Asian Journal of Electrical Sciences
ISSN: 2249-6297 (P); 2583-7923 (E)
Vol.14 No.2, 2025, pp.28-34

www.crijournals.org
DOI: https://doi.org/10.70112/ajes-2025.14.2.4292

A Systematic Review of Machine Learning Models for Predicting Type 2 Diabetes Mellitus Using Electronic Health Records

Oduware C. Odigie[1]*, Folasade Y. Ayankoya[2], Shade O. Kuyoro[3] and Ayodeji G. Abiodun[4]
[1*, 2, 3 & 4]Department of Computer Science, Babcock University, Nigeria
E-mail: ayankoyaf@babcock.edu.ng, kuyoros@babcock.edu.ng, abiodun0208@pg.babcock.edu.ng
*Corresponding Author: odigie0031@pg.babcock.edu.ng
(Received 26 July 2025; Revised 5 September 2025; Accepted 3 October 2025; Available online 20 October 2025)

Abstract **- This systematic review evaluates the use of machine learning models for predicting type 2 diabetes mellitus using electronic health record data. The global increase in the prevalence of type 2 diabetes underscores the need for reliable early prediction methods that can identify individuals at risk before disease onset. Machine learning provides an opportunity to improve predictive performance by uncovering complex relationships in clinical data that traditional statistical approaches may not capture. To assess progress in this area, a comprehensive search of the Scopus and PubMed databases was conducted to identify relevant studies published between January 2020 and October 2025. A total of 329 records were retrieved, and 13 studies met the inclusion criteria following a structured screening and quality assessment process. Data were extracted on model type, dataset characteristics, and reported outcomes. The reviewed studies showed that ensemble models and deep learning architectures generally achieved stronger predictive performance than single classifiers. Common predictors identified across studies included fasting plasma glucose, HbA1c, triglycerides, body mass index, age, and lipid measures. Although most models demonstrated high discrimination, key methodological limitations persisted, including insufficient external validation, inconsistent performance reporting, and limited transparency in data processing. The findings suggest that machine learning applied to electronic health record data offers significant potential for the early detection of type 2 diabetes; however, clinical adoption will require standardized evaluation frameworks, robust validation across diverse populations, and improved model interpretability to ensure trustworthy and equitable implementation in healthcare settings.**

Keywords: **Machine Learning, Type 2 Diabetes Mellitus, Electronic Health Records, Predictive Modeling, Ensemble Models, Deep Learning**

I. INTRODUCTION

Type 2 diabetes mellitus (T2D) is a chronic metabolic disorder characterized by impaired insulin secretion and resistance to insulin action, leading to persistent elevation of blood glucose levels [1]. It develops gradually and causes progressive damage to major organs such as the heart and kidneys, resulting in increased mortality and reduced quality of life [2]. The International Diabetes Federation (IDF) reports that approximately 537 million adults worldwide were living with diabetes in 2021, and this number is projected to rise to around 783 million by 2045 [3]. The growing prevalence of T2D presents a major global health challenge, placing significant social and economic strain on healthcare systems [4], [5].

Early detection and accurate risk prediction are essential for controlling the progression and complications of T2D, yet they remain difficult to achieve in practice. Many patients are diagnosed only after substantial organ damage has occurred, limiting the effectiveness of preventive interventions [6]. The rate of disease progression varies widely among individuals due to genetic, metabolic, lifestyle, and environmental differences [7]. This variability complicates efforts to forecast disease trajectories and hinders the personalization of treatment. Therefore, predictive models that can identify individuals at high risk of developing T2D are needed [8].

Traditional statistical approaches, such as survival analysis and Cox proportional hazards models, have provided valuable insights into diabetes risk factors [9]. However, these models depend on predefined mathematical assumptions and often fail to capture the nonlinear and multidimensional relationships present in real-world clinical data [10]. The increasing availability of detailed patient information through electronic health records (EHRs) has revealed the limitations of traditional approaches in exploiting data complexity and temporal dynamics [11].

Machine learning (ML), a subset of artificial intelligence (AI), offers an effective way to overcome these challenges. ML algorithms learn patterns directly from data without explicit programming and can model complex interactions among multiple variables [12], [13]. In diabetes research, ML has been applied to risk prediction, glycemic trend forecasting, and complication detection [14]. Compared with conventional models, ML methods can integrate large and diverse datasets, identify nonlinear relationships, and enhance predictive performance [15]. When applied to EHR data, ML enables dynamic, individualized risk prediction and supports a shift from reactive to preventive healthcare [16].

