Journal of Engineering Applications 12[33]1-7: e21233107
Article

Predictive model for Type 2 Diabetes Risk using Artificial Intelligence and multivariate data

Modelo predictivo de riesgo de DM2 usando Inteligencia Artificial y datos multivariados

Rodríguez-Silva, Jesús Rolando [a], Pérez-Briones, Nancy Griselda [b], Ventura-Sobrevilla, Janeth Margarita [c] and Boone-Villa, Victor Daniel * [d]

[a] ROR Universidad Autónoma de Coahuila • 0000-0001-8410-9174• 492229
[b] ROR Universidad Autónoma de Coahuila • 0000-0001-6903-4039 • 370311
[c] ROR Universidad Autónoma de Coahuila • 0000-0001-6304-5749 • 207772
[d] ROR Universidad Autónoma de Coahuila • 0000-0003-4220-7926 • 171109

Classification:

Area: Engineering
Field: Engineering
Discipline: Systems Engineering
Subdiscipline: Automation

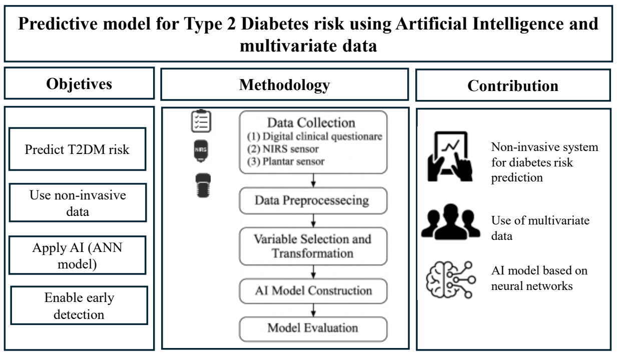
https://doi.org/10.35429/JEA.2025.12.33.2.1.7
History of the article:
Received: March 30, 2025
Accepted: August 30, 2025

Check for updates

* ✉ [danielboone@uadec.edu.mx]

Abstract

Type 2 Diabetes Mellitus (T2DM) is a global public health concern due to its growing prevalence and silent progression. This study presents the development and validation of a predictive model based on artificial neural networks (ANN) using multivariate, non-invasive data. The model integrates three data sources: near-infrared spectroscopy (NIRS) for estimated glucose levels, plantar pressure sensors (FSR) for neuropathy indicators, and a digital lifestyle questionnaire (IMEVID). After preprocessing and training using a stratified dataset of adults without previous diagnosis, the ANN achieved 82 % accuracy, with high sensitivity and a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.89. Compared to traditional methods like logistic regression, the proposed model demonstrated superior performance in classifying individuals at high or low risk. These results demonstrate the potential of non-invasive, AI-based screening tools to identify early-stage metabolic dysfunction, enabling timely preventive interventions. This approach may be especially useful in community-based or resource-limited settings to support public health decision-making.

