Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

Machine learning applied to infrared spectra of saliva for non-invasive diabetes detection

Aprendizaje automático aplicado a espectros infrarrojos de saliva para la detección no invasiva de la diabetes

Rosales-Martínez, O.[a], Granda-Gutiérrez, E.E.[b]*, Matías-Mendoza, G.A.[a], García-Hernández, R.A.[a], Flores-Fuentes, A.A.[b]

[a] Centro Universitario UAEM Tianguistenco; Universidad Autónoma del Estado de México; C.P. 52640, Tianguistenco, México. ORCID: https://orcid.org/0000-0002-6889-6355, https://orcid.org/0000-0003-1052-0619, https://orcid.org/0000-0001-7941-377X.

[b] Centro Universitario UAEM Atlacomulco; Universidad Autónoma del Estado de México; C.P. 50450, Atlacomulco, México. ORCID: https://orcid.org/0000-0002-9316-9627*, https://orcid.org/0000-0002-6233-1708.

orosalesm001@alumno.uaemex.mx; eegrandag@uaemex.mx*; gmatiasm@uaemex.mx; reagarciah@uaemex.mx; aafloresf@uaemex.mx

Technological innovation: Automated, non-invasive identification of salivary biomarkers for detecting type 2 diabetes mellitus using FTIR spectroscopy and machine learning.

Industrial application area: Technology applicable to public and private health sectors.

Received: December 4, 2025
Accepted: April 21, 2026

Resumen

La creciente prevalencia de la diabetes mellitus tipo 2 (DMT2) demanda métodos diagnósticos rápidos y no invasivos. La espectroscopia infrarroja por transformada de Fourier con reflectancia total atenuada (ATR-FTIR) en saliva ha surgido como una alternativa prometedora, capaz de capturar la "huella molecular" que refleja el estado metabólico del individuo. Este estudio tuvo como objetivo desarrollar y validar un modelo de clasificación basado en aprendizaje automático para diferenciar eficazmente entre individuos con y sin DMT2, utilizando un conjunto de datos público de 1040 espectros ATR-FTIR de saliva (540 de pacientes con DMT2 y 500 no diabéticos). Se aplicó un análisis de selección de características basado en información mutua para cuantificar la relevancia de cada número de onda e identificar las regiones espectrales más predictivas. Posteriormente, se entrenó un modelo *Ridge* Classifier utilizando únicamente el subconjunto de características filtradas por puntuación de información mutua. El análisis consiguió identificar una "huella molecular" de la DMT2 detectable en la saliva, la cual se concentra en tres regiones clave para su identificación: 450-400 cm^{-1}, 1800-1200 cm^{-1} y 3600-2800 cm^{-1}. El modelo de

clasificación final demostró un rendimiento sobresaliente, alcanzando una exactitud general del 95.19%, con una sensibilidad del 97.22% y una especificidad del 93%. Estos resultados validaron que las alteraciones metabólicas de la diabetes provocan cambios consistentes y medibles en la saliva, posicionando a la espectroscopia ATR-FTIR combinada con el aprendizaje automático como una herramienta potente y viable para la asistencia diagnóstica no invasiva. Estos hallazgos proporcionan una base empírica para que especialistas en bioquímica y biología molecular realicen estudios futuros orientados a identificar las biomoléculas específicas responsables de estas variaciones espectrales.

Palabras clave: Aprendizaje Automático, Biomarcadores Espectrales, Diabetes Mellitus Tipo 2, Espectroscopia ATR-FTIR, Huella Espectral en Saliva.

Abstract

The growing prevalence of type 2 diabetes mellitus (T2DM) demands rapid and non-invasive diagnostic methods. Attenuated total reflectance Fourier-transform infrared (ATR-FTIR) spectroscopy of saliva has emerged as a promising alternative, capable of capturing the "molecular fingerprint" that reflects the individual's metabolic state. We developed and validated a machine learning model to effectively differentiate between individuals with and without T2DM. For this purpose, we used a public dataset of 1040 saliva ATR-FTIR spectra, comprising 540 samples from individuals with T2DM and 500 from non-diabetic individuals. A feature selection analysis based on *mutual information* was applied to quantify the relevance of each wavenumber and identify the most predictive spectral regions. Subsequently, a *Ridge* Classifier model was trained using only the subset of features filtered by *mutual information* score. A molecular fingerprint of T2DM in saliva was identified in three key regions: 450-400 cm^{-1}, 1800-1200 cm^{-1}, and 3600-2800 cm^{-1}. The final classification model demonstrated outstanding performance, achieving an overall accuracy of 95.19%, with a sensitivity of 97.22% and a specificity of 93%. These results indicate that the metabolic alterations of diabetes cause consistent and measurable changes in saliva, positioning ATR-FTIR spectroscopy combined with machine learning as a powerful and viable tool for non-invasive diagnostic support. These findings provide an empirical basis for specialists in biochemistry and molecular biology to conduct future studies aimed at identifying the specific biomolecules responsible for these spectral variations.

Keywords: ATR-FTIR Spectroscopy, Machine Learning, Spectral Biomarkers, Spectral Fingerprint in Saliva, Type 2 Diabetes Mellitus.

