Biomedical Signal Processing and Control 69 (2021) 102855

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Type 2 diabetes diagnosis assisted by machine learning techniques through the analysis of FTIR spectra of saliva

Miguel Sanchez-Brito [a], Francisco J. Luna-Rosas [a], Ricardo Mendoza-Gonzalez [a], Gustavo J. Vazquez-Zapien [b], Julio C. Martinez-Romo [a], Monica M. Mata-Miranda [b,*]

[a] TecNM/Instituto Tecnológico de Aguascalientes, Aguascalientes 20256, Mexico
[b] Escuela Militar de Medicina, Centro Militar de Ciencias de la Salud, Secretaría de la Defensa Nacional, Ciudad de México 11200, Mexico

ARTICLE INFO

Keywords:
Artificial intelligence techniques
Artificial neural network
Fourier Transform Infrared (FTIR) spectroscopy
Human saliva
Diabetes

ABSTRACT

Diabetes is one of the four main non-communicable diseases worldwide. Despite not being a fatal disease, many complications derive from this illness that causes a drastic deterioration in the patient's health. Diabetes is a silent disease that, on many occasions, causes symptoms until the disease is already advanced, and due to the lack of education in health prevention, only a small part of the population undergoes routine laboratory studies. If this disease is detected on time, the quality of life could be improved. However, the simple facts of taking a blood sample, control studies are omitted. Besides, there is a need to sample the patient many times according to its severity and control. In the present work, we provide a novel technique based on the FTIR spectra of saliva samples to diagnose this disease. After analyzing the samples of 1,000 people, we found that it is possible to identify patients with this pathology through artificial neural networks and SVMr reliably. As it is not invasive and does not require reagents or complex processes, the proposed technique could be more agile and cheaper than traditional ones.

1. Introduction

Diabetes is one of the four main non-communicable diseases; it is estimated that this disease affects approximately 422 million people worldwide. Despite not being a fatal disease by itself, its effects on patients' health could cause it. According to the World Health Organization (WHO), adults with diabetes also have a two to three-fold increased risk of heart attacks and strokes. In pregnancy, poorly controlled diabetes increases the risk of fetal death and other complications. Some other possible complications include kidney failure, leg amputation, vision loss, and nerve damage [1]. Although there is no treatment to cure this disease, its early detection is essential to allow patients to have a good quality of life through medication and the adoption of healthy habits. The four principal methodologies for making the diagnosis of diabetes announced by the American Diabetes Association (ADA) are A1C, Fasting Plasma Glucose (FPG), Oral Glucose Tolerance Test (OGTT), and Random (also called Casual) Plasma Glucose Test. All of them are based on blood analysis [2]. Unlike the other three tests, A1C has the advantage that the patient does not necessarily have to be fasting; besides, its results reflect the patient's condition up to three months in advance because it focuses its analysis on the level of glycosylation of hemoglobin and this is the estimated life span for this protein [3]. Moreover, it is essential to mention that although it is vital to diagnose this disease early, the lack of education in health prevention makes to omit control laboratory studies by the simple fact to think in a blood sample, reason by which some other sample fluids have been considered. In this sense, although hemoglobin is a protein found in the bloodstream, it is possible to find different proteins with the capacity for glycosylation in addition to glucose in saliva [4–6].

As [7] points out, there is a need for a reagent to perform the A1C test, which price range between 7 and 9 dollars, allowing to know the result from 6 to 9 min. However, techniques like Fourier Transform Infrared Spectroscopy (FTIR) are impressive. Thanks to the interaction of a sample with different frequencies (Hertz) belonging to the mid-infrared region, it is possible to analyze its molecular structure without using a reagent. The interaction of the sample with different frequencies of the infrared spectrum causes vibrations in the bonds of the different molecules that make it up, allowing a map of the sample's chemical structure; this map is called the infrared (IR) spectrum. On the IR spectrum, several regions have already been identified depending on

* Corresponding author.
E-mail address: mmcmaribel@gmail.com (M.M. Mata-Miranda).

https://doi.org/10.1016/j.bspc.2021.102855
Received 27 March 2021; Received in revised form 28 May 2021; Accepted 5 June 2021
Available online 21 June 2021

the molecules that make up the sample; regions around 3500–3000 cm^{-1}, 1700–1600 cm^{-1}, 1560–1500 cm^{-1} are attributed to protein vibrations, specifically to amide A, amide I, and amide II respectively [7–9]. Therefore, they are fundamental for an initial analysis to identify morphological changes in the spectrum associated with the glycosylation process similar to the A1C test and hemoglobin.

Using FTIR spectroscopy to assist in diagnosing diabetes, it is possible to find some contributions [10–19]. In [10], the authors process the 1500–1200 cm^{-1} region of FTIR spectra obtained directly from the lips of 28 patients to estimate hemoglobin values using a partial least squares (PLS) regression analysis model. Based on this estimate, they characterize diabetic and non-diabetic patients depending on whether the estimated value is greater or less than 6.5%, a value used by the A1C test to make the diagnosis. The technique proposed by the authors allowed obtaining an 87.5% accuracy, correctly detecting 14 of the 16 confirmed patients with diabetes. Analyzing the region 1500–1000 cm^{-1}, and using an Extreme Gradient Boosting (XGBoost) model, the authors of [11] report values of sensitivity, specificity, and accuracy of 95.23%, 96.00%, and 95.65%, respectively, analyzing blood samples from 113 patients. Unlike [10] and [11], in [12], the authors use the full spectrum of 112 peripheral blood samples to study two methodologies to characterize patients with and without diabetes. The Classification and Regression Trees (CART) and XGBoost models elaborated allow the authors to obtain a percentage of specificity, sensibility, and accuracy of 80%, 95%, 86.67%, and 100%, 85%, and 93%, respectively.

In [13], the authors analyze 100 plasma samples from women in the gestation stage to propose a methodology that identifies women with gestational diabetes. The authors studied the region called the biological fingerprint (1800–900 cm^{-1}) using models of Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machines (SVM). According to the results reported by the authors, the LDA model allows obtaining values of 100% for accuracy, sensibility, and specificity. Despite analyzing the entire fingerprint region, the authors highlight the vibrations attributed to lipids and proteins as key to discerning between populations. Although the authors of [14] and [15] do not propose a methodology to discriminate between patients with and without diabetes, they do compare the FTIR spectra of blood samples of controlled and uncontrolled diabetic patients to identify spectral regions that could be associated with changes in the hemoglobin values of the 74 participants (36 controlled and 38 uncontrolled). In addition to sub-regions in the biological fingerprint, the authors point out the importance of including the vibrations recorded between 3300 and 2800 cm^{-1} associated with proteins and lipids. According to their study, there are also significant changes in these regions associated with the glycosylation values of the proteins.