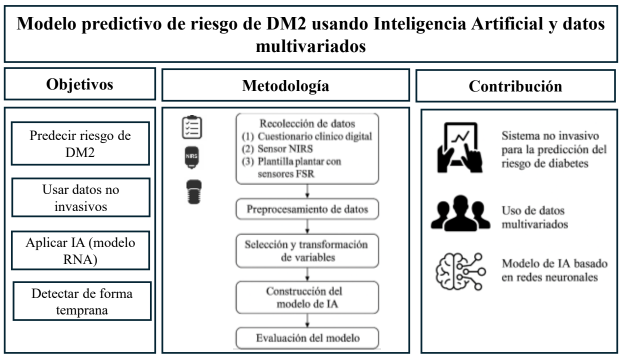


Type 2 Diabetes mellitus, artificial intelligence, non-invasive prediction

Area: Advocacy and attention to national problems

Resumen

La diabetes mellitus tipo 2 (DM2) es un problema de salud pública a nivel mundial debido a su creciente prevalencia y progresión silenciosa. Este estudio presenta el desarrollo y validación de un modelo predictivo basado en redes neuronales artificiales (RNA), utilizando datos multivariados y no invasivos. El modelo integra tres fuentes de datos: espectroscopía de infrarrojo cercano (NIRS) para estimar niveles de glucosa, sensores de presión plantar (FSR) como posibles indicadores de neuropatía y un cuestionario digital de estilo de vida (IMEVID). Tras el preprocesamiento y el entrenamiento con un conjunto de datos estratificado de adultos sin diagnóstico previo, la RNA alcanzó una precisión del 82 %, con alta sensibilidad y un área bajo la curva ROC (AUC, por sus siglas en inglés) de 0.89. En comparación con métodos tradicionales como la regresión logística, el modelo propuesto demostró un mejor desempeño en la clasificación de individuos con riesgo alto o bajo. Estos resultados respaldan el potencial de herramientas de tamizaje no invasivas basadas en IA para intervenciones preventivas oportunas.


Modelo predictivo de riesgo de DM2 usando Inteligencia Artificial y datos multivariados
Objetivos
Metodología
Contribución
Predecir riesgo de DM2
Usar datos no invasivos
Aplicar IA (modelo RNA)
Detectar de forma temprana
NIRS
Recolección de datos
(1) Cuestionario clínico digital
(2) Sensor NIRS
(3) Plantilla plantar con sensores FSR
Preprocesamiento de datos
Selección y transformación de variables
Construcción del modelo de IA
Evaluación del modelo
Sistema no invasivo para la predicción del riesgo de diabetes
Uso de datos multivariados
Modelo de IA basado en redes neuronales

Diabetes mellitus tipo 2, Inteligencia artificial, predicción no invasiva

Citation: Rodríguez-Silva, Jesús Rolando, Pérez-Briones, Nancy Griselda, Ventura-Sobrevilla, Janeth Margarita and Boone-Villa, Victor Daniel. [2025]. Predictive model for Type 2 Diabetes Risk using Artificial Intelligence and multivariate data. Journal of Engineering Applications. 12[33]1-7: e21233107.

RENIECYT
Registro Nacional de Instituciones y Empresas Científicas y Tecnológicas
1702902 SECIHTI

Introduction

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disorder characterized by persistent hyperglycemia due to defects in insulin secretion, insulin action, or both. It currently represents one of the leading causes of morbidity and mortality worldwide, with an alarming increase in low- and middle-income countries [World Health Organization, 2016]. According to the International Diabetes Federation (IDF), the number of affected adults is projected to reach 783 million by the year 2045, highlighting the urgent need for effective prevention strategies [International Diabetes Federation, 2021].

Traditionally, the diagnosis of T2DM has been based on clinical tests such as fasting plasma glucose (FPG), oral glucose tolerance test (OGTT), and glycated hemoglobin (HbA1c) [American Diabetes Association, 2023].

While these tests are useful for confirming diabetic status, they have limitations for early risk detection, as they tend to identify the disease in more advanced stages of metabolic dysfunction [Bonora & Tuomilehto, 2011; Guerrero-Romero & Rodríguez-Morán, 2005]. This has driven the search for alternative methods capable of predicting the risk of T2DM at earlier stages, when interventions can be more effective.

In this context, the growing availability of biomedical data—from electronic health records and medical imaging to wearable and ambient biosensors—has laid the foundation for the application of artificial intelligence (AI) and machine learning in health prediction tasks. These technologies enable the development of multimodal models capable of capturing complex interactions among clinical, physiological, and behavioral variables, surpassing the limitations of traditional statistical approaches [Acosta et al., 2022].

The use of supervised and unsupervised algorithms enables the modeling of complex interaction patterns among clinical, physiological, and lifestyle variables, surpassing the limitations of traditional statistical methods [Beam & Kohane, 2018].

Recent studies have demonstrated the effectiveness of techniques such as logistic regression, artificial neural networks (ANN), and unsupervised clustering (K-means) in predicting the risk of chronic diseases, including T2DM [Kavakiotis et al., 2017; Lai et al., 2019].