Introduction

Type 2 diabetes mellitus (T2DM) is a major public health challenge (Khan et al., 2019). Saliva sampling is being investigated as a non-invasive approach to T2DM detection that avoids needle-based blood collection (Caixeta et al., 2020). Saliva contains proteins, sugars, and other molecules that reflect systemic physiological processes. When analyzed using infrared radiation, these molecules produce a characteristic spectral signature. This approach, known as attenuated total reflectance Fourier-transform infrared (ATR-FTIR) spectroscopy, is rapid

and requires minimal sample preparation (Caixeta et al., 2023).

Globally, the prevalence of all types of diabetes among adults aged 20 to 79 has risen steadily over recent decades. The International Diabetes Federation (2025) report, which includes data compiled by the World Bank and processed by Our World in Data, indicates that the proportion of the population affected by the disease expanded across most regions between 2000 and 2024. Figure 1 illustrates this worldwide prevalence as of 2024. This rising trend emphasizes the scale of the public health issue and highlights the importance of creating accessible and effective diagnostic tools for early detection and monitoring.

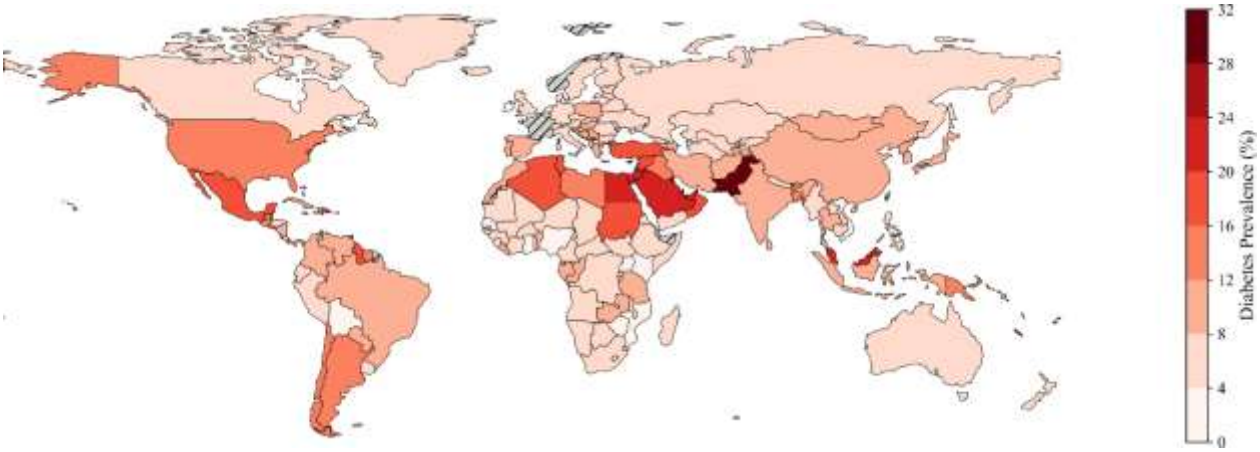


Figure 1. World prevalence of diabetes in adults aged 20 to 79 in 2024. The shades of red reflect the percentage of the population affected by all types of diabetes, where darker tones indicate higher prevalence. Chart created by the authors based on data from the International Diabetes Federation (2025) and processed by Our World in Data.

Previous research explored this technique for T2DM detection using saliva, paving the way for its clinical application (Andrade et al., 2024; Caixeta et al., 2023). Sanchez-Brito et al. (2022) made a significant contribution to the field by publishing an extensive open dataset of saliva FTIR spectra from T2DM patients, fostering reproducible research. Subsequently, Wiratama (2024) used this same dataset to establish a baseline performance using machine learning models. This study further explores the dataset through exploratory data analysis and develops a machine learning classification model to distinguish between spectra from individuals with and without T2DM. This study provides an empirical basis and reinforces the importance of this technology as a diagnostic support tool. However, an important question remains: can we identify which spectral regions are the most informative? What distinguishes this work from previous studies is the introduction of a *mutual information* gain filter that identifies the spectral regions with the highest predictive power, thereby reducing noise and improving model accuracy.

Materials and equipment

This study used the FTIR Diabetes dataset published by Sanchez-Brito et al. (2022), available in the open-access Figshare[1] repository. The data were obtained from samples collected at the Medical Specialties Unit (UEM, by its Spanish acronym), whose

[1] Available at: https://figshare.com/articles/dataset/FTIR_Dataset/19450916/1 (DOI: 10.6084/m9.figshare.19450916.v1).

spectra were captured and preprocessed at the Military Medical School (EMM); both institutions belong to the Mexican Ministry of National Defense (SEDENA).

First, the dataset's composition was analyzed, revealing a slight class imbalance (Figure 2): the 540 samples from diabetic patients accounted for 51.92% of the total, while the 500 samples from non-diabetic individuals accounted for 48.08%.

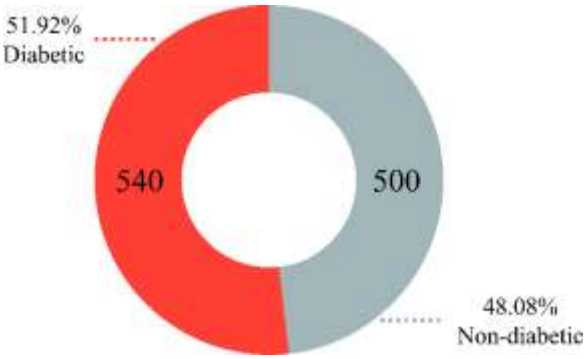


Figure 2. Distribution of study participants, where ~52% of the participants were people with diabetes (orange), and ~48% were people without diabetes (gray), representing a slight imbalance of just ~2% relative to the mean. Chart created by the authors based on the dataset from Sanchez-Brito et al. (2022).