Similarly, in [16], the authors focus only on analyzing FTIR spectra of blood samples from people with diabetes. The purpose of this work was to determine the feasibility of using FITR spectroscopy to detect bands that could be associated with diabetic people who develop retinopathy. The study involves different combinations of spectral features with linear, quadratic, and cubic Support Vector Machine, and according to the reported results, it was possible to obtain values of 90% sensitivity, 92.7% specificity, and 90.5% overall accuracy highlighting the importance of the spectral variations registered in the bands associated with lipids. In [17], a methodology is proposed for the diagnosis of diabetes from the analysis of FTIR spectra of fingernail clippings. The authors analyze the region between wavenumbers 1140–970 cm^{-1} in samples from 127 patients; in this region, vibrations attributed mainly to carbohydrates are reported. Although the characterization model focuses on the mentioned region, the authors highlight the differences between the populations in the 3140–2095 cm^{-1} region associated with lipids. Through a Receiver Operating Characteristic (ROC) analysis, the authors report values of 82% and 90% for specificity and sensitivity, respectively.

Although studies from [10] to [17] use FTIR to characterize populations with and without diabetes, they use blood samples. Closely linked to the current project proposal are the works [18] and [19], which analyze saliva sample spectra to propose a non-invasive technique similar to the present study. Authors of [18] propose an LDA model that receives six subregions of the biological fingerprint of saliva samples; however, it also highlights that the most significant differences in the lipid region occur in the range between 3000 and 2800 cm^{-1}. The six subregions were obtained from subtracting the average spectra of the populations with and without diabetes, made up of 39 and 22 individuals, respectively. Of the 66 spectra that made up the work database, approximately two-thirds were used in the model training process, and the remaining third was used in the evaluation process; the metrics of sensitivity, specificity, and accuracy for the trained model was 100% while in the evaluation process they were 88.2%. In [19], saliva samples from patients with psoriasis (35), diabetes (10), and patients without any of the above conditions (20) are compared. The authors indicate three main subregions that could differentiate populations, the first one centered at 3316 ± 41 cm^{-1}, the second at 647 ± 3 cm^{-1}, and the third at 1543 ± 3 cm^{-1}. However, they just study the region between 1700 and 1500 cm^{-1} associated with amides I and II of the samples that make up their database through Principal Component Analysis (PCA) and adopt a set of seven principal components that describe 99.99% of the variance of their data. Although it does not formally propose a methodology to use PCA in diagnosing diabetes, the graphing of the 1vs6 components shows (although with several overlapping points between the populations) the differences between the groups.

The diversity of the methodologies used to characterize the spectra sets derives from the sample complexity (components) that is analyzed as indicated in [20] and [21]. To reduce the number of methodologies that allow segmenting populations based on IR spectra, authors of [22] and [23] list the ones that have allowed the best results to be obtained in the characterization spectra for patients with different pathologies. In this work, we experimented with the techniques proposed in [22] and [23] to determine if it is possible to characterize people with and without diabetes in a reliable way using the only saliva and identifying the regions in which the most representative vibrations of both are present.

2. Materials and methods

2.1. Population and spectra capture

This study discriminated between diabetic and healthy patients by analyzing FTIR spectra of saliva assisted by machine learning techniques. For which purpose, between February 2019 and February 2020, 1000 samples of approximately 1 ml of saliva were collected in microcentrifuge tubes in the Unidad de Especialidades Médicas (UEM) of the Secretaría de la Defensa Nacional (SEDENA). The control group was integrated by 500 healthy ambulatory volunteers who reported glucose levels less than 100 mg/dl and no other biochemical alteration in their laboratory studies. The diabetic group was integrated by 500 diabetic patients previously diagnosed with medical follow-up in the UEM. The population characteristics are summarized in Table 1.

The inclusion criteria were patients over 18 years of age who had a fasting period of at least 8 h. Similarly, the exclusion criteria were patients who had brushed or rinsed the oral cavity with mouthwash before sampling and patients with orthodontic treatment or other dental

Table 1
Population characteristics.

Group	Gender		Age (years)	Time of evolution of the disease (years)
Control	Male	172	53 ± 14	–
	Female	328		
Diabetic	Male	188	60 ± 11	11 ± 8
	Female	312		

treatment. Both groups were informed that their samples would be used for different diagnostic assays as a reference control or to try other diagnosis types.

Written informed consent for the obtention of the saliva sample and its analysis were obtained. The Clinical Research Ethics Committee of the Unidad de Especialidades Médicas of the SEDENA approved the protocol and the informed consent forms. All experiments were examined and approved by the appropriate ethics committee and followed the ethical standards laid down in the 1964 Declaration of Helsinki.

The spectrometer used in the present study was Jasco FTIR-6600, equipped with attenuated total reflectance (ATR) accessory. From each sample, 3 μl of saliva was placed directly on the ATR by pipetting. Once the sample was dry, the spectrum was captured using the following parameters, the resolution of 4 cm^{-1} and a total of 120 scans as suggested [20] for liquid samples. The captured spectra were normalized using standard normal variate (SNV) methodology to correct scattering as suggested [23].

2.2. Methodologies for characterization

The authors of [18] and [19] point out a set of methodologies based on the accuracy values they are allowed to achieve in their implemented jobs. The proposed techniques are LDA, K-Nearest Neighbors (KNN), Multivariable Linear Regression Models (MLRM), SVM, and Artificial Neural Networks (ANN). The ANN and SVM techniques, unlike the others, allow us to try to solve a problem from the classification (ANNc and SVMc) or regression (ANNr and SVMr) paradigms; in this study, we analyzed both forms for these methodologies. The k value for which we configured the KNN classifier was 10. For SVM, we employ a grade two polynomial kernel; the initial cost of constraint violation was 1, and the epsilon in the insensitive-loss function 0.1. Considering [21], we configured the ANN as the most straightforward model; this is, a feed-forward network with a backpropagation rule, and as an activation function, we use the hyperbolic tangent. The MLRM is similar to the simple linear regression but with as many coefficients as absorbance values the spectrum region evaluated has.