EHR data contain longitudinal information on patient demographics, laboratory results, diagnoses, prescriptions, and lifestyle indicators [17]. These records capture disease progression over time and allow ML models to learn from repeated observations. Integrating ML with EHR data has shown considerable promise in identifying subtle physiological and behavioral changes that precede clinical

diagnosis, enabling earlier intervention and improved patient outcomes [18].

A wide range of ML algorithms has been used for T2D prediction, including logistic regression, decision trees, random forests, support vector machines, neural networks, and ensemble methods such as gradient boosting [19], [20]. Deep learning architectures, particularly convolutional and recurrent neural networks, extend this capability by identifying hierarchical and temporal patterns in longitudinal data [21]. Although many models demonstrate strong predictive performance, challenges such as missing data, model interpretability, overfitting, and lack of external validation limit their translation into clinical practice [22].

Despite these advances, existing studies on ML-based prediction of T2D using EHR data remain fragmented and methodologically diverse. Variations in study design, validation strategies, and reporting standards make it difficult to compare findings or determine which approaches are most reliable for clinical application [23]. A systematic review is therefore warranted to consolidate current evidence, assess methodological rigor, and identify key gaps that limit clinical application.

A. Rationale

T2D is a multifactorial condition characterized by substantial variation in onset, progression, and outcomes. Conventional statistical methods, while valuable, have limited capacity to capture the complex and dynamic patterns present in clinical data. EHRs contain detailed longitudinal information that reflects the temporal evolution of patient health, creating new opportunities for building more accurate predictive models of disease risk. ML models can use such data to detect complex interactions, handle incomplete or heterogeneous information, and produce individualized risk predictions. Despite growing interest in this area, existing evidence remains dispersed across studies that differ in datasets, algorithms, and evaluation strategies, making comparisons difficult and conclusions uncertain. A systematic review focused specifically on ML models using EHR data for T2D prediction is therefore needed to consolidate current knowledge, evaluate methodological quality, and identify research gaps that must be addressed to advance clinically applicable predictive systems.

B. Objectives

This systematic review aims to evaluate existing research on the application of machine learning models for predicting T2D using EHR data.
The specific objectives are:

1. To identify the machine learning techniques used for predicting T2D using EHR data.
2. To evaluate the predictive performance of these models across reported studies.
3. To summarize the key predictors of T2D identified across studies.
4. To assess the methodological strengths and limitations that influence the reliability and clinical applicability of these models.

II. METHODOLOGY

This review followed a systematic and transparent approach to identify, select, and analyze studies that developed ML models for predicting T2D using HER data, in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

A. Eligibility Criteria

The selection of studies was guided by the PICOS (Population, Intervention, Comparison, Outcomes, and Study Design) framework to ensure a structured and transparent inclusion process.

1. *Population:* Adults aged 18 years and older, with or without a diagnosis of T2D, who's clinical, demographic, or laboratory data were available in EHRs.
2. *Intervention:* Studies that applied machine learning algorithms to predict the risk of T2D using EHR data.
3. *Comparison:* Studies comparing different machine learning algorithms for predicting T2D.
4. *Outcome:* Studies that reported predictive performance results for T2D prediction.
5. *Study Design:* Observational studies that developed or validated ML models using EHR data.

B. Inclusion Criteria

1. Adults aged 18 years and older.
2. Studies applying ML algorithms for T2D prediction using EHR data.
3. Studies reporting predictive performance metrics.
4. Peer-reviewed, full-text articles published in English between 2020 and 2025.

C. Exclusion Criteria

1. Studies on type 1 or gestational diabetes.
2. Studies not using ML methods or not based on EHR data.
3. Reviews, editorials, conference abstracts, or unpublished works.
4. Non-English articles or studies without full-text access.

D. Information Sources

A wide-ranging literature search was carried out using Scopus and PubMed to obtain a comprehensive collection of relevant studies for this systematic review. This approach ensured a broad and balanced representation of research from both biomedical and computational fields. The last search date was 19 October 2025.

E. Search Strategy

A comprehensive literature search was conducted to identify relevant studies on ML models developed for predicting T2D using EHR data. The search strategy combined predefined keywords with Boolean operators to capture all studies within this scope. The main search terms included type 2

diabetes, machine learning, deep learning, artificial intelligence, data mining, predictive model, classification model, supervised learning, and electronic health records. Filters were applied to restrict results to articles published between January 2020 and October 2025, written in English, and classified as original research.