Figure 3 shows the distribution of participants by gender, with 664 samples from women and 376 from men.

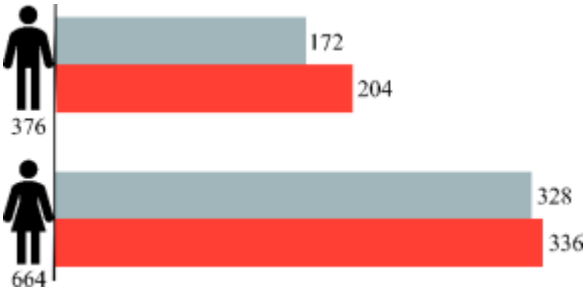


Figure 3. Distribution of diabetic and non-diabetic participants according to gender. Among men, 204 participants had diabetes and 172 did not; among women, the corresponding counts were 336 and 328. Chart created by the authors based on the dataset from Sanchez-Brito et al. (2022).

The gender distribution shown in Figure 4 indicates a higher prevalence of diabetes among men (54.26%) than among women (50.60%).

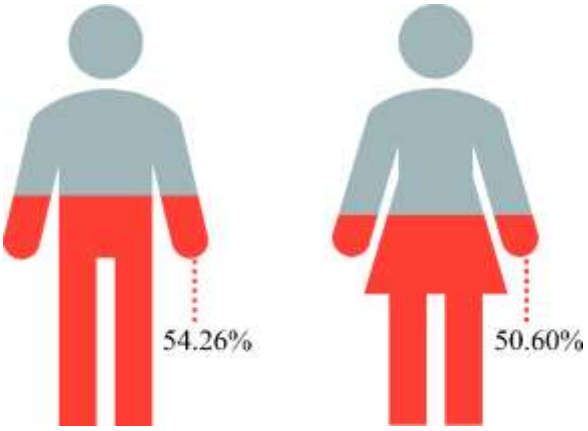


Figure 4. Percentage comparison of diabetes prevalence by gender. The red section of each figure represents the proportion of individuals with diabetes within that group (men on the left, women on the right). Chart created by the authors based on the dataset from Sanchez-Brito et al. (2022).

Subsequently, the age distribution was analyzed by gender and diabetes status, as shown in Figure 5. The analysis revealed a wide age range among the participants, from 9 to 98 years. For both men and women, the earliest recorded T2DM cases in the FTIR Diabetes dataset occurred at age 20. Additionally, non-diabetic women were represented at both extremes of the age range, including those aged 9–18 years and those older than 90 years.

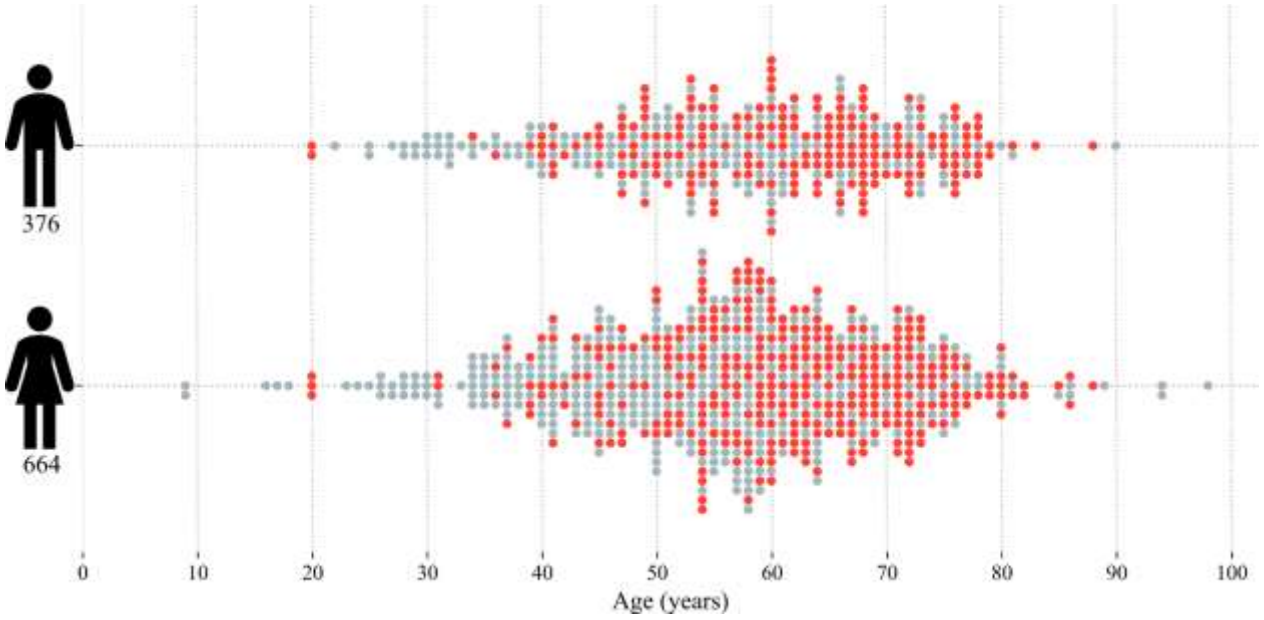


Figure 5. The swarm plot displays participants by age and diabetes status (non-diabetic: gray; diabetic: orange), spanning ages from 9 to 98 years. Figure created by the authors based on the dataset from Sanchez-Brito et al. (2022).

