabetes. Furthermore, Focal loss, a type of cross-entropy loss, has been used in deep learning models, particularly CNN-based diabetic retinopathy detection. Consequently, Focal Loss learns dynamically to reduce the weight of hard diabetic classes for the improved exploration of difficult cases, thus providing more robust predictions and smaller false negatives, which are crucial factors in clinical applications.

Together, these strategies affect the model performance in terms of maintaining sensitivity while decreasing false negatives; thus, diabetic cases were accurately identified. Additional data augmentation, adaptive loss functions (i.e., learned loss function), and ensemble learning added to overcome class imbalance in diabetes prediction models are future research horizons.

4.4.3. Data quality and consistency

It is more important for the quality of the datasets that go into the building of these models and for their consistency. Probable sources of bias include differences in data collection techniques and dissimilarities in the definitions of diabetes used in different studies. Some of these problems can be avoided when data collection protocols are standardized. For instance, a cross-sectional study of the patient database of Aizawa Hospital, Japan, and a large population-based cross-sectional AusDiab study showed that the methods of data collection may differ. This limitation is compounded by the fact that many of the studies available do not provide follow-up data, thus limiting the possibility of assessing the long-term value of predictive models. In addition, issues related to heterogeneity in the diagnosis and assessment of diabetes-related variables, including fasting glucose levels, HbA1c thresholds, and diagnostic criteria, also make it challenging to synthesize evidence from different studies.

4.4.4. Standardization and integration

It is crucial to note the attempts to establish international guidelines for the collection of data and reports concerning diabetes. To increase the reliability of the prediction models, it is necessary to unify the methods of data collection and combine different datasets. This approach can assist in addressing current drawbacks such as inconsistencies in demographic and clinical variables that can influence the results of the model. Combining datasets from different regions and population groups may offer a broader understanding of diabetes, and thus enable the creation of models that are viable in different populations and clinical situations. In addition, the combination of EHRs, genetic information, and CGM data with clinical and demographic data can improve the accuracy and comprehensiveness of the model.

4.4.5. Machine learning algorithms and training

The **RQ**$_2$ was on the approach used in the machine learning algorithms in predicting diabetes as well as the training approaches with independent variables. The studies showed how algorithms such as CNNs, SVMs, and XGBoost perform with different data types and different types of predictions. This shows that independent variables and training strategies, such as cross validation and data augmentation, are critical to improving the performance of the model. However, the choice of features and training methods influences their effectiveness and transferability. The use of these ML algorithms in different studies proves their efficiency in providing accurate predictions in different clinical areas.

4.4.6. Algorithm selection

The SLR reveals several ML algorithms utilized in diabetes prediction; they are CNNs, SVMs, Logistic Regression, and XGBoost. Each has advantages and is used for different data and prediction problems. For instance, CNNs are efficient in analyzing retinal images for DR, whereas SVM are employed for analyzing glycemic indices and demographic data. Despite the presence of more complex models, logistic regression continues to be used because of its simplicity and ease of interpreting results while analyzing structured clinical and demographic data. XGBoost is preferred owing to its demonstrated superiority with tabular datasets and flexibility in handling missing values and feature interactions.

4.4.7. Feature selection and data preprocessing

The selection of independent variables is appropriate when developing these models. The required inputs were demographic data, clinical measurements, and medical images of patients. Some of the critical techniques in the development of any model include feature selection and data pre-processing. Some of these are the choice of features for CNNs for retinal image analysis and training techniques for SVMs for glycemic indices. Feature selection techniques such as Recursive Feature Elimination and Wrapper methods guarantee that only the variables that are most beneficial to the model are used, thereby decreasing the noise level. Other preparations that may help enhance work on the project include data scaling, handling of missing values, and converting categorical data to numerical data.

4.4.8. Training strategies

Training strategies like cross-validation and data augmentation and feature selection are important for increasing the model reliability. The complexity and accuracy, as well as the independence of the data, were checked using methods such as k-fold cross-validation to avoid overfitting. Data augmentation, especially in image-based investigations, enhances the performance of the model because of variations in the dataset. For instance, the variation in retinal images by rotation and changing the brightness of the images can improve the generalization of the model. Pre-processing needs involve feature selection techniques that help identify the best variables to be used in the model to eliminate noisy ones. Bagging and boosting are other techniques that are also used to enhance the performance of models by using multiple models and obtaining one final result that is more accurate than the individual models.

4.4.9. Applications and use cases

The same ML algorithms have been applied in many research studies implying their efficiency and usefulness. For example, CNNs have been applied in the identification of diabetic retinopathy from retinal images, SVM in the analysis of ECG data for Diabetic and Non-Diabetic Heart Rate Variability and XGBoost to large-scale EHR data for diabetes onset and complication prediction. The above applications prove that with the execution of ML algorithms, predictions are accurate and reliable in different clinical practices. Further, incorporation of ML models with clinical decision support systems (CDSS) may help clinicians make better and timely decisions that may help to improve the quality of life of a patient and decrease the impact of complications due to diabetes.

4.4.10. Evaluation setups and metrics

As derived from **RQ**$_3$, which aimed at identifying the evaluation setups employed in the assessment of the machine learning models for diabetes prediction, including the kinds of validation employed and the measures used to measure the performance of the models. The studies used other techniques to assess the validity

and portability of the model, including k-fold cross-validation, external validation, and bootstrap sampling (Table 5). Measures such as accuracy, AUC, sensitivity, specificity, precision, and F1 score offered satisfactory measures of model performance, and thus underlined their value in clinical use. These evaluation setups and metrics ensure that models are not only valid in providing a range of clinical applications but also in terms of delivering rich and comprehensive information concerning model performance, thus helping to distinguish the most suitable predictive models. Key findings related to RQ$_3$ are presented in Table 6, which further provides a detailed information about the evaluation setups and metrics used in the literature.