A search of articles using the following query string on Scopus resulted in 204 documents. TITLE-ABS-KEY (("machine learning" OR "deep learning" OR "artificial intelligence" OR "data mining" OR "predictive model*" OR "classification model*" OR "supervised learning") AND ("type 2 diabetes" OR "T2DM" OR "diabetes mellitus type 2") AND ("electronic health record*" OR "EHR" OR "clinical record*" OR "medical record*" OR "healthcare data" OR "patient data") AND ("prediction" OR "risk assessment" OR "diagnosis" OR "prognosis")) AND PUBYEAR > 2019 AND PUBYEAR < 2026 AND (LIMIT-TO (SUBJAREA , "MEDI") OR LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (LANGUAGE , "English"))

A search of articles using the following query string on PubMed database resulted in 125 documents. (("machine learning"[All Fields] OR "deep learning"[All Fields] OR "artificial intelligence"[All Fields] OR "data mining"[All Fields] OR "predictive model*"[All Fields] OR "classification model*"[All Fields] OR "supervised learning"[All Fields]) AND ("type 2 diabetes"[All Fields] OR "T2DM"[All Fields] OR "diabetes mellitus type 2"[All Fields]) AND ("electronic health record*"[All Fields] OR "EHR"[All Fields] OR "clinical record*"[All Fields] OR "medical record*"[All Fields] OR "healthcare data"[All Fields] OR "patient data"[All Fields]) AND ("prediction"[All Fields] OR "risk assessment"[All Fields] OR "diagnosis"[All Fields] OR "prognosis"[All Fields])) AND ((medline[Filter]) AND (fha[Filter]) AND (humans[Filter]) AND (english[Filter]) AND (2020:2025[pdat]))

F. Data Management

All articles retrieved from the database searches were exported in RIS (Research Information Systems) and NBIB (National Library of Medicine Bibliographic) file formats from Scopus and PubMed, respectively, and imported into Rayyan, a web-based platform designed for systematic review management. The platform features an integrated artificial intelligence component that facilitates efficient article screening and organization.

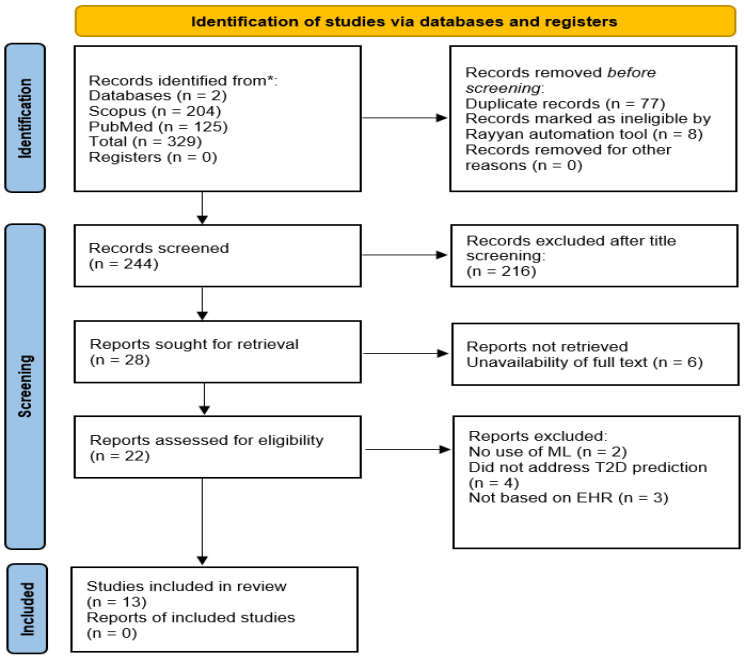

Fig.1 The Screened Studies Documented Using PRISMA Flow Diagram

G. Study Selection

The study selection process was conducted using Rayyan, a web-based tool for managing systematic reviews. Titles and abstracts were screened for relevance based on the inclusion criteria, and full-text articles of potentially eligible studies were assessed for final inclusion.

H. Data Extraction

The study's data collection process focused on gathering essential information from all included research articles. This involved extracting details such as study characteristics, ML algorithms applied, and key findings. Each study was examined carefully to ensure that the extracted information accurately reflected the reported methodology and outcomes. The purpose of this process was to compile reliable evidence on how ML models have been developed and applied for predicting T2D using EHR data.

I. Risk of Bias

A detailed assessment of potential bias was conducted to ensure the reliability of this review. As the sole reviewer, the evaluation followed recognized standards relevant to the included study designs. Studies that did not meet the eligibility criteria or showed poor alignment between objectives, methods, and outcomes were excluded. A broad search across multiple databases minimized selection bias, and any uncertainties were resolved through careful re-examination and reference to supporting literature.