Each spectrum consists of 3,736 absorbance measurements across a wavenumber range from 4,000 to 399 cm^{-1}, presented in the conventional spectroscopy format from highest to lowest wavenumber. For preprocessing, statistical analysis, visualization, and modeling, we used the Python programming language (Python Software Foundation, 2025). The analysis relied on established open-source libraries (Raschka et al., 2020), such as Pandas for data manipulation (McKinney, 2022), Matplotlib (Hunter, 2007), and Seaborn (Waskom, 2021) for graphics, and Scikit-learn (Pedregosa et al., 2011) for modeling. The computational analyses can be conducted on a standard computer; a supercomputer is not required.

A descriptive statistical analysis of the spectra assessed overall variability across the dataset by calculating the minimum, maximum, mean, and standard deviation. As shown in Figure 6, the greatest variability is concentrated in two main regions, 1700 to 1500 cm^{-1}, and 3500 to 3000 cm^{-1}, where the standard deviation is greatest. These zones coincide with the highest average absorbance peaks, suggesting they are the most effective regions for sample differentiation. This observation led to the hypothesis that the most useful information for distinguishing T2DM would be concentrated in these areas. While it is logical to expect changes where the signal is most intense, this assumption required verification through a more rigorous analysis.

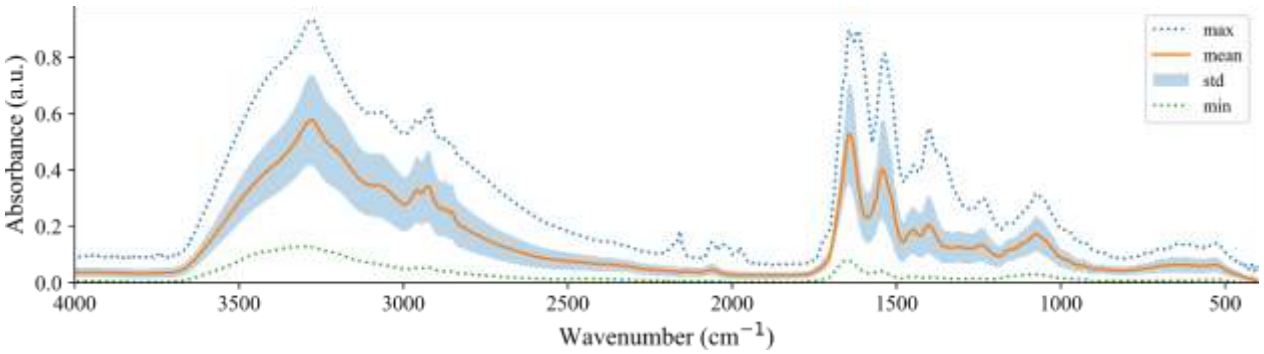


Figure 6. Average and variability of the absorbance spectra. The solid orange line represents the average absorbance of the sample set, while the shaded area corresponds to the standard deviation. The upper and lower dotted lines indicate the maximum and minimum absorbance values recorded, respectively. Figure created by the authors based on the dataset from Sanchez-Brito et al. (2022).

To validate this hypothesis and quantify the relevance of each spectral variable, a feature selection analysis based on *mutual information* was applied. This technique has been utilized in various studies, such as selecting discriminative wavelengths in hyperspectral tumor imaging (Su et al., 2026), identifying relevant features for predicting wine quality (Kulasiri et al., 2024), and filtering informative spectral bands in the analysis of latex gloves (Jongyingcharoen et al., 2024). Similarly, in our study, this method served as a relevance filter to measure how much information each wavenumber provides for differentiating between diabetic and non-diabetic individuals.

Mutual information quantifies the relationship between each wavenumber and T2DM status. A genetic algorithm was used to determine the cutoff value that separates useful information from noise. The resulting cutoff value of 0.009 retained only the most informative signals. The results presented in Figure 7 indicate that the hypothesis is partly supported but incomplete: the regions with the greatest statistical variability are also the most informative in the 1800–1200 cm^{-1} and 3600–2800 cm^{-1} ranges. Counterintuitively, despite the low statistical variability observed in Figure 6 at the lower-wavenumber end of the spectrum (near 450–400 cm^{-1}), this region exhibits the highest information gain.