4.4.11. Deciding evaluation methods

The evaluation setups that are incorporated in the diabetes prediction research are aimed at achieving the ML model validity and reliability. Cross validation, external validation, and bootstrap sampling are techniques that offer strong guidelines for the performance of the model. They assist in detecting overfitting, guaranteeing the generalization of models, and providing information about their advantages and limitations. For example, k-fold cross validation, where the set of collected data is split into k sets, where one is used for validation data and the other k-1 are used as training data, offers a more comprehensive result of the model.

4.4.12. Selecting key metrics

To assess the performance of the model, the evaluation measures of

Table 5
Key findings from RQ_2: on the trainings strategies, independent variables and ML algorithms.

Studies	Independent variables	Training strategies	ML algorithms
[34,35,41,63]	Retinal images	Cross-validation	CNN
[51]	Glycemic Indices	Recursive feature elimination, SVM training/testing split	
[26,37]	Demographic and Five-fold Cross-validation	Logistic regression health-related variables	
[14,64]	Fundus Images	Data augmentation	CNN (PhelcomNet)
[30,45,52,60,65,66,69,86,87]	EHR data	Feature selection, five-fold cross-validation	XGBoost
[53]	Fundus images	Data augmentation, cross validation	CNN
[67]	Demographic and clinical features	Parameter tuning, wrapper approaches	Random Forest
[50]	Metabolomics profiles	Multivariate logistic regression, cross-validation	Regularized Least Squares Regression
[63,68]	Retinal images	Testing on new dataset without retraining	Various FDA-approved models
[48]	Glucose Levels	Grid search, 10-fold cross validation	ANN, SVM
[27,46,54,55,57,59,64,68,70–75]	HRV signals from ECG	–	Hybrid Models(CNN-LSTM-SVM)
[56]	OCT measurements	–	Inception-V3
[29,48,49,76–84]	Demographic, clinical, lifestyle factors	–	Bayesian Network

Table 6
Key findings from RQ_3: evaluation techniques and metrics.

Studies	Evaluation setups	Evaluation metrics
[36,37,51]	10 fold cross validation, external validation	Accuracy, Precision, Recall, F1 Score
[63]	External validation	PPV, NPV
[27,30,49,54,65,67,68,76,77,87]	5-fold cross validation	Accuracy
[48,50,52,53,56,60,66,75,83,84]	Cross validation	AUC
[34,35,41,45–47,58,59,70,71,81,85]	Cross validation	Sensitivity, Specificity
[26,57,64,69,79,80,82,86]	Cross-validation	Precision, Recall, F1 Score
[63]	External validation	PPV, NPV

accuracy, Area Under the Curve (AUC), sensitivity, specificity, precision, and F1 score are essential. These metrics provide detailed information on how well a model identifies diabetes, reduces false-positive and false-negative rates, and works in practice. For instance, accuracy quantifies the number of true results (both true positives and true negatives) per total analyzed cases, whereas AUC provides information on the ability of the model to classify classes. Sensitivity and specificity: These are critical metrics for analyzing the capacity of a model to diagnose positive and negative cases. Accuracy is not recommended for imbalanced datasets because it tends to favor the majority class, whereas Precision and F1 score, which take into account both precision and recall, are recommended for use with imbalanced datasets.

4.4.13. *Ensuring reliability*

The validity of the models can be confirmed through the use of datasets other than the ones used in training as a way of testing if the models developed will work well on new data. For instance, the study conducted using the Optumő EHR dataset was externally validated by comparing model predictions to the scores of professional graders of new images from the screening programs, as well as other diverse population groups. Bootstrap sampling, which trains the model multiple times with random samples drawn with replacements from the dataset, provides an empirical measure of the variability of the performance measures. These evaluation setups help ensure that the models developed are more reliable, valid, and transportable to a broad spectrum of clinical practice. In addition, the interpretability and explainability of the models are essential, especially in a clinical environment, where the ability to understand the decision-making process of the system will increase acceptance by practitioners.

4.4.14. *Advances in AI for diabetes prediction*

Emerging technologies, especially in the domain of ML and DL, have enriched the idea of diabetes prediction in the recent past. These include the accuracy of diagnosis, features extracted from the data, and the ability to analyze data in real time for early diagnosis, risk evaluation, and management. Data pre-processing is an important step that directly affects the performance of the model. Techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are used to minimize computational intensity while optimizing significant biomarkers [27]. This will help in handling class imbalance, and techniques such as the synthetic minority oversampling technique (SMOTE) will be employed to enhance the generalization of the models [62]. Some of these methods include contrast enhancement to adjust the image contrast before feeding the data into the deep learning model and GANs to generate many new image datasets to enhance the capability of the deep learning model in feature extraction [26]. Deep learning techniques exhibit better performance than standard ML methods in the case of diabetes-related tasks. CNNs are widely used in Diabetic Retinopathy (DR) because of their efficiency in identifying diabetic retinopathy with high accuracy in analyzing retinal images [55]. There is also an enhancement of target lesions in these models by focusing on the attention mechanisms. Therefore, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are suitable for working with continuous glycogen measurements and for calculating real-time risk assessments, as is the case with CGM [30]. The potential of AI in this field remains wide as it has been used for diabetes risk appraisal and control. For example, the current tools approved by the Food and Drug Administration include Google DeepMind and IDx-DR, and employ CNNs for DR screening [68]. Moreover, it is efficient in predicting diabetes onset several years prior to diagnosis [82]. AI in wearable and mobile health technologies has also been applied to glucose monitoring and data sharing using transfer learning to forecast risks [70].