Moreover, recent advances in non-invasive technologies have enabled the detection of glucose-related biomarkers through near-infrared (NIR) spectroscopy. Han et al. [2017] proposed a differential correction method designed to reduce background spectral variations—such as those caused by light source drift, temperature, and skin interface—thus improving the accuracy of glucose measurement using NIR signals.

Among these non-invasive approaches, plantar pressure analysis has gained increasing attention due to its potential to detect early-stage diabetic foot complications. A recent study demonstrated that smart insoles embedded with pressure sensors, combined with machine learning algorithms, can accurately distinguish between healthy and diabetic feet during dynamic activities such as walking. This capability not only improves early risk stratification but also opens the door to real-time monitoring and preventive interventions in home settings [Agrawal et al., 2024].

This paper presents the construction and validation of a multivariate predictive model based on machine learning techniques, using data obtained through non-invasive methods and demographic characteristics associated with T2DM risk.

The main objective is to evaluate the technical feasibility of an early prediction system that can be implemented in community screening strategies or personalized preventive monitoring.

Methodology

This study is classified as applied and quantitative research, with an observational, cross-sectional, and predictive design, focused on the development of an artificial intelligence model to estimate T2DM risk using multivariate data collected through non-invasive sensors and a digital clinical questionnaire.

The methodology was structured into five stages: data collection, preprocessing, variable selection, model construction, and performance evaluation.

1. Data collection

The target population consisted of adult individuals aged between 30 and 50 years, without a previous diagnosis of T2DM, residing in the state of Coahuila, Mexico. Sampling was non-probabilistic by convenience, and data collection was carried out in a controlled environment at the Faculty of Mechanical and Electrical Engineering (FMEE) of the Autonomous University of Coahuila. Participants were evaluated using a structured clinical questionnaire and sensory devices designed to capture key physiological information.

Multivariate data were obtained from three sources, as follows:

- A modified version of the digital IMEVID questionnaire, collecting information on family history, dietary habits, physical activity level, substance use, and clinical symptoms.
- A NIRS sensor, used to estimate capillary glucose levels.
- A plantar insole equipped with FSR sensors, designed to measure pressure distribution in key foot areas as an indicator of peripheral neuropathy.

The full data acquisition and processing flow is represented in Figure 1, which graphically summarizes the components of the implemented predictive system.

Box 1

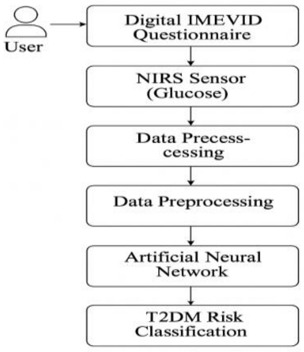


Figure 1

Data capture and analysis system diagram for T2DM risk prediction

Source: Own elaboration

2. Data preprocessing

The collected data were organized and prepared for analysis using Python tools (pandas, NumPy, and scikit-learn). Categorical variables such as sex and presence of symptoms were encoded using LabelEncoder, while numerical variables like glucose and plantar pressure were standardized using StandardScaler. Incomplete records were removed, and consistency and cleanliness of the final dataset were verified.

Subsequently, the data were split into two subsets: 80% for training and 20% for testing, using the train_test_split function with a fixed random seed, to ensure reproducibility and evaluate the model under objective conditions.

3. Variable selection and transformation

Based on the collected variables, a continuous output variable was generated to represent the T2DM risk index. This variable was constructed from a weighted combination of the IMEVID score, estimated glucose values, and plantar sensor readings. It was then transformed into a binary category ("low risk" and "high risk") using the average of the training set as the threshold.