The means of each study group (control and diabetic) were compared to detect the most relevant regions in which changes in biochemical compounds could be detected and determine if one of them is essential for its characterization or necessary to analyze the entire spectrum. After that, the seven techniques mentioned above (LDA, KNN, MLRM, SVMc, SVMr, ANNc, and ANNr) were developed as is suggested by [23]. Additionally, we also evaluated the region called biological fingerprint (B.F. from 1800 to 900 cm^{-1}). The performance classification and regression techniques were carried out through the Leave One Out Cross Validation (LOOCV) and Hold Out cross-validation techniques.

3. Results

In this research, we used LDA, KNN, MLRM, SVMc, SVMr, ANNc, and ANNr techniques through FTIR spectra to discriminate people with diabetes from healthy people employing saliva. By representing our

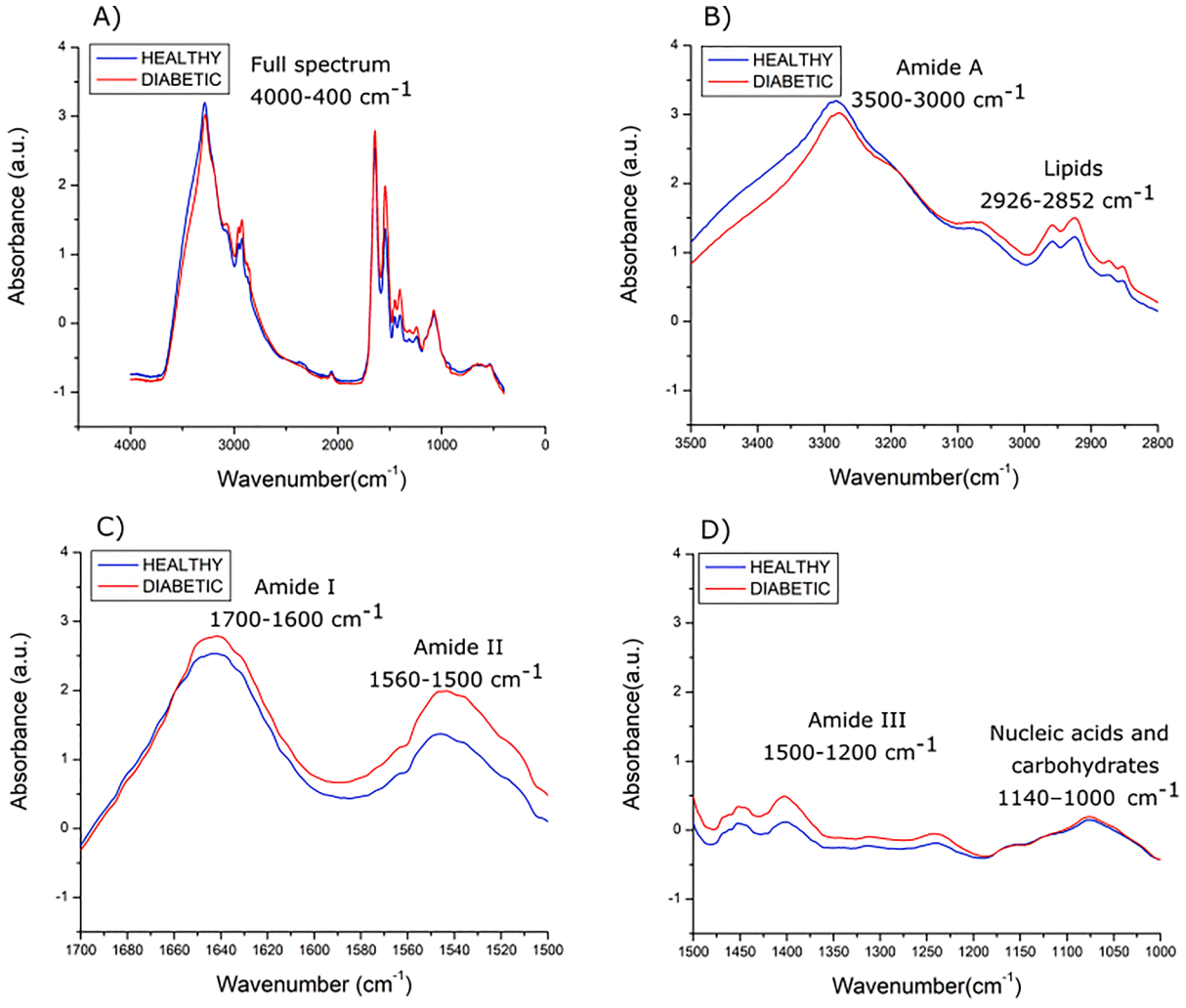

Fig. 1. Spectral differences between the means of healthy and diabetic populations.

1000 spectra through a mean, we contrast the authors' regions [10–13,18,19]. The results are presented in Fig. 1, where it is possible to find several differences, mainly in absorbance levels in the lipid (Fig. 1B), Amides I and II (Fig. 1C) regions, as well as in the region of 1500 to 1000 cm^{-1} (Fig. 1D) suggested by[11], where vibrations are associated to amide III, carbohydrates, and nucleic acids [24]. However, the region associated with amide A (Fig. 1B, 3500–3000 cm^{-1}) [7] presented a change in the healthy population's slope, which could be relevant when implementing the characterization methodologies.

In the following subsections, we present the results obtained by applying the seven methods (LDA, KNN, SVMc, ANNc, MLRM, SVMr, and ANNr) considering first a LOOCV evaluation and subsequently a Hold out 80–20 (training-test) strategy. We use circles to represent the outputs of the methods; in blue, we point out the HEALTHY population, and in red, the DIABETIC one. To determine the relevance of one specific region for population characterization, we perform six tests analyzing different spectrum subsets with each method. The sub-regions analyzed correspond to the full spectrum (4000–400 cm^{-1}), Amide A together with lipids (3500–2800 cm^{-1}), B.F. (1800–900 cm^{-1}), amide I (1700–1600 cm^{-1}), amide II (1560–1500 cm^{-1}), and amide III and nucleic acids (1500–1200 cm^{-1}; the objective of this strategy is to detect the sub-region that more excellent separability allows obtaining between the blue and red points.

3.1. Linear discriminant analysis (LDA)-LOOCV

LDA uses the pooled variance–covariance matrix in the distance calculation; hence, the distance between a test sample and a given class centroid is weighted according to each spectral variable's overall variance. The results that we obtained when applying this methodology are presented in Fig. 2.