In the future, the issue of Explainable AI (XAI) will remain important for clinical implementation because it addresses how the models make decisions. Another promising approach is federated learning, which allows obtaining a model updated on a server without uploading the patient data. Moreover, applying multimodal AI models that use genomic, imaging, and lifestyle data will provide a framework for the enhanced precision of diabetes prediction, and hence, provide a better treatment option.

The discussion also focuses on the selection of datasets, the choice of machine learning algorithms, and the evaluation frameworks in the creation of reliable diabetes prediction models. Hence, despite the progress made in the field, solving problems regarding data quality, consistency, and privacy is crucial for future development. Here, interdisciplinary cooperation and compliance with the standardized procedure of data collection and the use of sophisticated algorithms will allow the potential of machine learning to be realized and contribute to the efficient treatment of diabetes. The use of multiple sources of data,

appropriate selection of features, and better training and validation paradigms will improve the robustness and transferability of predictive models and thus contribute to the betterment of lives as well as the field of diabetes prediction. Nevertheless, *what are the main limitations that have been found in this SLR, and how can future studies deal with these issues to enhance diabetes prediction models?* This question will be discussed in the next section to identify the current limitations of diabetes prediction and future directions for the improvement of the models.

5. Research limitations and implications

Diabetes prediction using artificial intelligence, specifically machine learning (ML), has presented a method of early diagnosis and effective control of the disease. Our study identified some issues, such as data quality, feature selection, model complexity, and ethical implications, which make it difficult to achieve in the healthcare domain. Solving these problems is vital for building stable and accurate ML models that can be incorporated into clinical settings. In this section, we discuss the limitations and their consequences for researchers and stakeholders and outline the steps to enhance the field and diabetes prognosis and management.

5.1. Data quality and standardization issues

The datasets used to build the AI-based diabetes prediction models were reliable and consistent. Some frequently used databases include NHANES, PIDD, Optumő EHR, EyePACS, MIMIC-III, KNHANES, REPLACE-BG, Humedica, Practice Fusion EHR, and the Itabuna Diabetes Campaign Dataset, which have issues of missing data, data collection methods, and variability in definitions. These differences arise from the differences in the diagnostic approach, training strategy, and evaluation technique, which have implications for the accuracy and transferability of the model.

5.1.1. Demographic and ethnic bias

As the name suggests, the Pima Indians Diabetes Database (PIDD) is a database of adult females of Pima Indian origin belonging to one demographic population and, therefore, not generalizable to other populations. Consequently, the modeling risk hypothesis arising from this dataset may not perform optimally with other ethnic groups or male patient data, and can lead to poor accuracy when healthcare is predicated on the findings of the model in diverse healthcare facilities. The EYEPACS used for diabetic retinopathy classification images has a significant concern regarding ethnic bias, as some ethnic groups retain a high representation, while others have a low representation. This leads to performance disparity because CNN-based models trained on this dataset cannot accurately classify retinal images from such populations. Thus, such models cannot be applied in diverse clinical settings and are not generalizable.

5.1.2. Class imbalance

NHANES has more subjects not screened for diabetes, causing low sensitivity and misclassification of the screened diabetic subjects. There are class imbalance problems for a sizeable portion of female Indian patients with PIDD. Early and advanced diabetes cases are not well distributed in Optumő EHR, which hinders the prediction of disease progression. CNN-based models for constructing different risk factors from datasets such as EyePACS and the Kaggle Diabetic Retinopathy dataset are relatively ineffective for the detection of severe cases owing to the imbalance in DR severity levels. The MIMIC III and Humedica datasets were skewed toward type 2 diabetes, decreasing the model accuracy for Type 1 prediction.

5.1.3. Missing and inconsistent data

It is important to note that the data in the Optumő EHR sample type are missing some records, and the medical codes are not uniform. Some basic aspects, such as HbA1c levels, BMI, and blood pressure, may not be available at certain times, and this requires the use of imputation techniques that introduce artificial bias that affects the results generated by AI models of a particular disease.

5.1.4. Retrospective vs. prospective data limitations

Many AI models rely on retrospective datasets such as NHANES, Optumő EHR, and MIMIC-III, which contain historical patient records. These datasets provide large-scale information, but lack longitudinal follow-up, making it difficult to assess disease progression or predict long-term outcomes. Conversely, prospective studies, such as the Botnia Prospective Study, offer real-time data collection but suffer from smaller sample sizes, longer collection periods, and high resource demands, limiting their feasibility for large-scale predictive modeling. The choice of retrospective versus prospective data significantly impacts model generalizability, where retrospective models may struggle with real-world applications, whereas prospective models risk overfitting owing to limited data availability.

These dataset limitations compromise model fairness, robustness, and clinical applicability, requiring targeted interventions, such as bias-correction techniques, multi-source data integration, external validation, and hybrid modeling approaches, to ensure that AI-driven diabetes prediction models perform reliably across diverse populations.

☞ Better quality and variety of data, consistent methods of data gathering, and balancing classes in the models are important for improving the machine learning models dependability and credibility. This is why healthcare organizations and policymakers should ensure the development of large databases with ethnic, demographic and geographical characteristics of minorities; this would increase the relevance of the models used.