Box 2

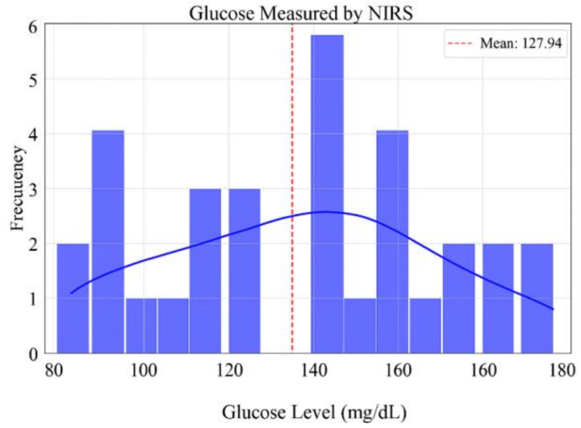


Figure 2

Glucose distribution for T2DM risk prediction

Source: Own elaboration

4. AI Model construction

The model was built using a multilayer artificial neural network (ANN) programmed in Keras.

The architecture consisted of:

- An input layer with a number of neurons equal to the number of variables.

- Three dense hidden layers with ReLU activation (50, 30, and 15 neurons respectively).

An output layer with 3 neurons and Softmax activation for multiclass classification (no risk, tendency, confirmed diagnosis).

The model was trained using the Adam optimizer and the binary_crossentropy loss function. To prevent overfitting, EarlyStopping was implemented, stopping training when the validation loss did not improve for 10 consecutive epochs. The final model was saved in HDF5 format under the name modelo_diabetes.h5, allowing future reuse without retraining.

Box 3

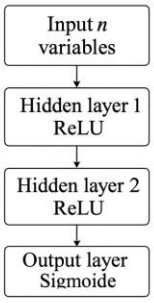


Figure 3

Neural network architecture diagram used in the model.

Source: Own elaboration

5. Model evaluation

The model's performance was evaluated using two approaches. First, regression metrics such as Mean Squared Error (MSE) and the coefficient of determination (R^2) were calculated. Second, the predicted values were transformed into binary labels to perform a classification analysis based on accuracy, sensitivity, specificity, confusion matrix, and ROC Area Under the Curve (AUC). These metrics were compared with the results of a logistic regression model used as a baseline, highlighting the advantage of the neural network approach for this type of multivariate problem.

Additionally, to visually illustrate the model's behavior during evaluation, Figure 4 presents the ROC curve generated from the test set, showing the model's ability to discriminate between the "low risk" and "high risk" classes, with an AUC close to 0.89, thus quantitatively supporting its predictive performance.

Box 4

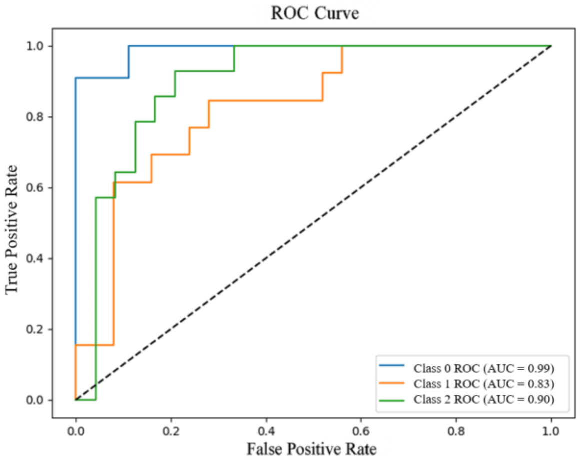


Figure 4

ROC curve of the binary predictive model for T2DM risk

Source: Own elaboration

Results

After the training and validation process, the ANN-based model achieved strong overall performance in classifying the risk of developing T2DM. The implemented architecture, designed with three hidden layers and an output layer using the Softmax function, was trained using multivariate data from non-invasive sensors (NIRS and FSR) and the IMEVID digital clinical questionnaire.

The model was evaluated through five-fold cross-validation and compared with conventional algorithms such as logistic regression, support vector machines (SVM), and random forests.

The ANN exhibited the best performance, achieving a mean squared error (MSE) of 0.015 and a sensitivity of 0.729. These metrics support its selection as the final model, offering a superior balance between accuracy and early risk detection capacity.

In multiclass classification, the model identified three risk categories: no apparent risk (class 0), tendency toward T2DM (class 1), and compatible diagnosis with T2DM (class 2). The detailed performance by class is summarized in Table 1, with an overall accuracy of 82 %, recall of 82 %, and F1-score of 0.81, indicating a solid balance between identifying positive cases and reducing false positives.