III. RESULTS

A total of 329 studies were identified through database searches, comprising 204 from Scopus and 125 from PubMed. After removing 77 duplicates and 8 automatically ineligible records, 244 studies remained for title and abstract screening. During this stage, 216 studies were excluded for not meeting the inclusion criteria, leaving 28 studies for full-text review. Following detailed assessment, 6 studies could not be retrieved, and 9 were excluded for not meeting the eligibility requirements. Consequently, 13 studies met all inclusion criteria and were included in this systematic review.

The included studies, published between 2020 and 2025, examined the use of machine learning techniques for predicting type 2 diabetes using electronic health record data. Table I presents a summary of the included studies, outlining their algorithms, performance metrics, and key findings.

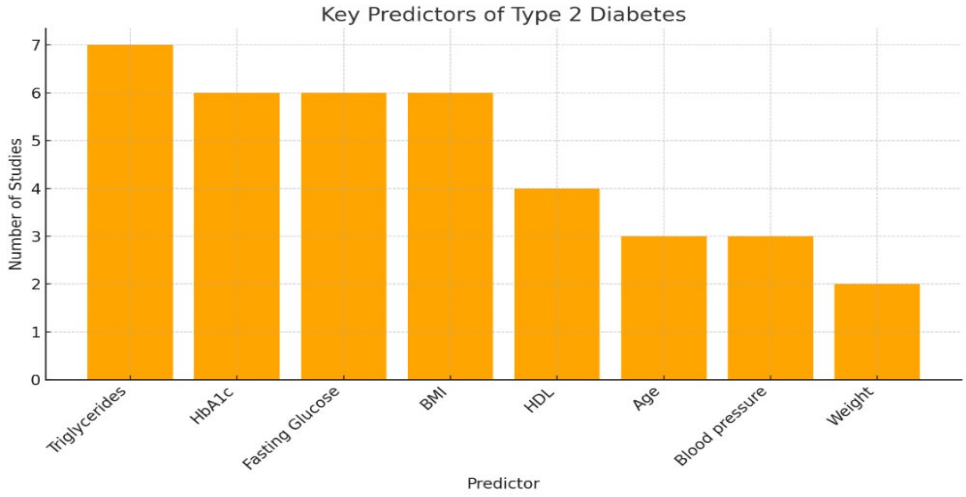

Fig.2 Frequency of Key Predictors of T2D Reported Across Included Studies

A. Summary of Findings

This review analyzed 13 studies published between 2020 and 2025 that applied machine learning to predict T2D using EHR data. The reviewed evidence shows a steady evolution in modeling techniques and methodological rigor. Earlier studies mainly used logistic regression and support vector machines because of their simplicity and interpretability [27], [29]. More recent work adopted ensemble methods such as random forest and XGBoost, which achieved stronger discrimination and handled complex relationships within clinical data more effectively [26], [28]. Deep learning models, including convolutional and recurrent neural networks, were also applied to longitudinal datasets, allowing them to capture temporal patterns that reflect disease progression. However, these models required large and diverse datasets to remain stable and often demanded substantial computational resources. Model performance varied across the reviewed studies. Ensemble and deep learning approaches generally performed better than single classifiers, particularly when supported by feature selection, data balancing, or augmentation methods. Studies that combined multiple data sources, such as laboratory results, clinical notes, and wearable sensor records, achieved improved robustness and reliability [24], [33]. Despite these advances, most models were validated only internally, with very few performing external validation using independent datasets. This limitation reduces confidence in their generalizability. In addition, variations in performance metrics and the absence of calibration analyses made it difficult to compare model reliability across studies. The studies also identified a group of predictors that appeared consistently and aligned with established evidence on diabetes risk. Triglycerides, HbA1c, fasting glucose, body mass index, high-density lipoprotein cholesterol, age, and blood pressure were the most frequently reported predictors. These findings demonstrate that machine learning can detect both established and emerging risk indicators of T2D.

Methodological quality varied across studies. Those using large datasets, addressing class imbalance, and clearly describing variable selection tended to achieve more stable and accurate outcomes [29], [30]. However, others provided limited detail on data preparation, handling of missing values, or bias management, reducing transparency and reproducibility. Deep learning studies rarely addressed computational efficiency or implementation feasibility [27]. Although explainable artificial intelligence improved interpretability in some models, inconsistent reporting and limited validation remain key challenges that restrict the clinical application of these approaches.