5.2. Feature engineering and selection

Diabetes prediction models generally involve feature engineering and selection since they use demographic and clinical aspects such as age, BMI, blood glucose level, and others. However, the disease is complex, and long-term models that do not include genetic and lifestyle factors are very simplified. Feature selection is always a problem; researchers usually choose features that are either redundant or irrelevant, which is bad for the model.

☞ Researchers should use genetic and other related factors as well as lifestyle and behavior data in their research and should use dimensionality reduction and feature importance analysis. Clinicians were the main targets involved in the process of model creation and refinement to achieve accuracy and validity. Therefore, stakeholders should promote programs that increase the number of data assets and select better features for predictions.

5.3. Model complexity and interpretability

The complexity of the models also poses a problem in the prediction of diabetes. Large neural networks are precise; however, they are overparameterized and overfit the training data. These models are often “black boxes”, which make it difficult to explain the cause of a decision to a patient or another health care worker. This lack of transparency is a drawback in terms of adoption and patient management. As diabetes prediction and as well as diabetic retinopathy detection rely on AI-based models, it is important to focus more on Explainable AI techniques for clinical transparency and trust. Methods such as Grad-CAM, SHAP, and LIME have been successfully incorporated into CNN-based- and tree-based models.

As diabetes prediction as well as diabetic retinopathy detection rely on AI-based models, it is important to focus more on Explainable AI (XAI) techniques for clinical transparency and trust. Methods such as Grad-CAM, SHAP, and LIME have been successfully incorporated into CNN-based and tree-based models; however, their adoption is limited to

other styles of deep-learning models that do not consist of CNNs. As part of future research, to fill this gap, hybrid AI frameworks that integrate leading deep learning models with more interpretability methods should be created. In addition to the reliability of AI-driven diagnostics, the integration of multimodal interpretability, attention, and attribute attribution methods can be beneficial. Furthermore, cooperation between AI researchers and clinicians is of key importance in defining standardized XAI guidelines so that model explanations correspond to reality in clinical practice and regulations.

Ultimately, XAI techniques should be expanded to diabetes prediction and other healthcare domains to foster trust, encourage adoption, and lead to models based on AI being used effectively in clinical decision-making.

☞ Models of moderate complexity should be easier to interpret and that explainability should then be applied as an intervention. If efficient models cannot be used, more complex models should be used in the predictions. Clinicians should consider using usability models to enhance the implementation of clinical practice. Training and other resource supplies to healthcare personnel need to be encouraged.

5.4. Training and validation limitations

Different methods are used for better model performance, such as crossvalidation, data augmentation, and hyperparameter tuning. However, these techniques also have certain disadvantages. Validation confirms the model performance on a specific dataset but does not ensure its generalizability. However, models can be adapted to new datasets and new datasets can be accommodated by the models. Furthermore, deep learning models rely on sizeable datasets, and when there are inadequate training samples, these overfits are impractical for real-life applications.

☞ Data-augmentation and cross-validation should be employed by researchers to boost the model performances. Standard procedures for training and validation should be set by all stakeholders to minimize the variations between the different studies and applications. It is suggested that principles and best practices for ML models can increase their dependence and make outcomes more similar and comparable; in this case, institutions should collaborate for better research in this regard.

5.5. Lack of standardized evaluation metrics

Different approaches adopt the use of measures, including accuracy, area under the curve, sensitivity, and specificity; therefore, there is no proper way of comparing modes. Some models are aimed at high accuracy, which is not appropriate for imbalanced datasets, while others tend to focus on different aspects of precision-recall curves, which also leads to inconsistencies in the results.

☞ The used evaluation metrics should include accuracy, AUC, sensitivity, specificity, and clinically oriented performance. Stakeholders should ensure that standard evaluation tools are used to serve clients and meet the market standards. A criterion for the improvement of predictive models based on statistical results and clinical relevance was developed. This will improve the credibility of the ML models since openness will be promoted.

5.6. Computational and resource constraints

Complex deep learning models such as CNNs require substantial computational power, making them impractical for resource-limited clinical settings. Federated learning and model compression techniques have been proposed as solutions, but they remain underexplored in diabetes prediction research.

☞ Researchers have to design a model for health care provision, irrespective of the available resources. Efforts must be made to compute assets by key players, and standardization of the same must be made. Governments should allocate resources for technology and training distributions to reduce the disparity of healthcare quality.

5.7. Ethical and legal considerations

There are ethical and legal issues that come with ML models in diabetes prediction such as bias in the training data and measures such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) that limit data access. These problems can aggravate health disparities, restrict data protection and security, and impede data sharing and model building.

☞ The models that have been developed require high reliability, and the ethical norms such as HIPAA and GDPR contribute to it. This guarantees appropriate use of data making statistical models more precise and helps to eliminate discrimination and unfair practices in the sphere which in its turn will benefit society and provide equal treatment for patients.

5.8. Lack of collaboration on diabetes prediction efforts

Collaboration in the diabetes prediction hamper the data accessibility for constructing a model, its testing, moral analysis, and its practical use in real-world settings, which slows down research, development, and patient outcomes and benefits, while there is a requirement for diverse and non-bias data.

☞ Clinicians and stakeholders should engage researchers in creating communication channels for new models in order to improve performance and fit the demands of the healthcare sectors. Involvement of stakeholders helps in producing clinically meaningful, technologically realistic, and ethically sound models that enhance patients experience.