As we can appreciate in Fig. 2, analyzing all the regions, a mix of red and blue points is obtained for both categories, diabetic (D) and healthy (H). The optimal performance would have allowed obtaining a complete horizontal blue line for the horizontal axis "H" and a complete horizontal red line for the axis "D," however, analyzing the B.F. region, it was possible to obtain an accuracy of 93.5%, a sensibility of 93.6%, and 93.7% specificity.

3.2. K-nearest neighbors (KNN)-LOOCV

KNN assigns a category to a spectrum based on the spectra' K closest measurements previously recorded in the database. The obtained results from this methodology are shown in Fig. 3.

By the results presented in Fig. 3, we can appreciate a lack performance of KNN against LDA; this is based on the fact that the blue points (derived from processing samples of people without diabetes with K-NN) are largely occluded by red points misclassified on the horizontal line "H." The best performance was achieved considering the entire spectrum, allowing us to obtain 77%, 75.2%, and 79.2% for accuracy, sensibility, and specificity, respectively.

3.3. Support Vector Machine classification (SVMc)-LOOCV

A Support Vector Machine (SVM) can be imagined as a surface that creates a boundary between data plotted multidimensional points representing examples and their feature values. An SVM aims to create a flat boundary called a hyperplane, which divides the space to create relatively homogeneous partitions on either side [25].

The main difference between the classification and regression models for both SVM and ANN is that the classification models' dependent variable is categorical. In contrast, the variable is numerical and used to solve the equations formed together with the regression model's independent variables. Using SVMc, we obtained the results presented in Fig. 4.

With this methodology, it is possible to obtain good results analyzing the amide A and lipids regions. However, the best percentages of accuracy, sensibility, and specificity using this technique were achieved by analyzing the entire spectrum, 94.1%, 94.2%, and 94%, respectively; these percentages are higher than those obtained by the KNN and LDA methods.

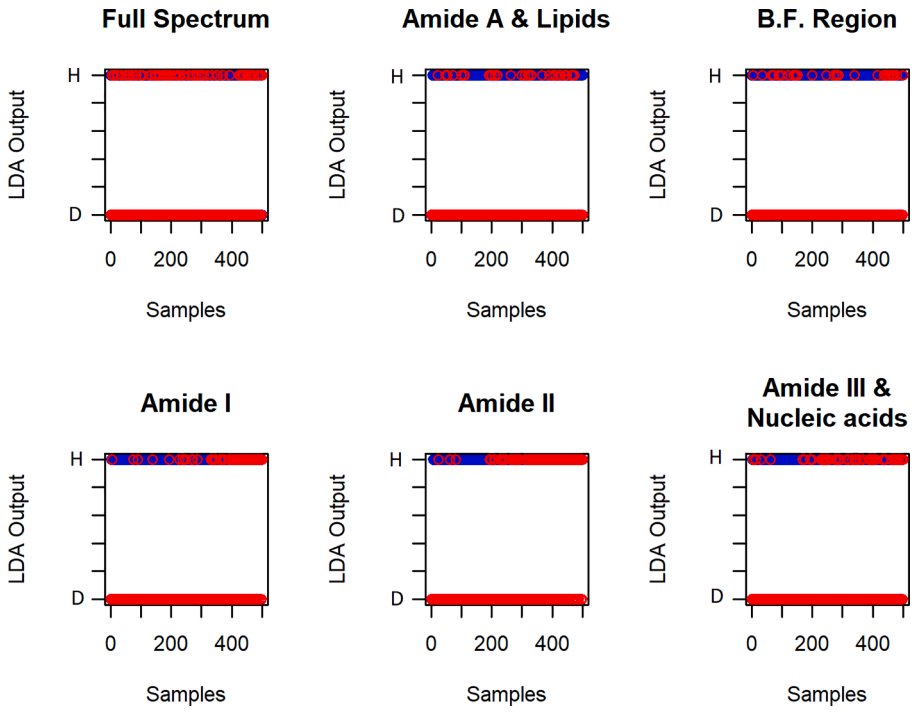

Fig. 2. Results obtained by LDA for patients' characterization process with (red points-D) and without diabetes (blue points-H) analyzing different IR spectrum regions.

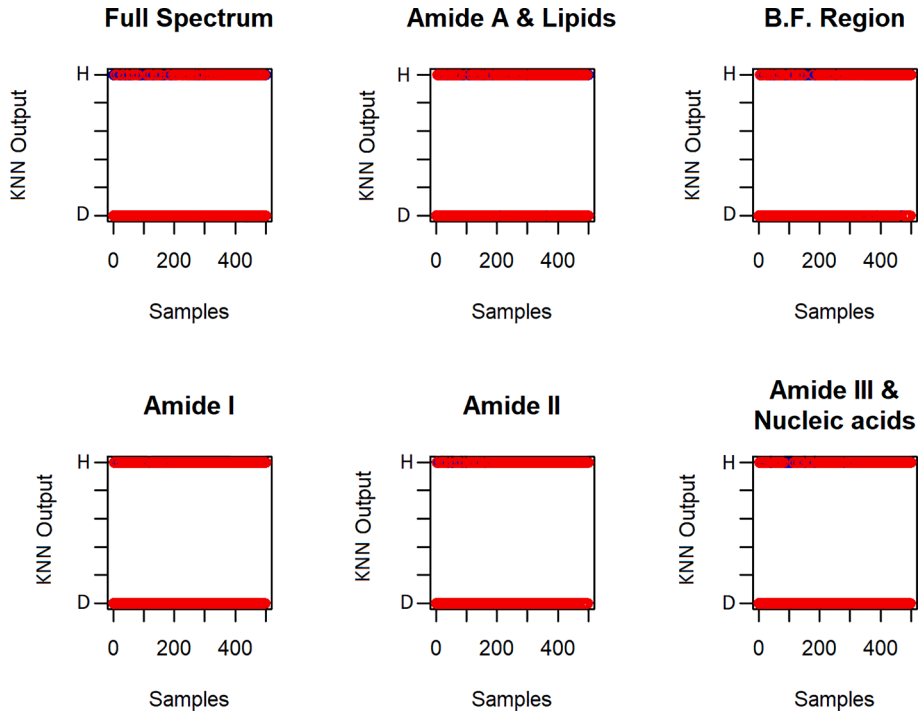

Fig. 3. Results obtained by KNN for patients' characterization process with and without diabetes analyzing different IR spectrum regions.