Box 5

Table 1

Class-wise results of the ANN model

Class	Precision	Recall	F1-score
0 – no apparent risk	0.91	0.91	0.91
1 – tendency toward T2DM	0.80	0.62	0.70
2 – compatible diagnosis with T2DM	0.76	0.93	0.84

Source: Own elaboration

The model's discrimination capacity was confirmed through ROC curve analysis, which showed AUC values of 0.99 for the no-risk class, 0.83 for the intermediate class, and 0.90 for the diagnosed T2DM class (Figure 4). These values reflect the model's excellent ability to distinguish between different risk profiles.

The confusion matrix in Figure 5 revealed accurate classification in classes 0 and 2, while classification errors were concentrated in class 1. From a preventive perspective, this misclassification is even favorable, as it tends to classify individuals with borderline profiles as positive, thereby fulfilling the system's early detection objective.

Box 6

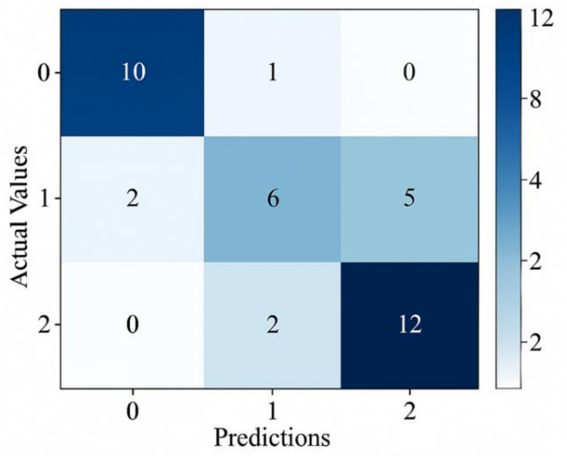


Figure 5

Confusion matrix of the model

Source: Own elaboration

The results obtained are consistent with the findings of Agrawal et al. [2024], who demonstrated that plantar pressure sensors combined with machine learning algorithms enable accurate differentiation between diabetic and healthy feet, highlighting their usefulness for real-time monitoring.

Similarly, the multivariable approach used in this study aligns with that proposed by Lai et al. [2019], where the integration of clinical, demographic, and lifestyle data through supervised models significantly improved predictive capacity. This convergence indicates that the model developed in the present research is aligned with current best practices and reinforces its validity for real-world applications. Furthermore, unlike traditional studies based solely on clinical or biochemical parameters, the present work incorporates real-time physiological inputs captured through wearable plantar pressure sensors, offering a novel approach for early detection outside of clinical settings [Totaganti et al., 2023].

Conclusions

This study demonstrated the feasibility of implementing a multivariate predictive model supported by artificial intelligence to estimate the risk of developing Type 2 Diabetes Mellitus (T2DM) using non-invasive data. The model, based on an artificial neural network (ANN) and fed with information obtained from NIRS and FSR sensors, along with the IMEVID questionnaire, achieved competitive performance—with an overall accuracy of 82% and appropriate discrimination across different risk levels.

The results confirm that the combination of emerging biomarkers and lifestyle variables improves early detection of T2DM risk. The model's accurate classification of individuals with no risk and those with a confirmed diagnosis, along with a consistent tendency to identify intermediate-stage subjects as positive, reinforces the system's utility in preventive contexts. Moreover, the cross-validation applied to the model and its comparison with other algorithms (such as logistic regression, SVM, and random forests) showed that the proposed ANN offers a more robust solution to the complexity of multivariate data. This approach represents a step forward toward accessible, personalized, and non-invasive screening tools with potential for implementation in community settings or as a complement to public health strategies.

For future work, it is recommended to expand the population sample, incorporate new physiological variables, and explore more complex architectures to further increase model precision.

Declarations

Conflict of interest

The authors declare no interest conflict. They have no known competing financial interests or personal relationships that could have appeared to influence the article reported in this article.