Addressing these limitations requires future studies to standardize the dataset selection, implement external validation strategies, and develop more interpretable AI models to ensure real-world applicability.

6. Threats to validity

Like any other systematic literature review (SLR), the present study has some limitations that could have potentially affected the validity of the observed results. This section presents these limitations and measures taken to manage them.

6.1. Literature selection

The study might have limited its effectiveness due to the exclusion of studies from sources beyond IEEE, PubMed, and ScienceDirect without adding the Web of Science and Scopus databases. The restricted search databases might have reduced both the amount of research materials included and the general review thoroughness. One of the crucial issues in conducting a systematic literature review is how to find sufficient papers to provide a general understanding of the state of the art of a given research area. In this regard, the present study formulated a comprehensive search question with no temporal restrictions to acquire as many papers as possible concerning the application of machine learning for diabetes prediction. Although this approach is time consuming, it is used to achieve exhaustiveness. It is important to note that synonyms and alternative spellings of the terms commonly used in the literature to define the search query were identified. Furthermore, we searched for these search terms among the systematic literature reviews on diabetes prediction to determine if there are other suitable terms. To further enhance data collection in the research area, a backward snowballing session was conducted on the papers obtained after the exclusion/inclusion criteria were applied. To ensure credibility, all processes to arrive at the choice of the primary studies were cross-checked by at least one of the authors. The implementation of these actions allows us to gain confidence in the comprehensiveness of the selection of literature sources. To ensure that all steps and intermediary results of the analyses reported here can be verified and independently

replicated, all of them are presented in the online appendix.

6.2. Literature analysis and synthesis

Following the selection process, the following exclusion criteria were used to remove papers that could not make a significant contribution in the summarization of the state of the art about the defined research questions. We did not restrict the list of primary studies to articles that met the inclusion criteria but also performed an extra quality check to confirm their relevance. To ensure that no resources that do not meet the objectives of the study are included, this manual assessment posed an additional layer to the process.

More broadly, the literature synthesis was performed according to the results of manual analyses, which are known to be prone to human factors. In this regard, two observations are necessary. First, the two main authors were involved in the process, which reduced the subjectivity and possible mistakes. Second, a third author was consistently involved, and he provided input on how to perform the different phases of the systematic literature review whenever necessary.

These combined efforts go a long way toward reducing the threats to validity and provide a thorough and comprehensive review of the current state of affairs regarding the use of machine learning techniques in the prediction of diabetes.

7. Conclusion

This systematic review demonstrated the future progress and productivity of ML in the diagnosis and management of diabetes, a major global health concern. In this way, while comparing 53 studies, this review offers an overview of the datasets, ML algorithms, training methods, independent variables, and evaluation metrics used in diabetes prediction. Some of these datasets include the Singapore National Diabetic Retinopathy Screening Program, REPLACE-BG, National Health and Nutrition Examination Survey (NHANES), the Pima Indians Diabetes Database (PIDD), Optumő EHR, EyePACS, MESSIDOR, Kaggle Diabetic Retinopathy Dataset, KNHANES, and the Humedica database, which come with their peculiarities, some of which are class imbalance. This review highlights the positive impact of several ML algorithms, such as CNN, SVM, Logistic Regression, and XGBoost, in diagnosing diabetes. The interpretability of AI-driven diabetes prediction and diabetic retinopathy detection models depends largely on explainable AI (XAI) methods for clearing up their functionality. The increasing use of Grad-CAM, SHAP, and LIME techniques provides medical professionals with more insights into the decision-making process of machine learning models with enhanced clinical reliability and transparency levels. The restricted use of XAI methods for CNN networks requires researchers to conduct additional studies to achieve interpretability across all types of machine learning applications. Other attributes that are often used as independent variables include age, body mass index, blood glucose concentrations, genetic polymorphisms, and lifestyle, which are instrumental in building forecasting models. In addition, this review provides insights into methods such as cross-validation, data augmentation, and feature selection, which improve the flexibility and stability of the models. Therefore, it is crucial to use assessment indicators such as accuracy, AUC, sensitivity, and specificity to provide a comprehensive assessment of the model. In the future, it will be necessary to overcome the current weaknesses to enhance the utilization of ML for diabetic prediction. Future studies should focus on the quality and variability of data, methods of handling class imbalance, interpretability of the model, and computational complexity. Multi-center studies involving various population groups and standardizing the metrics for the evaluation and validation of the models are the few important steps that need to be taken. If these challenges are addressed, ML has the potential to enhance the accuracy of diagnosis, health of patients, and effectiveness of the healthcare system, thereby lowering the global impact of diabetes. In light of these findings, this review calls for the integration of ethicists and other stakeholders in formulating recommendation policies involving the application of ML-based diabetes prediction models aimed at enhancing the quality of life of people globally, through the use of AI technology in the delivery of healthcare services. The findings and recommendations of this review are useful in the current drive toward the use of AI and ML in combating one of the biggest challenges to health in the modern world.

CRediT authorship contribution statement

Pir Bakhsh Khokhar: Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Carmine Gravino:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis. **Fabio Palomba:** Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis.

Funding

This work has been partially supported by the European Union through the Italian Ministry of University and Research, Project PNRR "*D3-4Health*: Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care". PNC 0000001. CUP B53C22006090001.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgements

The authors would like to express their sincere gratitude to the Department of Informatics, University of Salerno, for providing the research environment and institutional support necessary for this work. We also thank our colleagues for their constructive feedback and discussions that significantly contributed to the improvement of this study. Special thanks to the reviewers for their valuable comments and suggestions that helped refine the quality of this manuscript.