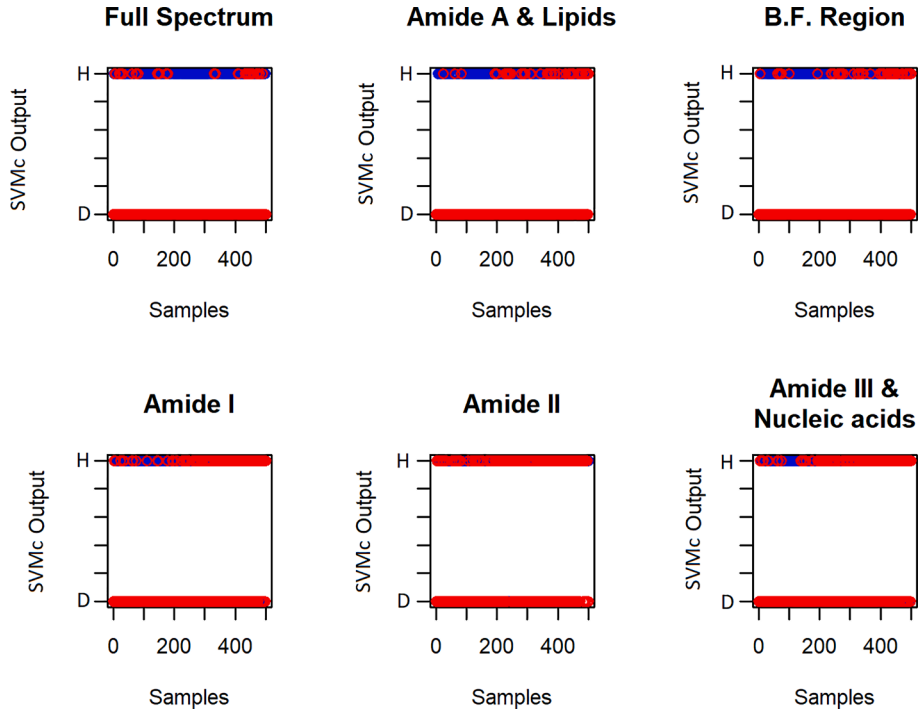

Fig. 4. Results obtained by SVMc for patients' characterization process with and without diabetes analyzing different IR spectrum regions.

3.4. Decision trees

Decision trees utilize a tree structure to model the relationships among the features and the potential outcomes. There are numerous implementations of decision trees, but one of the most well-known implementations is the C5.0 algorithm [25]. After applying the C5.0 methodology to our database, we obtained the results presented in Fig. 5; in this figure, it can be seen that the best distribution of the output values is obtained by analyzing the entire spectrum.

Considering the C5.0 methodology, the best accuracy, sensibility, and specificity results were obtained by analyzing the full spectrum, allowing to obtain the following percentages for each metric 79.8%,81.4%, and 78.2%, respectively.

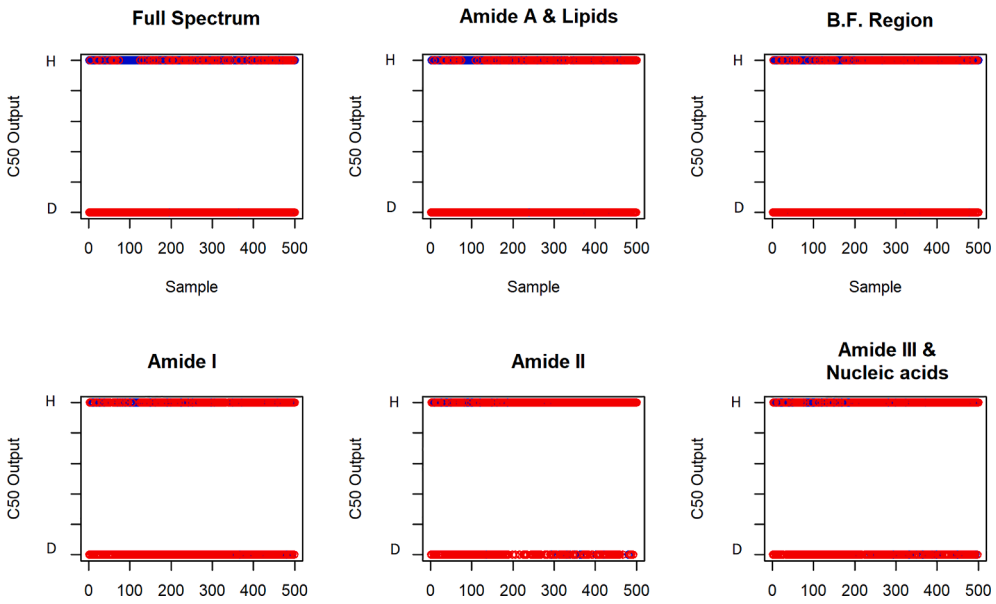

Fig. 5. Results obtained by C5.0 for patients' characterization process with and without diabetes analyzing different IR spectrum regions.

3.5. *Artificial neural network classification (ANNc)-LOOCV*

The neural network is constructed with an interconnected group of nodes, which involves the input, connected weights, processing element, and output. Like SVM, it can be used in regression or classification problems depending on the dependent variable.

To get an initial idea of the performance that ANNs could have in characterizing FTIR spectra of patients with and without diabetes, we used the simpler model suggested by [25], which consists of a single neuron. As an activation function, we use the hyperbolic tangent function. The results obtained using this methodology are presented in Fig. 5; this employment technique may allow us to segment both populations by analyzing the lipid region in conjunction with Amide A (3500–2800 cm^{-1}). Fig. 6 presents the results obtained directly from the activation function; the final classification would be carried out by implementing the last step function with a limit of 0.5, which would correctly classify our population of 1000 spectra.