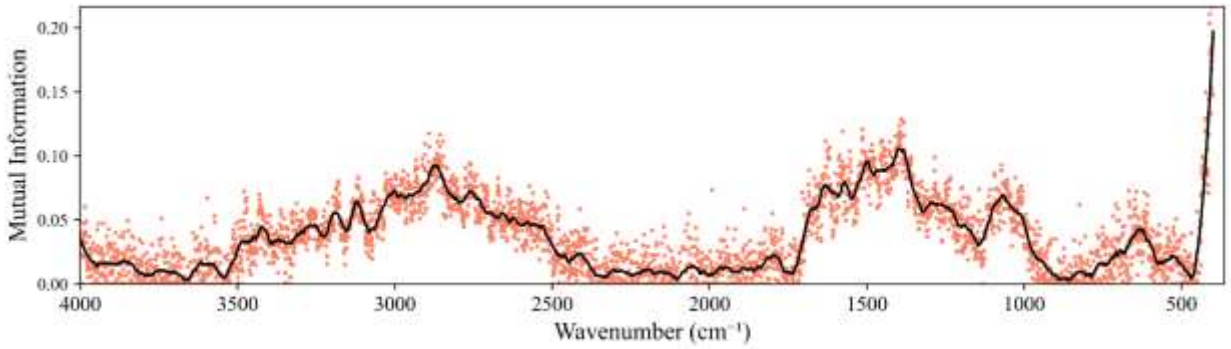


Figure 7. Orange dots represent the *mutual information* value for an individual wavenumber, while the black line denotes the moving average (81-point window, selected via genetic algorithm). Higher values indicate greater relevance for classifying T2DM. Figure created by the authors based on the dataset from Sanchez-Brito et al. (2022).

Once the most informative spectral regions were identified, we performed feature selection to optimize a classification model. Through genetic algorithm optimization, a *mutual information* threshold of 0.009 was determined, retaining only the most relevant wavenumbers. Figure 8 shows these features as vertical red lines, with color intensity indicating their importance.

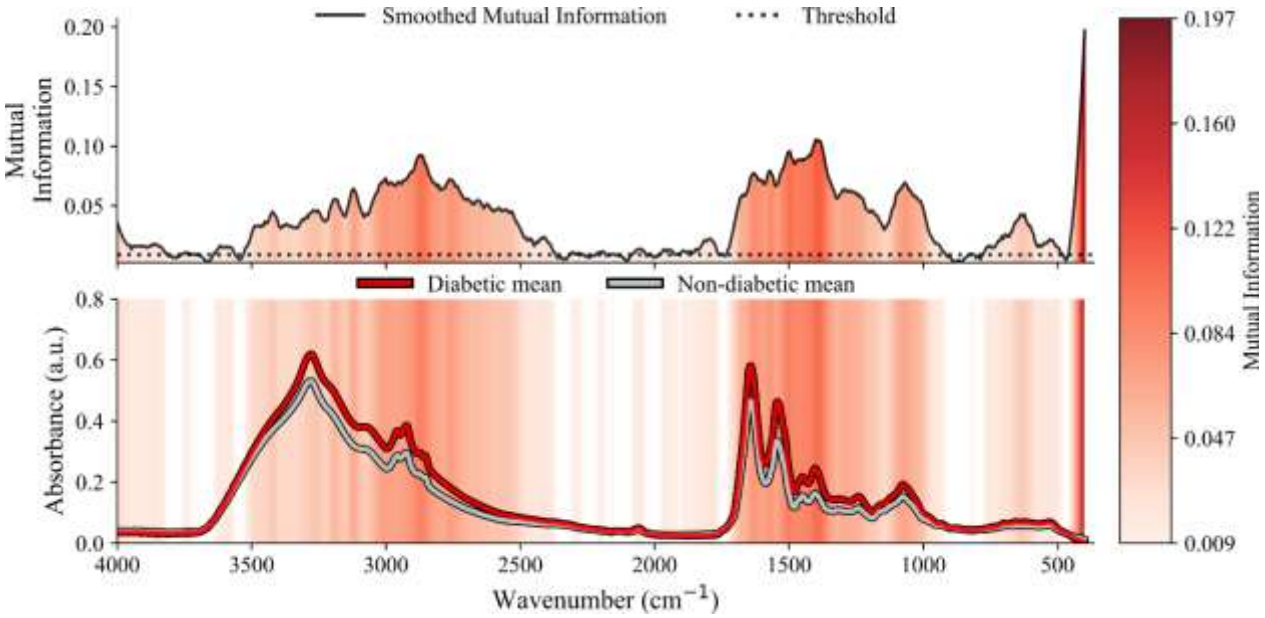


Figure 8. Mean spectra of diabetic (orange) and non-diabetic (gray) groups. Vertical red bars highlight the wavenumbers selected for the final classification model based on their high *mutual information* scores (≥ 0.009). Figure created by the authors based on the dataset from Sanchez-Brito et al. (2022).

Once the key areas of the spectral fingerprint were identified, we applied the Ridge regularization approach introduced by Hoerl and Kennard (1970). This algorithm is often evaluated within broader benchmarks and has proven effective in applications such as spectroscopic material identification for plastic recycling (Q. Albuquerque et al., 2026). Additionally, the Ridge classifier has demonstrated strong performance in neurological applications, achieving perfect discrimination (AUC = 1.00) when predicting cerebrospinal fluid erythrocyte load in Alzheimer's disease (Christodoulou et al., 2025). For other complex classification tasks, such as predicting antihypertensive drug candidates, the Ridge classifier has demonstrated accuracies of up to 99% (Marisca E. Gondokesumo et al., 2026). Ridge regularization limits model complexity rather than fitting every detail of the training data. This mechanism mitigates overfitting and enables the model to distinguish samples from individuals with and without diabetes by balancing relevant signal and background noise.