References

[1] Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018;20:e10775.
[2] Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. Ob- taining data from electronic health records, in: Tools and technolo- gies for registry interoperability, registries for evaluating patient out- comes: A users guide. In: 3rd edition, Addendum 2 [Internet]. Agency for: Healthcare Research and Quality (US); 2019.
[3] Liu K, Li L, Ma Y, Jiang J, Liu Z, Ye Z, et al. Machine learning models for blood glucose level prediction in patients with diabetes mellitus: systematic review and network meta-analysis. JMIR Med Inform 2023;11:e47833.
[4] Costanzo MC, von Grotthuss M, Massung J, Jang D, Caulkins L, Koesterer R, et al. The type 2 diabetes knowledge portal: an open access genetic resource dedicated to type 2 diabetes and related traits. Cell Metab 2023;35:695–710.
[5] Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr 2021;13:148.
[6] Ahmed SF, Alam MSB, Hassan M, Rozbu MR, Ishtiak T, Rafa N, et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review 2023;56:13521–617.
[7] Zhao Z, Alzubaidi L, Zhang J, Duan Y, Gu Y. A comparison review of transfer learning and self-supervised learning: definitions, applications, advantages and limitations. Expert Systems with Applications 2023;122807.
[8] Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, et al. Machine learning models for data- driven prediction of diabetes by lifestyle type. Int J Environ Res Public Health 2022;19:15027.
[9] Chicco D, Starovoitov V, Jurman G. The benefits of the Matthews correlation coefficient (mcc) over the diagnostic odds ratio (dor) in binary classification assessment. Ieee Access 2021;9:47112–24.
[10] Cabrera, D., Cabrera, L.L., 2023. The steps to doing a systems literature review (slr). Journal of Systems Thinking Preprints.
[11] Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. The lancet 2017;389:2239–51.
[12] Alsaigh A, Almalki FAM, Alnefaie MAM, et al. The role of physical therapists in fighting the type 2 diabetes epidemic patients attending primary healthcare centers