Using ANNc, it is possible to segment populations with accuracy, sensibility, and specificity of 100%, analyzing the region from 3500 to 2800 cm^{-1}. Addressing the problem from the classification paradigm and considering the LOOCV model, the best results were obtained through ANNc, as shown in Figs. 2-6, with the Amide A region and lipids being the most relevant to characterize both populations. We approach the regression paradigm using the MLRM, SVMr, and ANNr models in the following subsections. Unlike the classification models, the regression models' output is a non-categorical numerical variable, so it is

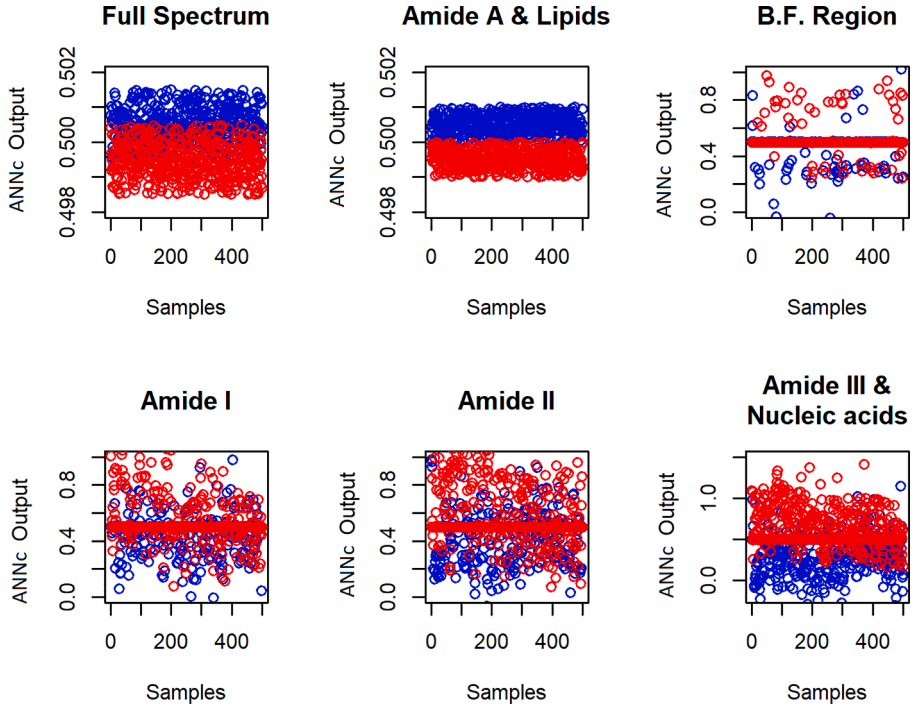

Fig. 6. Results obtained by ANNc for patients' characterization process with and without diabetes analyzing different IR spectrum regions.

necessary to define a limit from which the results are associated with a particular class.

3.6. Multivariable linear regression models (MLRM)-LOOCV

This model is similar to the simple linear regression model, but multiple independent variables contribute to the dependent variable's computing. Here the dependent variable was established as 1 for the diabetes population and 2 for the control one. The obtained results are presented in Fig. 7, where it is possible to observe that the best results are obtained when analyzing the amide I region.

Once the decision frontier was defined at 1.5, it was possible to obtain 83.3% accuracy, 86.6% of sensibility, and 90% specificity analyzing the amide I region.

3.7. Support Vector Machine regression (SVMr)-LOOCV

The results of analyzing the database by SVMr are presented in Fig. 8; analyzing the regions of the complete spectrum, Amide A and Lipids region (3500–2800 cm^{-1}), B.F., and amide I spectra are classified in the right way; however, the best results are obtained by analyzing the B.F.

By analyzing the B.F. with SVMr and setting a boundary at 1.5, it is possible to obtain 90%, 95%, and 85.6% accuracy, sensibility, and specificity.

3.8. Artificial neural network regression (ANNr)-LOOCV

As we can see in Fig. 9, through ANNr, it is possible to reliably segment the populations analyzing either the full spectrum or the Amide A and lipids region (3500–2800 cm^{-1}). Although using ANNc, it was also possible to characterize both populations correctly by analyzing the 3500–2800 cm^{-1} region; the results obtained by ANNr show less dispersion, which suggests that it is the best option and that the main morphological differences between the spectra of patients with and without diabetes are found in this region.

Using ANNr, it is possible to segment populations with an accuracy, sensibility, and specificity of 100%, defining a decision boundary in 1.5.

As noted in Fig. 9, it is possible to characterize people with and without diabetes in a reliable way by analyzing the region 3500–2800 cm^{-1} related to amide A and lipids of the FTIR spectra of saliva samples. The distribution of the outputs obtained by ANNr are presented in Fig. 10; on it, we can see that despite the occurrence of several cases that exceed a standard deviation from the mean of each population, the outputs obtained by ANNr are not far from this limit, besides, the separation between classes allows obtained a reliable margin to carry out the diagnosis.

The metrics obtained from the model are presented in Table 2, including standard deviation for the outputs of each population and the coefficient of determination R^2.

In addition to the accuracy percentages, it is essential to mention that the proposed methodology does not require reagents or devices to extract the fluid to be analyzed. The materials necessary to perform the proposed test are only a plastic tip of the pipette and a microcentrifuge tube of approximately 3 ml and also that the processing time, once the model is delivered, is approximately 30 s per sample.

Morais and Ghimire have already pointed out the advantages and disadvantages of using the LOOCV methodology to evaluate the performance of a classifier [23,26]; however, they have also stated that the main weakness of this technique is that constructing as many models as samples in the database, the model could have generalization problems with new data. Reason by which, in this research, we fragmented the database into two randomly constructed subsets. The first subset considered 80% (800) of the spectra, and this subset was used to train the seven classifiers analyzed in this work. The remaining 20% (200) of the database was used to evaluate the trained models as suggested by [23,26] for Hold out cross-validation. The results are presented in Fig. 11 and Table 3, where it is possible to appreciate that the most promising methodologies and spectral regions to perform the characterization of patients with and without diabetes change for the results obtained by LOOCV.

Considering the results presented in Fig. 11 and Table 3, we observed that through the Hold Out methodology, the best results are obtained