To ensure consistent and reproducible results, the data were scaled to a mean of zero and a standard deviation of one. The random seed was set to 42 for all processes involving randomness, including data splitting and model initialization. This setting ensures that calculations are performed identically in every execution, thereby promoting reproducibility. We used the *Ridge* classifier and the *mutual information* filter with their default parameters from the Scikit-learn library.

To ensure a fair evaluation, the dataset was split into stratified training (80%) and test (20%) sets. Figure 9 summarizes the workflow from saliva-sample processing to the final classification prediction, including information filtering and machine-learning classification.

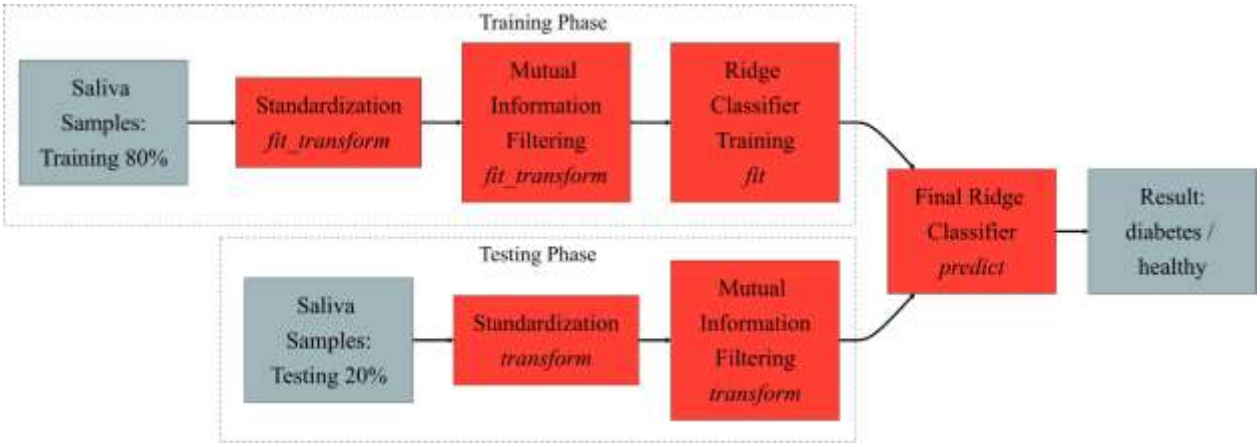


Figure 9. The integrated machine learning workflow for T2DM classification follows a two-stream architecture specifically designed to prevent data leakage and ensure an unbiased analysis. The process begins with a training phase where 80% of the data are used to fit standardization and mutual-information filtering parameters before training the Ridge classifier. Finally, the testing phase evaluates the remaining 20% of unseen data by applying the previously fitted parameters through transform to obtain the final classification prediction. Figure created by the authors.

As illustrated in Figure 9, the methodology separates the fitting of preprocessing parameters from the evaluation phase. The standardization metrics and mutual information thresholds are derived exclusively from the training subset (80%). Consequently, the test set (20%) remains completely isolated until the final prediction stage. During this phase, the previously fitted parameters are applied via transform operations to evaluate the classifier's performance.

This work focuses on pattern recognition from a computational perspective, so we identified the key information located in the spectral regions of 450–400 cm^{-1}, 1800–1200 cm^{-1}, and 3600–2800 cm^{-1}. Identifying the specific molecules, such as proteins, sugars, or lipids, associated with these spectral regions remains a task for future studies in chemistry and molecular biology.

Results and discussion

This study examined whether the most informative regions of the spectrum could be identified, and the results demonstrate that this is feasible. In the dataset, each spectrum is a dense signal composed of 3,736 wavenumbers, each representing an absorbance measurement. In data science, increasing the number of features does not necessarily improve performance, a phenomenon known as the “curse of dimensionality”. Many of these wavenumbers are redundant or irrelevant, acting as background "noise" that obscures the true signal and makes it difficult for the algorithm to identify T2DM patterns. Therefore, working with all the data is not only computationally expensive but also counterproductive for building an accurate model.

To address this challenge, we applied a method called *mutual information* as a filter to quantify which of the 3736 wavenumbers

were most informative for distinguishing between healthy and diabetic individuals. We selected a *mutual information* score greater than or equal to 0.009, which was the threshold previously determined through a genetic algorithm. This filtering reduced the feature space from 3736 to 3024 wavenumbers, ensuring that only the most informative spectral regions were used.

The key finding, illustrated in Figure 8, confirmed the presence of the fingerprint in the 1800-1200 cm^{-1} and 3600-2800 cm^{-1} regions. These bands align with statistical variability and show consistent differences between the two groups. In addition to confirming these regions, the analysis revealed an unexpected finding: a zone at the lower-wavenumber end of the spectrum, near 450–400 cm^{-1}, which at first glance did not seem relevant due to its low variability, was, counterintuitively, one of the regions with the highest predictive power.

The performance comparison provides evidence of the model's effectiveness. To validate the impact of our relevance filter, we compared the model's performance using the raw, full spectrum versus the optimized features selected via *mutual information*. As shown in Table 1, where the metrics were calculated using the independent test set to evaluate performance on unseen data, the relative improvement is substantial. These results demonstrate that applying the filter to select only the most significant spectral regions improved the system's accuracy by nearly 10%. This suggests that analyzing the entire salivary signal is unnecessary for high precision, as focusing on the most informative features yields superior results.