in Makkah city, Saudi Arabia in 2022. Annals of the Romanian Society for Cell Biology 2022;26:3135–48.
[13] Association, A.D.. Physical activity/exercise and diabetes mellitus. Diabetes Care 2003;26:s73–7.
[14] Bloom DE, Canning D, Fink G. Program on the global demography of aging. Harvard University. Oct; 2009.
[15] DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, et al. Type 2 diabetes mellitus. Nat Rev Dis Primers 2015;1:1–22.
[16] Daneman D. Type 1 diabetes. The Lancet 2006;367:847–58.
[17] Ahmad E, Lim S, Lamptey R, Webb DR, Davies MJ. Type 2 diabetes. The Lancet 2022;400:1803–20.
[18] Wadghiri MZ, Idri A, El Idrissi T, Hakkoum H. Ensemble blood glucose prediction in diabetes mellitus: a review. Comput Biol Med 2022;147:105674.
[19] Bidwai P, Gite S, Pahuja K, Kotecha K. A systematic literature review on diabetic retinopathy using an artificial intelligence approach. Big Data and Cognitive Computing 2022;6:152.
[20] Usman TM, Saheed YK, Nsang A, Ajibesin A, Rakshit S. A systematic literature review of machine learning based risk prediction models for diabetic retinopathy progression. Artif Intell Med 2023;143:102617.
[21] Wijoseno MR, Permanasari AE, Pratama AR. Ma- chine learning diabetes diagnosis literature review, in: 2023 10th international conference on information technology, computer, and electrical engineering (ICITACEE). IEEE pp 2023; 304–308.
[22] Saxena R, Sharma SK, Gupta M, Sampada G. [retracted] a comprehensive review of various diabetic prediction models: a literature survey. Journal of Healthcare Engineering 2022;2022:8100697.
[23] Felizardo V, Garcia NM, Pombo N, Megdiche I. Data- based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review. Artif Intell Med 2021;118:102120.
[24] Idrissi TE, Idri A, Bakkoury Z. Systematic map and review of predictive techniques in diabetes self-management. International Journal of Information Management 2019;46:263–77.
[25] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C. D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al., 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. Bmj 372.
[26] Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124:962–9.
[27] Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data 2019;6:1–19.
[28] Centers for Disease Control and Prevention. National Diabetes Statistics Report. CDC: Technical Report; 2023.
[29] Chetoui M, Akhloufi MA, Kardouchi M. Diabetic retinopa- thy detection using machine learning and texture features, in: 2018 IEEE Canadian conference on electrical & computer engineering (CCECE). IEEE. pp. 2018;1–4.
[30] Swapna G, Vinayakumar R, Soman K. Diabetes detection using deep learning algorithms. ICT express 2018;4:243–6.
[31] Hall MA. Correlation-based feature selection for machine learning. Ph.D. thesis. The University of Waikato; 1999.
[32] Naidu G, Zuva T, Sibanda EM. A review of evaluation metrics in machine learning algorithms. Computer Science On- line Conference, Springer 2023:15–25.
[33] Ghadikolaei HS, Ghauch H, Fischione C, Skoglund M. Learning and data selection in big datasets. International Confer- ence on Machine Learning, PMLR 2019: 2191–200.
[34] Ting DSW, Cheung CYL, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Jama 2017;318:2211–23.
[35] Sangi M, Win KT, Shirvani F, Namazi-Rad MR, Shukla N. Applying a novel combination of techniques to develop a predictive model for diabetes complications. PloS One 2015;10:e0121569.
[36] Kim J, Kim J, Kwak M, Bajaj M. Genetic prediction of type 2 diabetes using deep neural network. Clin Genet 2018;93:822–9.
[37] Vangeepuram, N., Liu, B., Chiu, P.h., Wang, L., Pandey, G., 2021. Predicting youth diabetes risk using nhanes data and machine learn- ing. Sci Rep 11, 11212.
[38] Pima Indians Diabetes Database, 2023. Pima indians diabetes dataset (pidd) documentation. Available: https://www.kaggle.com/datasets/ uciml/pima-indians -diabetes-database.
[39] Li X, Jiang Y, Zhang J, Li M, Luo H, Yin S. Lesion- attention pyramid network for diabetic retinopathy grading. Artif Intell Med 2022;126:102259.
[40] Aleppo G, Ruedy KJ, Riddlesworth TD, Kruger DF, Peters AL, Hirsch I, et al. Replace-bg: a randomized trial comparing con- tinuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes. Diabetes Care 2017;40:538–45.
[41] Yokota N, Miyakoshi T, Sato Y, Nakasone Y, Yamashita K, Imai T, et al. Predictive models for conversion of prediabetes to diabetes. J Diabetes Complications 2017; 31:1266–71.
[42] Dunstan DW, Zimmet PZ, Welborn TA, Cameron AJ, Shaw J, De Courten M, et al. The australian diabetes, obesity and lifestyle study (ausdiab)methods and response rates. Diabetes Res Clin Pract 2002;57:119–29.
[43] Pires R, Avila S, Wainer J, Valle E, Abramoff MD, Rocha A. A data-driven approach to referable diabetic retinopathy detection. Artif Intell Med 2019;96:93–106.
[44] Sambo F, Facchinetti A, Hakaste L, Kravic J, Di Camillo B, Fico G, et al. A bayesian network for probabilistic reasoning and imputation of missing risk factors in type 2 diabetes, in: Artificial intelligence in medicine. In: 15th conference on artificial intelligence in medicine, AIME 2015, Pavia, Italy, June 17–20, 2015. Proceedings 15, springer. Pp; 2015. p. 172–6.
[45] Li L, Lee CC, Zhou FL, Molony C, Doder Z, Zalmover E, et al. Performance assessment of different machine learning approaches in predicting diabetic ketoacidosis in adults with type 1 diabetes using electronic health records data. Pharmacoepidemiol Drug Saf 2021;30:610–8.
[46] Nguyen BP, Pham HN, Tran H, Nghiem N, Nguyen QH, Do TT, et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput Methods Programs Biomed 2019;182:105055.
[47] Malerbi FK, Andrade RE, Morales PH, Stuchi JA, Lencione D, de Paulo JV, et al. Diabetic retinopathy screening using artificial intelligence and handheld smartphone-based retinal camera. J Diabetes Sci Technol 2022;16:716–23.
[48] Tang F, Luenam P, Ran AR, Quadeer AA, Raman R, Sen P, et al. Detection of diabetic retinopathy from ultra-widefield scanning laser ophthalmoscope images: a multicenter deep learning analysis. Oph- thalmology Retina 2021;5:1097–106.
[49] Moreno, E.M., Lujan, M.J.A., Rusinol, M.T., Fernandez, P.J., Man- rique, P.N., Trivino, C.A., Miquel, M.P., Rodriguez, M.A., Burguil- los, M.J.G., 2016. Type 2 diabetes screening test by means of a pulse oximeter. IEEE Transactions on Biomedical Engineering 64, 341–351.
[50] Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 2021;44:1168–75.
[51] Herrero P, Reddy M, Georgiou P, Oliver NS. Identifying continuous glucose monitoring data using machine learning. Diabetes Technol Ther 2022;24:403–8.
[52] Anggraeni Z, Wibawa HA. Detection of the emergence of exudate on the image of retina using extreme learning machine method, in: 2019 3rd international conference on informatics and computational sciences (ICICoS). IEEE. pp. 2019; 1–6.
[53] Buccheri E, Dell'Aquila D, Russo M. Artificial intelligence in health data analysis: the darwinian evolution theory suggests an extremely simple and zero-cost large-scale screening tool for predi- abetes and type 2 diabetes. Diabetes Res Clin Pract 2021;174:108722.
[54] Karkuzhali S, Manimegalai D. Distinguising proof of dia- betic retinopathy detection by hybrid approaches in two dimensional retinal fundus images. J Med Syst 2019;43:1–12.
[55] Arcadu F, Benmansour F, Maunz A, Michon J, Haskova Z, McClintock D, et al. Deep learning predicts oct measures of diabetic macular thickening from color fundus photographs. Invest Ophthalmol Vis Sci 2019;60:852–7.
[56] Debédat J, Sokolovska N, Coupaye M, Panunzi S, Chakaroun R, Genser L, et al. Long-term relapse of type 2 diabetes after roux-en-y gastric bypass: prediction and clinical relevance. Diabetes Care 2018;41:2086–95.
[57] Chen W, Chen S, Zhang H, Wu T. A hybrid prediction model for type 2 diabetes using k-means and decision tree, in: 2017 8th IEEE international conference on software engineering and ser- vice science (ICSESS). IEEE. pp. 2017;386–390.
[58] Kotfila C, Uzuner Ö. A systematic comparison of feature space effects on disease classifier performance for phenotype iden- tification of five diseases. J Biomed Inform 2015;58:S92–102.
[59] Qiu H, Yu HY, Wang LY, Yao Q, Wu SN, Yin C, et al. Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy. Sci Rep 2017;7: 16417.
[60] Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D. The impact of oversampling with smote on the performance of 3 classifiers in prediction of type 2 diabetes. Med Decis Making 2016;36:137–44.
[61] Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current techniques for diabetes prediction: review and case study. Applied Sciences 2019;9:4604.
[62] Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, Sherazi HHR. Machine learning based diabetes classification and pre- diction for healthcare applications. Journal of healthcare engineering 2021;2021:9930985.
[63] Choi, S.B., Kim, W.J., Yoo, T.K., Park, J.S., Chung, J.W., Lee, Y.h., Kang, E.S., Kim, D.W., 2014. Screening for prediabetes using machine learning models. Comput Math Methods Med 2014, 618976.
[64] Seiglie J, Platt J, Cromer SJ, Bunda B, Foulkes AS, Bassett IV, et al. Dia- betes as a risk factor for poor early outcomes in patients hospitalized with covid-19. Diabetes Care 2020;43:2938–44.
[65] Bhuvaneswari G, Manikandan G. A novel machine learning framework for diagnosing the type 2 diabetics using temporal fuzzy ant miner decision tree classifier with temporal weighted genetic algorithm. Computing 2018;100:759–72.
[66] Peddinti G, Cobb J, Yengo L, Froguel P, Kravić J, Balkau B, et al. Early metabolic markers identify potential targets for the prevention of type 2 diabetes. Diabetologia 2017;60:1740–50.
[67] Chen H, Tan C, Lin Z, Wu T. The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. Comput Biol Med 2014;50:70–5.
[68] Gadekallu TR, Khare N, Bhattacharya S, Singh S, Maddikunta PKR, Ra IH, et al. Early detection of diabetic retinopathy using pca-firefly based deep learning model. Electronics 2020;9:274.
[69] Almutairi ES, Abbod MF. Machine learning methods for diabetes prevalence classification in Saudi Arabia. Modelling 2023;4:37–55.
[70] Chowdary PBK, Kumar RU. An effective approach for detecting diabetes using deep learning techniques based on convo- lutional lstm networks. International Journal of Advanced Computer Science and Applications 2021;12:519–25.
[71] Deberneh HM, Kim I. Prediction of type 2 diabetes based on machine learning algorithm. Int J Environ Res Public Health 2021;18:3317.
[72] Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning tech- niques. BMC Endocr Disord 2019;19:1–9.
[73] Nijalingappa P, Sandeep B. Machine learning approach for the identification of diabetes retinopathy and its stages, in: 2015 international conference on applied and theoretical computing and communication technology (iCATccT). IEEE. pp. 2015;653–658.