TABLE I SUMMARY OF INCLUDED STUDIES

S/N	Author(s)	Year	ML Technique(s)	Key Findings	Accuracy	AUC
1	Ding *et al.* [24]	2024	Large Language Multimodal Model (LLMM)	The proposed LLMM integrated textual clinical notes with laboratory data from 5 of EHRs to predict new-onset T2D. The multimodal approach outperformed unimodal ML models. It provided interpretable predictions using SHAP analysis, revealing clinically meaningful features such as glucose, HbA1c, hypertension, and age as key risk factors for early diabetes detection.	0.93	0.93
2	Shrestha *et al.* [25]	2022	Support Vector Machine (SVM) with Radial Base Function (RBF) Kernel and Long Short-Term Memory (LSTM) Layer	The SVM with RBF and LSTM layer model improved prediction of T2D onset using EHR datasets compared with standard ML models. Integrating temporal learning through the LSTM layer and nonlinear mapping via the RBF kernel enhanced classification accuracy and AUC while reducing processing time. The model achieved stable performance across multiple datasets, demonstrating its suitability for practical diabetes risk prediction	0.86	0.83
3	Bernstorff *et al.* [26]	2024	XGBoost and regularized Logistic Regression	The XGBoost model accurately predicted T2D onset in patients with mental illness using routine EHR data, identifying high-risk individuals up to 5 years before diagnosis. It outperformed logistic regression, showing stable performance across age and sex groups. Key predictive factors included HbA1c, triglycerides, weight, and HDL levels, reflecting strong alignment with known metabolic indicators of diabetes risk.	N/A	0.84
4	Alix *et al.* [27]	2021	Logistic Regression	A logistic regression model was developed using laboratory EHR data from over 13,000 Canadian patients to support an online diabetes risk prediction tool. The model identified fasting glucose, BMI, triglycerides, and HDL as the strongest predictors of T2D. It demonstrated reliable calibration and discrimination, enabling early risk estimation in primary care through an accessible, clinically interpretable web interface.	N/A	0.74
5	Liu *et al.* [28]	2025	Random Forest, Logistic Regression, and XGBoost	Using a 10-year EHR dataset of healthy adults, the XGBoost model achieved the highest performance for predicting future T2D incidence. The model identified HbA1c, fasting glucose, weight, free thyroxine (fT4), and triglycerides as the most influential predictors. The findings highlighted the previously underexplored role of thyroid hormone in diabetes risk assessment and demonstrated the clinical potential of ML-based prediction for early identification and prevention in primary care.	0.98	0.92
6	Bernardini *et al.* [29]	2020	Sparse Balanced Support Vector Machine (SB-SVM), Support Vector Machine, Random Forest, Decision Tree, K Nearest Neighbor, Logistic Regression, MultiLayer Perceptron, Deep Belief Network	The SB-SVM model predicted high-risk and potentially undiagnosed T2D cases from EHR data by managing class imbalance and high-dimensional features without additional feature selection. It achieved higher recall and AUC compared with standard ML and deep learning models. The model used sparsity to highlight clinically relevant predictors such as HbA1c, blood pressure, and lipid disorders, supporting clearer understanding of key factors associated with T2D risk.	N/A	0.91
7	García-Domínguez *et al.* [30]	2023	Random Forest with Generative Adversarial Network (GAN)	GANs were applied to generate synthetic clinical data for patients with and without T2D, addressing data scarcity in EHR-based research. The augmented dataset was used to train a Random Forest model, which achieved improved diagnostic performance compared with models trained on real data alone. The study demonstrated that GAN-based data augmentation enhanced classification robustness and sensitivity while maintaining data integrity, suggesting its potential for strengthening ML-driven T2D diagnosis in limited clinical datasets.	0.96	0.96
8	Kang *et al.* [31]	2025	Logistic Regression-based nomogram with LASSO feature selection	Using EHR data from patients with Chronic Obstructive Pulmonary Disease, LASSO regression identified 7 significant predictors of T2D, including PCO_2, neutrophil count, C reactive protein, erythrocyte sedimentation rate, bilirubin, triglycerides, and BMI. These variables were used to develop a Logistic Regression model that showed good calibration, discrimination, and stable performance after internal validation.	N/A	0.80
9	Perveen *et al.* [32]	2020	Hybrid Hidden Markov Model enhanced with Newton's Divided Difference Method (HMM–NDDM)	The study developed a hybrid HMM NDDM model using EHR data to predict the risk of developing T2D over multiple time periods. The NDDM component addressed irregular and sparse clinical data before HMM training. Logistic regression was used to identify key predictors including HbA1c, fasting glucose, triglycerides, HDL, LDL, and BMI. The hybrid model achieved higher discrimination than the standard HMM across all prediction horizons, showing reliable estimation of diabetes risk in longitudinal patient records.	N/A	0.81
10	Ha *et al.* [33]	2025	Ensemble model combining XGBoost, CNN, LSTM with GAN-based data augmentation	A stacking ensemble combining XGBoost, CNN, and LSTM improved early diabetes prediction using EHR, hospital, and wearable data. GAN based augmentation reduced data imbalance, and SHAP analysis identified glucose, BMI, C peptide, insulin, age, and blood pressure as key predictors. The model achieved higher accuracy and interpretability than single algorithms.	0.90	0.92
11	Deberneh and Kim [34]	2021	Logistic Regression, Random Forest, Support Vector Machine, XGBoost, and Ensemble Models (Soft Voting, Stacking)	Using 6 years of EHR data, features were selected through ANOVA, chi squared tests, and recursive feature elimination. The Soft Voting ensemble achieved the best performance, identifying fasting plasma glucose, HbA1c, triglycerides, BMI, gamma GTP, age, and uric acid as key predictors. It showed improved discrimination and stability for one-year T2D prediction compared with single models.	0.73	N/A