Table 1. Performance comparison of the *Ridge* Classifier model: Full Spectrum vs. Selected Features.

Metric	Full Spectrum (Baseline)	Selected Features Using Mutual Information (Optimized)	Relative Improvement
Accuracy	87.02%	**95.19%**	**9.39%**
Precision	83.47%	93.75%	12.32%
Sensitivity	93.42%	97.22%	4.07%
Specificity	80.00%	93.00%	16.25%

As observed, using the complete set of wavenumbers in the dataset resulted in a baseline accuracy of approximately 87%. However, by applying the *mutual information* threshold, the overall accuracy for detecting T2DM increased to 95.19%. This relative improvement of nearly 9.4% confirms that feature selection not only reduced data dimensionality but also removed noise that could confuse the classifier. This enhancement is reflected in the key performance metrics. The model achieved a high precision of 93.75%, meaning that positive classifications were correct in nearly 94 out of 100 cases. Furthermore, it demonstrated a sensitivity of 97.22%, successfully detecting 97 out of every 100 individuals with T2DM, and a specificity of 93%, correctly identifying 93 of every 100 non-diabetic individuals.

This targeted approach is consistent with recent advancements in other medical domains where the combination of ATR-FTIR spectroscopy of saliva and machine learning is being employed for automated screening. For instance, Avelar et al. (2024) successfully applied Support Vector Machine (SVM) algorithms and SHAP feature importance to identify ten specific molecular vibrational modes for screening Amelogenesis Imperfecta. Similarly, Contreras et al. (2025) utilized neural networks combined with SHAP analyses on salivary spectra to pinpoint the most influential spectral zones for detecting

Sjögren's disease. Consistent with these studies, our work supports the premise that extracting and interpreting only the most relevant spectral features is a critical step for building accurate, explainable, and reliable clinical tools.

This research goes beyond merely repeating what has been observed (Caixeta et al., 2020, 2023; Sanchez-Brito et al., 2022) regarding saliva and diabetes. Its primary contribution lies in the filtering tool used to identify the most significant signals within a large feature space, thereby providing a framework for future technological developments focused on essential features. While previous studies (Andrade et al., 2024; Wiratama, 2024) analyzed saliva as a whole, we developed a filter that identifies the most relevant spectral features. Ultimately, this work confirms that saliva contains a detectable signature of diabetes and demonstrates how this signature can be interpreted by focusing on the most informative spectral features.

Finally, these results demonstrate the effectiveness of the proposed method. We showed that a *mutual information* analysis can isolate a set of highly predictive spectral biomarkers. These findings indicate that analyzing the complete spectrum, which is noisy and computationally expensive, is unnecessary; instead, the analysis can focus on the most informative regions. These results confirm that metabolic alterations in diabetes cause consistent, measurable changes in the molecular composition of saliva. These findings indicate that FTIR spectroscopy combined with machine learning is sufficiently sensitive to detect these differences, supporting its potential as a tool for rapid, non-invasive diagnostic support. Determining which specific molecules change within these spectral regions requires analytical chemistry studies beyond the scope of this work.

Conclusions

This work confirms that the combination of ATR-FTIR spectra and machine learning can effectively detect diabetes-induced changes in saliva. With an accuracy of 95%, these results demonstrate the feasibility of developing a reliable, non-invasive diagnostic-support tool. While further studies with diverse populations are necessary, this advancement represents a significant step toward accessible, needle-free diagnostics. The 95% accuracy rate confirms that machine learning can detect diabetes-associated patterns in saliva. This finding provides a foundation for future developments in the field.

While these results confirm that diabetes leaves a measurable chemical signature in saliva, it is essential to maintain a balanced perspective. Although the current data are promising, this research is still in its early stages. The primary challenge is to validate this approach in more diverse populations and real-world clinical environments to determine its suitability as a diagnostic-support tool for routine medical use. Ultimately, these findings provide a foundation for future applications, such as automated non-invasive detection software and optimized diagnostic screening systems.

Acknowledgements

We thank Universidad Autónoma del Estado de México, through the Centro Universitario UAEM Tianguistenco. This work was supported by a doctoral scholarship from SECIHIT (CVU 932933, grant 4043887).

References

Andrade, L. M., Sabino-Silva, R., & Carneiro, M. G. (2024). Evolutionary Neural Architecture Search for Type 2 Diabetes Mellitus Diagnosis from Salivary ATR-FTIR Spectroscopy. *Anais Do XXIV Simpósio Brasileiro de*

Computação Aplicada à Saúde (SBCAS 2024), 459–470. https://doi.org/10.5753/sbcas.2024.2675

Avelar, F. M., Lanza, C. R. M., Bernardino, S. S., Garcia-Junior, M. A., Martins, M. M., Carneiro, M. G., de Azevedo, V. A. C., & Sabino-Silva, R. (2024). Salivary Molecular Spectroscopy with Machine Learning Algorithms for a Diagnostic Triage for Amelogenesis Imperfecta. *International Journal of Molecular Sciences*, *25*(17), 9464. https://doi.org/10.3390/ijms25179464