Author contribution

Rodríguez-Silva, Jesús Rolando: Conceived the project, designed and implemented the neural network architecture, developed and integrated the non-invasive sensors, and led the overall execution of the predictive system.

Pérez-Briones, Nancy Griselda: Collaborated in the design and adaptation of the IMEVID digital questionnaire used for lifestyle data collection.

Ventura-Sobrevilla, Janeth Margarita: Conducted statistical analysis, assisted in model evaluation, and contributed to the interpretation of results and performance metrics.

Boone-Villa, Victor Daniel: Contributed to the manuscript writing and review, prepared the figures and diagrams, and supported the comparative analysis with baseline models.

Availability of data and materials

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Due to privacy considerations, raw data containing personal or sensitive information have been anonymized prior to analysis.

Funding

This research was partially funded through the project FONCYT-COAH-2022-C19-C020, titled "Desarrollo de un dispositivo auxiliar no invasivo para la medición de presión plantar en el diagnóstico preventivo de la diabetes tipo 2 en adultos coahuilenses", supported by the Fondo para el Desarrollo de la Ciencia y la Tecnología del Estado de Coahuila (FONCYT), under the 2022 Call – Modality C, granted by the Consejo Estatal de Ciencia y Tecnología de Coahuila (COECYT). The funding supported the development and integration of the plantar pressure sensor system based on Force Sensing Resistor (FSR) technology.

Acknowledgements

We thank the Universidad Autónoma de Coahuila for the support and facilities provided for the completion of this project.

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
ANN	Artificial Neural Network
T2DM	Type 2 Diabetes Mellitus
FPG	Fasting Plasma Glucose
FSR	Force Sensing Resistor
HbA1c	Glycated Hemoglobin
IMEVID	Instrument for Measuring Lifestyle in Diabetics
MSE	Mean Squared Error
NIRS	Near-Infrared Spectroscopy
OGTT	Oral Glucose Tolerance Test
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
R^2	Coefficient of Determination
SVM	Support Vector Machine

References

Antecedents

World Health Organization. [2016]. Global report on diabetes.

International Diabetes Federation. [2021]. IDF diabetes atlas (10th ed.).

Basics

American Diabetes Association. [2023]. Classification and diagnosis of diabetes. Diabetes Care, 46(Suppl. 1), S19–S40.

Bonora, E., & Tuomilehto, J. [2011]. The pros and cons of diagnosing diabetes with A1C. *Diabetes care*, *34 Suppl 2*(Suppl 2), S184–S190.

Guerrero-Romero, F., & Rodríguez-Morán, M. [2005]. Diagnosis of diabetes mellitus. Revista Médica del IMSS, 43[4], 325–332.

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. [2022]. Multimodal biomedical AI. *Nature Medicine, 28*[11], 1773–1784.

Beam, A. L., & Kohane, I. S. [2018]. Big data and machine learning in health care. JAMA, 319[13], 1317–1318.

Supports

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

Han, G., Yu, X., Xia, D., Liu, R., Liu, J., & Xu, K. [2017]. Preliminary Clinical Validation of a Differential Correction Method for Improving Measurement Accuracy in Noninvasive Blood Glucose Using Near-Infrared Spectroscopy. Applied Spectroscopy, 71[9], 2177–2186.

Differences

Totaganti, N. V. S. K., Basha, S. M., Kumar, D. P., & Rao, G. V. P. [2023]. Prediction of type-2 diabetes mellitus using supervised machine learning algorithms. *Materials Today: Proceedings*, 84[Part 6], 2036–2041.

Discussions

Agrawal, D. K., Jongpinit, W., Pojprapai, S., Usaha, W., Wattanapan, P., Tangkanjanavelukul, P., & Vitoonpong, T. [2024]. Smart Insole-Based Plantar Pressure Analysis for Healthy and Diabetic Feet Classification: Statistical vs. Machine Learning Approaches. *Technologies*, *12*[11], 231.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. [2019]. Predictive models for diabetes mellitus using machine learning techniques. *BMC endocrine disorders*, *19*[1], 101.