12	Al-Hussein *et al.* [35]	2025	Multiple Linear Regression, Artificial Neural Network, Random Forest, Support Vector Regression, Decision Tree Regression	The Random Forest model achieved the best performance for predicting the age at onset of T2D using EHR data from 1,000 patients in Saudi Arabia. Feature importance analysis identified triglycerides, total cholesterol, HDL, BMI, systolic blood pressure, white blood cell count, ferritin, and vitamin D as key predictors.	0.97	N/A
13	Muthu and Suriya [36]	2023	K-Nearest Neighbor	The study applied the KNN algorithm to EHR and clinical datasets, including the Pima Indian Diabetes dataset, to classify individuals as high or low risk for T2D. Data were preprocessed through normalization, outlier removal, and k fold validation to improve reliability. The model demonstrated stable predictive performance, showing that KNN can effectively identify T2D risk when datasets are well scaled and balanced.	0.70	N/A

IV. DISCUSSION

The findings of this review indicate that ML has improved the prediction of T2D using EHR data. Ensemble and deep learning models achieved higher predictive accuracy than traditional algorithms because they can capture complex and time-dependent relationships among clinical variables. These models were able to identify nonlinear interactions that conventional statistical techniques often overlook. Despite these improvements in predictive performance, their application in clinical settings remains limited. Most studies relied on internal validation, in which models were tested on the same datasets used for their development. This limits confidence in how well they would perform in new populations or across different healthcare systems. Several studies also omitted calibration analysis, which assesses whether predicted probabilities align with observed outcomes. Without this evaluation, a model may display good discrimination yet provide inaccurate estimates of patient risk. The absence of consistent evaluation metrics and transparent performance reporting further constrains comparability across studies and limits the assessment of reliability.

Overall, the reviewed evidence suggests that the increase in predictive accuracy has not been accompanied by comparable methodological consistency or readiness for clinical implementation. Stronger external validation using independent datasets is required to confirm model stability and generalizability. Standardized reporting of performance metrics and calibration results would enable meaningful comparison between studies. Attention to model interpretability and clarity in describing data processing procedures is also essential to improve transparency and trust among clinical users. Collaborative development between data scientists and healthcare practitioners will be critical to producing ML models that are accurate, reproducible, and suitable for integration into healthcare practice.

V. CONCLUSION

This systematic review examined the use of machine learning models for predicting type 2 diabetes mellitus using electronic health record data, highlighting that ensemble and deep learning approaches demonstrated stronger predictive performance than traditional methods because they can capture complex and nonlinear clinical relationships. The findings revealed common risk indicators such as fasting glucose, HbA1c, triglycerides, and body mass index, confirming their consistent relevance across models. However, major methodological limitations were identified, including limited external validation, inconsistent performance reporting, and insufficient model calibration, which restrict confidence in reliability and hinder clinical translation. To advance practical application, future research should emphasize methodological standardization, transparent evaluation, inclusion of diverse populations, and the use of interpretable machine learning techniques to improve clarity and clinical trust. The study establishes that machine learning applied to electronic health records has strong potential for early diabetes risk prediction but requires rigorous validation and transparency to support dependable use in healthcare decision-making.