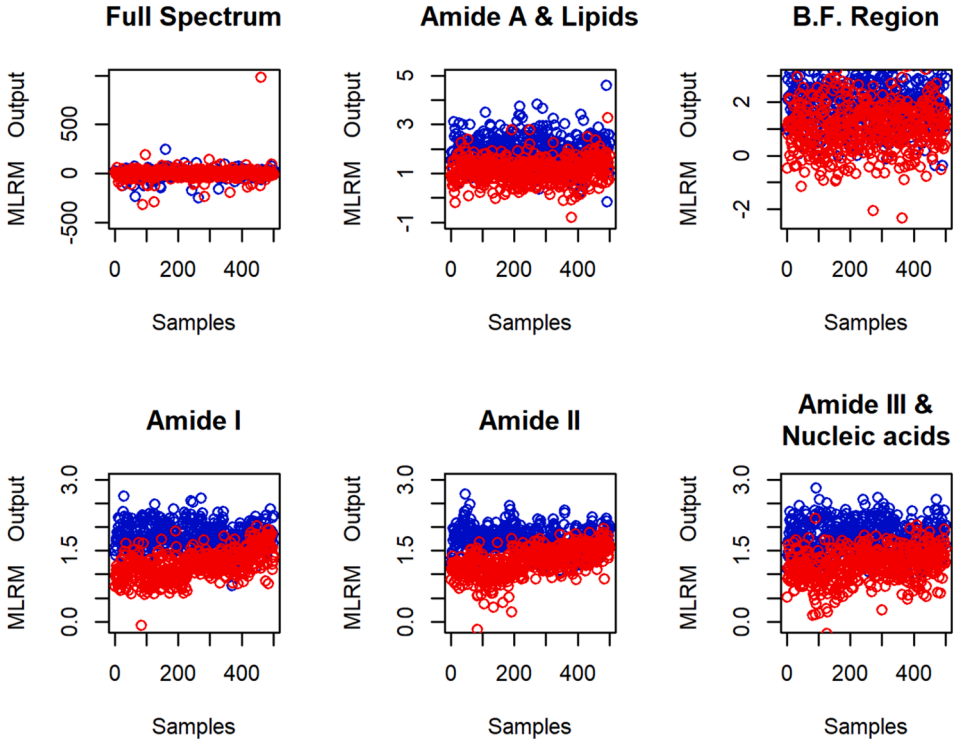

Fig. 7. Results obtained by MLRM for patients' characterization process with and without diabetes analyzing different IR spectrum regions.

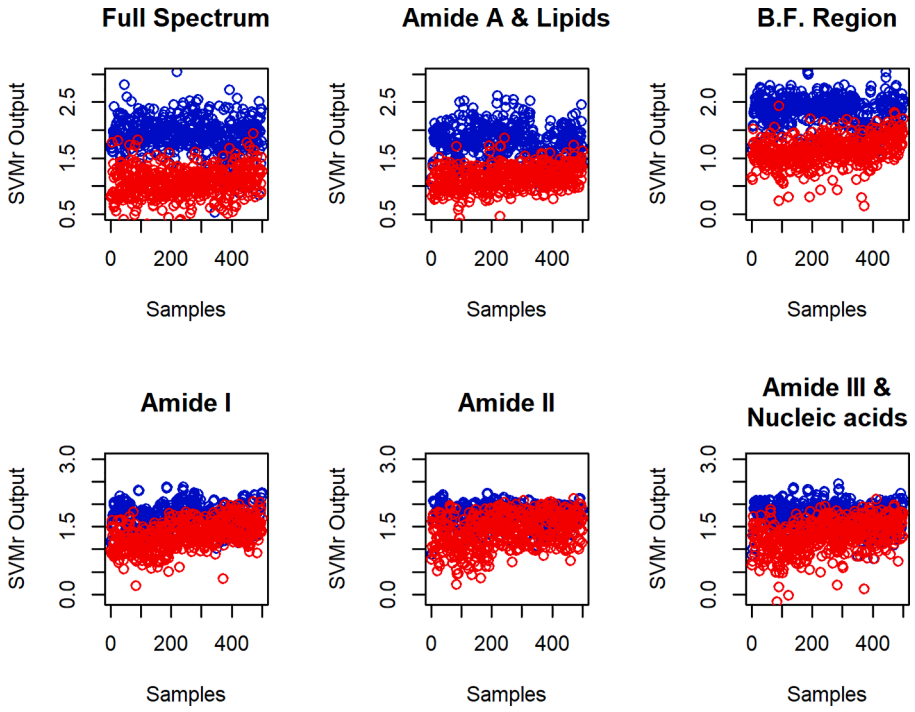

Fig. 8. Results obtained by SVMr for patients' characterization process with and without diabetes analyzing different IR spectrum regions.

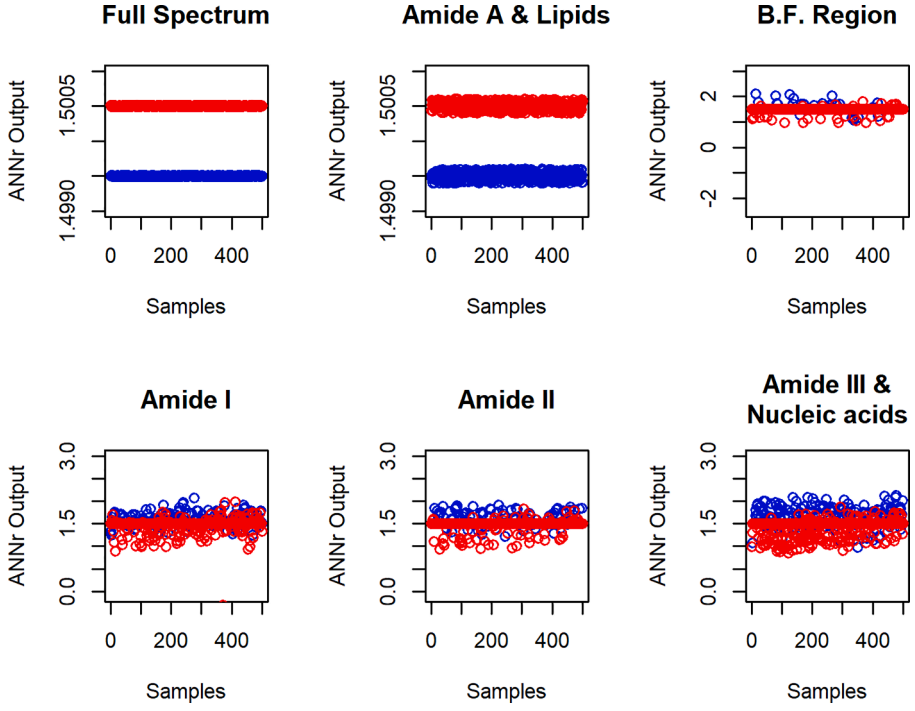

Fig. 9. Results obtained by ANNr for patients' characterization process with and without diabetes analyzing different IR spectrum regions.

through the analysis of the full spectrum using SVMr (Fig. 10 F), allowing to obtain percentages of accuracy, sensibility, and specificity of 95%, 99%, and 96% respectively as indicated in Table 3. This classifier also provided good results analyzing the full spectrum but considering the LOOCV methodology and the classification paradigm as indicated in section (*SVMc)-LOOCV* where we report percentages of accuracy, sensibility, and specificity of 94.1%, 94.2%, and 94%, respectively. However, using LOOCV, it was more convenient to analyze the region of amide A and lipids (3500–2800 cm^{-1}) with ANNr, which allows us to achieve 100% accuracy in population characterization. It is also worth highlighting LDA's performance, which showed percentages of sensibility, specificity, and accuracy of more than 90% with both LOOCV and Hold Out. Fig. 12, presents the means of both populations in this region (similarly to Fig. 1B) plus an upper and lower limit obtained from the standard deviation of the 500 spectra that make up each one of them, on it, it is possible to appreciate that the change in the slopes in the regions of amide A and lipids, in general, is maintained for the 1000 spectra analyzed.