[74] Pustozerov EA, Tkachuk AS, Vasukova EA, Anopova AD, Kokina MA, Gorelova IV, et al. Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus. Ieee Access 2020;8:219308–21.
[75] Swapna G, Kp S, Vinayakumar R. Automated detection of diabetes using cnn and cnn-lstm network and heart rate signals. Procedia computer science 2018;132: 1253–62.
[76] Al-Tarawneh M, Muheilan M, Al Tarawneh Z. Hand movement-based diabetes detection using machine learning tech- niques. International Journal on Engineering Applications (IREA) 2021;9:234–42.
[77] Anderson JP, Parikh JR, Shenfeld DK, Ivanov V, Marks C, Church BW, et al. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. J Diabetes Sci Technol 2016;10:6–18.
[78] Dinh A, Miertschin S, Young A, Mohanty SD. A data- driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 2019;19:1–15.
[79] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., Mous- takas, K., 2021. Machine learning tools for long-term type 2 diabetes risk prediction. Ieee Access 9, 103737–103757.
[80] Jahangir M, Afzal H, Ahmed M, Khurshid K, Nawaz R. An expert system for diabetes prediction using auto tuned multi-layer perceptron, in: 2017 intelligent systems conference (IntelliSys). IEEE. pp. 2017;722–728.
[81] Karthikeyan S, Sanjay KP, Madhusudan R, Sundaramoorthy S, Namboori PK. Detection of multi-class retinal diseases using artificial intelligence: an expeditious learning using deep cnn with minimal data. Biomedical & Pharmacology Journal 2019;12:1577.
[82] Nuankaew P, Chaising S, Temdee P. Average weighted objective distance-based method for type 2 diabetes prediction. IEEE Access 2021;9:137015–28.
[83] Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo Á, Barreto SM, et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes-elsa-brasil: accuracy study. Sao Paulo Medical Journal 2017; 135:234–46.
[84] Zhao C, Yu C. Rapid model identification for online subcutaneous glucose concentration prediction for new subjects with type i diabetes. IEEE Transactions on Biomedical Engineering 2015;62:1333–44.
[85] Casanova R, Saldana S, Simpson SL, Lacy ME, Subauste AR, Blackshear C, et al. Prediction of incident diabetes in the Jackson heart study using high-dimensional machine learning. PloS One 2016;11:e0163942.
[86] Khan MZ, Mangayarkarasi R, Vanmathi C, Angulakshmi M. Bio-inspired pso for improving neural based diabetes prediction system. Journal of ICT Standardization 2022;10:179–99.
[87] Samant P, Agarwal R. Analysis of computational techniques for diabetes diagnosis using the combination of iris-based features and physiological parameters. Neural Computing and Applications 2019;31:8441–53.