Declaration of Conflicting Interests
The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

Use of Artificial Intelligence (AI)-Assisted Technology for Manuscript Preparation
The authors confirm that no AI-assisted technologies were used in the preparation or writing of the manuscript, and no images were altered using AI.

ORCID
Oduware C. Odigie https://orcid.org/0009-0009-6544-2460
Folasade Y. Ayankoya http://orcid.org/0000-0003-0308-2753
Shade O. Kuyoro http://orcid.org/0000-0001-7235-7744
Ayodeji G. Abiodun https://orcid.org/0009-0000-1135-2907

REFERENCES

[1] U. Galicia-Garcia *et al.*, "Pathophysiology of type 2 diabetes mellitus," *Int. J. Mol. Sci.*, vol. 21, no. 17, pp. 1–34, 2020, doi: 10.3390/ijms21176275.
[2] C. Gavina *et al.*, "Premature mortality in type 2 diabetes mellitus associated with heart failure and chronic kidney disease: 20 years of real-world data," *J. Clin. Med.*, vol. 11, no. 8, p. 2131, Apr. 2022, doi: 10.3390/jcm11082131.
[3] H. Sun *et al.*, "IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 183, p. 109119, Jan. 2022, doi: 10.1016/j.diabres.2021.109119.
[4] C. H. Karugu *et al.*, "The economic burden of type 2 diabetes on the public healthcare system in Kenya: A cost of illness study," *BMC Health Serv. Res.*, vol. 24, no. 1, pp. 1–11, Dec. 2024, doi: 10.1186/s12913-024-11700-x.
[5] E. D. Parker *et al.*, "Economic costs of diabetes in the U.S. in 2022," *Diabetes Care*, vol. 47, no. 1, pp. 26–43, 2024, doi: 10.2337/dci23-0085.
[6] J. Zhang, Z. Zhang, K. Zhang, X. Ge, R. Sun, and X. Zhai, "Early detection of type 2 diabetes risk: Limitations of current diagnostic

criteria," *Front. Endocrinol.*, vol. 14, pp. 1–7, 2023, doi: 10.3389/fendo.2023.1260623.

[7] P. W. Franks and J. L. Sargent, "Diabetes and obesity: Leveraging heterogeneity for precision medicine," *Eur. Heart J.*, vol. 45, no. 48, pp. 5146–5155, 2024, doi: 10.1093/eurheartj/ehae746.

[8] A. Cahn *et al.*, "Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model," *Diabetes Metab. Res. Rev.*, vol. 36, no. 2, p. e3252, Feb. 2020, doi: 10.1002/dmrr.3252.

[9] Y. Edlitz and E. Segal, "Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards," *eLife*, vol. 11, pp. 1–24, 2022, doi: 10.7554/eLife.71862.

[10] S. Chen, J. Yu, S. Chamouni, Y. Wang, and Y. Li, "Integrating machine learning and artificial intelligence in life-course epidemiology: Pathways to innovative public health solutions," *BMC Med.*, vol. 22, no. 1, 2024, doi: 10.1186/s12916-024-03566-x.

[11] H. J. A. van Os *et al.*, "Developing clinical prediction models using primary care electronic health record data: The impact of data preparation choices on model performance," *Front. Epidemiol.*, vol. 2, pp. 1–8, 2022, doi: 10.3389/fepid.2022.871630.

[12] L. P. Nguyen *et al.*, "The utilization of machine learning algorithms for assisting physicians in the diagnosis of diabetes," *Diagnostics*, vol. 13, no. 12, 2023, doi: 10.3390/diagnostics13122087.

[13] H. A. Aliyu *et al.*, "Optimizing machine learning algorithms for diabetes data: A metaheuristic approach to balancing and tuning classifier parameters," *Franklin Open*, vol. 8, p. 100153, 2024, doi: 10.1016/j.fraope.2024.100153.

[14] W. Gong *et al.*, "Deep learning for enhanced prediction of diabetic retinopathy: A comparative study on the diabetes complications data set," *Front. Med.*, vol. 12, 2025, doi: 10.3389/fmed.2025.1591832.

[15] Y. Ye *et al.*, "Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: A retrospective cohort study," *J. Diabetes Res.*, vol. 2020, p. 4168340, 2020, doi: 10.1155/2020/4168340.

[16] T. Feng *et al.*, "Machine learning-based clinical decision support for infection risk prediction," *Front. Med.*, vol. 10, pp. 1–12, 2023, doi: 10.3389/fmed.2023.1213411.