Caixeta, D. C., Aguiar, E. M. G., Cardoso-Sousa, L., Coelho, L. M. D., Oliveira, S. W., Espindola, F. S., Raniero, L., Crosara, K. T. B., Baker, M. J., Siqueira, W. L., & Sabino-Silva, R. (2020). Salivary molecular spectroscopy: A sustainable, rapid and non-invasive monitoring tool for diabetes mellitus during insulin treatment. *PLOS ONE*, *15*(3), e0223461. https://doi.org/10.1371/journal.pone.0223461

Caixeta, D. C., Carneiro, M. G., Rodrigues, R., Alves, D. C. T., Goulart, L. R., Cunha, T. M., Espindola, F. S., Vitorino, R., & Sabino-Silva, R. (2023). Salivary ATR-FTIR Spectroscopy Coupled with Support Vector Machine Classification for Screening of Type 2 Diabetes Mellitus. *Diagnostics*, *13*(8), 1396. https://doi.org/10.3390/diagnostics13081396

Christodoulou, R. C., Vamvouras, G., Sarquis, M. D., Petrou, V., Papageorgiou, P. S., Rivera, L., Morales Gonzalez, C., Rivera, G., Papageorgiou, S. G., & Vassiliou, E. (2025). From Microbleeds to Iron: AI Prediction of Cerebrospinal Fluid Erythrocyte Load in Alzheimer's Disease. *Journal of Clinical Medicine*, *14*(20), 7360. https://doi.org/10.3390/jcm14207360

Contreras, J., Vyas, B., Ryabchykov, O., Larsen, M., Gildener-Leapman, N., Bocklitz, T., Popp, J., & Lednev, I. K. (2025). ATR-FTIR Spectroscopy of Saliva and Machine Learning as a Screening Test for Sjögren Disease. *Analytical Chemistry*, *97*(47), 26034–26044. https://doi.org/10.1021/acs.analchem.5c04238

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55. https://doi.org/10.2307/1267351

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

International Diabetes Federation. (2025). *Diabetes Prevalence*. https://archive.ourworldindata.org/20250916-102301/grapher/diabetes-prevalence.html

Jongyingcharoen, J. S., Howimanporn, S., Sitorus, A., Phanomsophon, T., Posom, J., Salubsi, T., Kongwaree, A., Lim, C. H., Phetpan, K., Sirisomboon, P., & Tsuchikawa, S. (2024). Classification of the Crosslink Density Level of Para Rubber Thick Film of Medical Glove by Using Near-Infrared Spectral Data. *Polymers*, *16*(2), 184. https://doi.org/10.3390/polym16020184

Khan, M. A. B., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., & Al Kaabi, J. (2019). Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. *Journal of Epidemiology and Global Health*, *10*(1), 107.

https://doi.org/10.2991/jegh.k.191028.001

Kulasiri, D., Somin, S., & Kumara Pathirannahalage, S. (2024). A Machine Learning Pipeline for Predicting Pinot Noir Wine Quality from Viticulture Data: Development and Implementation. *Foods*, *13*(19), 3091. https://doi.org/10.3390/foods13193091

Marisca E. Gondokesumo, Muhammad R. Rasyak, & Mansur Ibrahim. (2026). Antihypertensive Activity Prediction of Physalis angulata L. Through Computational Analysis. *Tropical Journal of Natural Product Research*, *9*(12). https://doi.org/10.26538/tjnpr/v9i12.22

McKinney, W. (2022). *Python for data analysis: Data wrangling with pandas, numpy, and jupyter*. " O'Reilly Media, Inc."

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel V. and Thirion, B., Grisel, O., Blondel, M., Prettenhofer P. and Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Python Software Foundation. (2025). *Python Language Reference, version 3.13.3*. https://docs.python.org/3/reference/index.html

Q. Albuquerque, R., Stephan, F., Pongratz, A., Brütting, C., Krause, K., & Ruckdäschel, H. (2026). Recycling of Thermoplastics with Machine Learning: A Review. *Advanced Functional Materials*, *36*(3). https://doi.org/10.1002/adfm.202509447

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information*, *11*(4), 193. https://doi.org/10.3390/info11040193

Sanchez-Brito, M., Vazquez-Zapien, G. J., Luna-Rosas, F. J., Mendoza-Gonzalez, R., Martinez-Romo, J. C., & Mata-Miranda, M. M. (2022). Attenuated total reflection FTIR dataset for identification of type 2 diabetes using saliva. *Computational and Structural Biotechnology Journal*, *20*, 4542–4548. https://doi.org/10.1016/j.csbj.2022.08.038

Su, Y., Sun, J., Fu, R., Li, X., Bai, J., Li, F., Yang, H., Cheng, Y., & Lu, J. (2026). Image Feature Fusion of Hyperspectral Imaging and MRI for Automated Subtype Classification and Grading of Adult Diffuse Gliomas According to the 2021 WHO Criteria. *Diagnostics*, *16*(3), 458. https://doi.org/10.3390/diagnostics16030458

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wiratama, J. (2024). *Diagnosis Diabetes Tipe 2 dengan Machine Learning melalui Analisis Spektroskopi ATR-FTIR pada Sampel Saliva = Diagnosis of Type 2 Diabetes with Machine Learning via ATR-FTIR Spectroscopic Analysis of Saliva Samples* (p. xiii + 41). https://lib.ui.ac.id/detail?id=9999920548865&lokasi=lokal#parentHorizontalTab1