[17] N. Huguet *et al.*, "Using electronic health records in longitudinal studies: Estimating patient attrition," *Med. Care*, vol. 58, suppl. 6, pp. S46–S54, Jun. 2020, doi: 10.1097/MLR.0000000000001298.

[18] R. Grout *et al.*, "Predicting disease onset from electronic health records for population health management: A scalable and explainable deep learning approach," *Front. Artif. Intell.*, vol. 6, p. 1287541, 2023, doi: 10.3389/frai.2023.1287541.

[19] S. G. Choi *et al.*, "Comparisons of prediction models for undiagnosed diabetes using machine learning versus traditional statistical methods," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-023-40170-0.

[20] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che, "Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model," *Healthcare*, vol. 8, no. 3, pp. 1–11, 2020, doi: 10.3390/healthcare8030247.

[21] H. Javidi *et al.*, "Identification of robust deep neural network models of longitudinal clinical measurements," *npj Digit. Med.*, vol. 5, no. 1, p. 106, 2022, doi: 10.1038/s41746-022-00651-4.

[22] H. Lee *et al.*, "Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts," *eClinicalMedicine*, vol. 80, p. 103069, 2025, doi: 10.1016/j.eclinm.2025.103069.

[23] P. T. Phuc *et al.*, "Early detection of dementia in populations with type 2 diabetes: Predictive analytics using a machine learning approach," *J. Med. Internet Res.*, vol. 26, p. e52107, 2024, doi: 10.2196/52107.

[24] J.-E. Ding *et al.*, "Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records," *Sci. Rep.*, vol. 14, no. 1, p. 20774, 2024, doi: 10.1038/s41598-024-71020-2.

[25] M. Shrestha *et al.*, "A novel deep learning solution enhancing support vector machines for predicting the onset of type 2 diabetes," *Multimed. Tools Appl.*, vol. 82, no. 4, pp. 6221–6241, 2023, doi: 10.1007/s11042-022-13582-9.

[26] M. Bernstorff *et al.*, "Development and validation of a machine learning model for prediction of type 2 diabetes in patients with mental illness," *Acta Psychiatr. Scand.*, vol. 151, no. 3, pp. 245–258, 2025, doi: 10.1111/acps.13687.

[27] G. Alix *et al.*, "An online risk tool for predicting type 2 diabetes mellitus," *Diabetology*, vol. 2, no. 3, pp. 123–129, 2021, doi: 10.3390/diabetology2030011.

[28] Y.-Q. Liu *et al.*, "Use of machine learning to predict the incidence of type 2 diabetes among relatively healthy adults: A 10-year longitudinal study in Taiwan," *Diagnostics*, vol. 15, no. 1, 2025, doi: 10.3390/diagnostics15010072.

[29] M. Bernardini *et al.*, "Discovering type 2 diabetes in electronic health records using the sparse balanced support vector machine," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 1, pp. 235–246, 2020, doi: 10.1109/JBHI.2019.2899218.

[30] A. García-Dominguez *et al.*, "Optimizing clinical diabetes diagnosis through generative adversarial networks: Evaluation and validation," *Diseases*, vol. 11, no. 4, 2023, doi: 10.3390/diseases11040134.

[31] X. Kang *et al.*, "Construction and validation of a prediction model for developing type 2 diabetes mellitus in patients with chronic obstructive pulmonary disease," *Front. Endocrinol.*, vol. 16, 2025, doi: 10.3389/fendo.2025.1560631.

[32] S. Perveen *et al.*, "A hybrid approach for modeling type 2 diabetes mellitus progression," *Front. Genet.*, vol. 10, 2020, doi: 10.3389/fgene.2019.01076.

[33] H.-H. Ha *et al.*, "Diabetes early prediction using machine learning and ensemble methods," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 15, no. 2, pp. 363–375, 2025, doi: 10.18517/ijaseit.15.2.20947.

[34] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, 2021, doi: 10.3390/ijerph18063317.

[35] F. Al-Hussein *et al.*, "Predicting type 2 diabetes onset age using machine learning: A case study in Saudi Arabia," *PLoS One*, vol. 20, no. 2, p. e0318484, 2025, doi: 10.1371/journal.pone.0318484.

[36] S. Suriya and J. J. Muthu, "Type 2 diabetes prediction using the k-nearest neighbor algorithm," *J. Trends Comput. Sci. Smart Technol.*, vol. 5, no. 2, pp. 190–205, 2023, doi: 10.36548/jtcsst.2023.2.007.