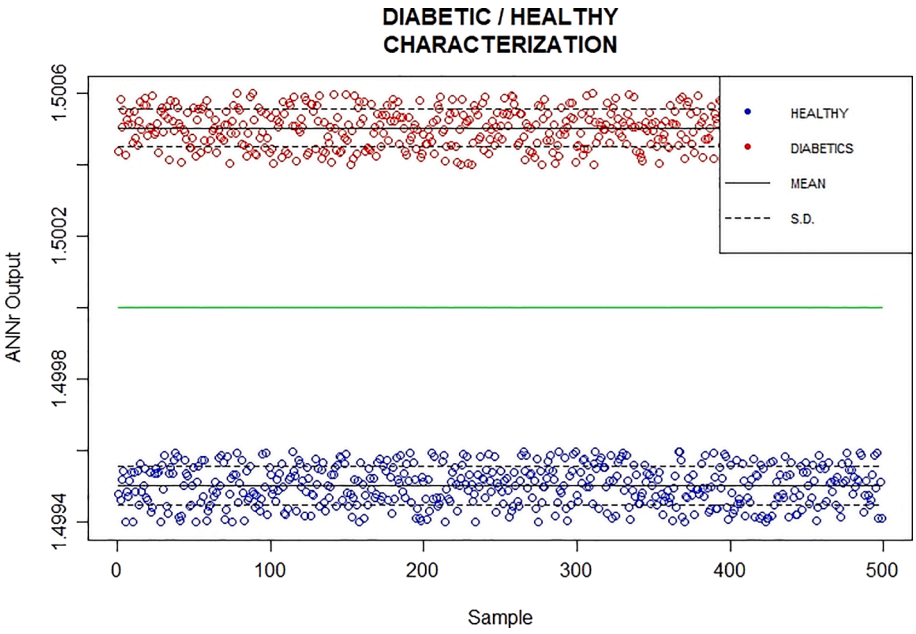

Fig. 10. Behavior of the output values obtained by ANNr.

Table 2
Output metrics obtained by ANNr.

Method	Population	Output range		S.D.	Misidentified	Accuracy (%)	Sensibility(%)	Specificity(%)	R^2
ANNr	HEALTHY	Max.	1.4996	5.25e-05	0/1000	100	100	100	0.9978
		Min.	1.4994						
	DIABETIC	Max.	1.5006	5.54e-05					
		Min.	1.5004						

4. Discussion

When evaluating a non-invasive methodology for the characterization of patients with and without diabetes from the analysis of IR spectra of saliva samples, the obtained results are presented in the present work. The percentages obtained for accuracy, sensibility, and specificity (100%) with LOOCV studying the regions of lipids and amide A in the region (3500–2800 cm^{-1}) [7]–[9], and the results obtained by evaluating the full spectrum through SVMr following the Hold Out cross-validation methodology (accuracy, sensibility, and specificity of 95%, 99%, 96% respectively), suggest that all amides regions on FTIR spectra play a significant role for the characterization of patients with and without diabetes since evaluating the B.F. (1800–900 cm^{-1}) which involves vibrations attributed to amide I, II and III; with SVMr and Hold Out cross-validation a lower sensitivity and accuracy percentages were obtained compared to complete spectrum analysis, which includes amide A in addition to the amides mentioned above.

The task of assisting in the diabetes diagnosis through the IR spectra analysis has been the subject of several studies [10–13,18,19], even in [13]. Using an LDA model called GA-LDA Test and emphasizing the importance of the lipid and protein regions, the authors manage to segment populations with and without diabetes with an accuracy, sensibility, and specificity similar to those obtained in the present work with LOOCV methodology and percentages of 5%, 1% and, 4% higher for accuracy, sensibility, and specificity after evaluating the spectra using SVMr and Hold Out. However, the authors of [13] analyzed blood samples (from 100 participants) in contrast to the saliva samples used here (1000), and in this fluid is the hemoglobin that, according to the ADA[3], is the crucial protein to assist in the diabetes diagnosis. This could be an essential factor for the designation of the regions of interest in their work (Amides I and II) in contrast to the need to use the entire spectrum, as pointed out by our experiments for our saliva-based methodology. Using an LDA (similar to [13]- [18]) model with the methodology proposed in the present work, important values of accuracy were also obtained (~93.5% using LOOCV and Hold Out section 3.1 and Table 3); however, the best results, as observed in Figs. 2-9, are obtained through ANNr and the vibrations attributed to amide A and lipids (3500–2800 cm^{-1}) considering LOOCV and SVMr studying the full spectrum following the Hold Out method. Scott et al. [18] also propose a method based on saliva analysis through FTIR spectroscopy to assist the diabetes diagnosis. They analyze some B.F. sub-regions using an LDA model, getting an 88.9% accuracy for a test subset composed of 17 spectra (nine from diabetic patients and eight from the control group). Despite getting good results using LDA and analyzing the B.F. region, our best results were obtained by analyzing the amide A and lipids regions (3500–2800 cm-1); in a test subset of 200 spectra, we got an accuracy of 93%, as we report on Table 3, this suggests the importance of considering this region in the study.

Although most related works have identified lipids, it is fundamental to carry out people's characterization with and without diabetes; the region attributed to amide A has been omitted in all the consulted works focusing only on amides I and II. The main vibrations recorded in the 3500–3000 cm^{-1} region attributed to amide A are due to the N–H bonds as mentioned [9] and, according to what [27,28] indicates, these bonds are fundamental in the glycation process. Through Fig. 1B, we can see an increase in the slope of the control group in the region of amide A which ends up leading to a higher level of absorbance in this region for the mentioned group, this change based on the obtained results is essential, to distinguish both populations since, despite there are differences in other regions, these are only in absorbance values.

From Fig. 12, we can conclude that in general, the behavior described is maintained, that is, people with diabetes have a higher

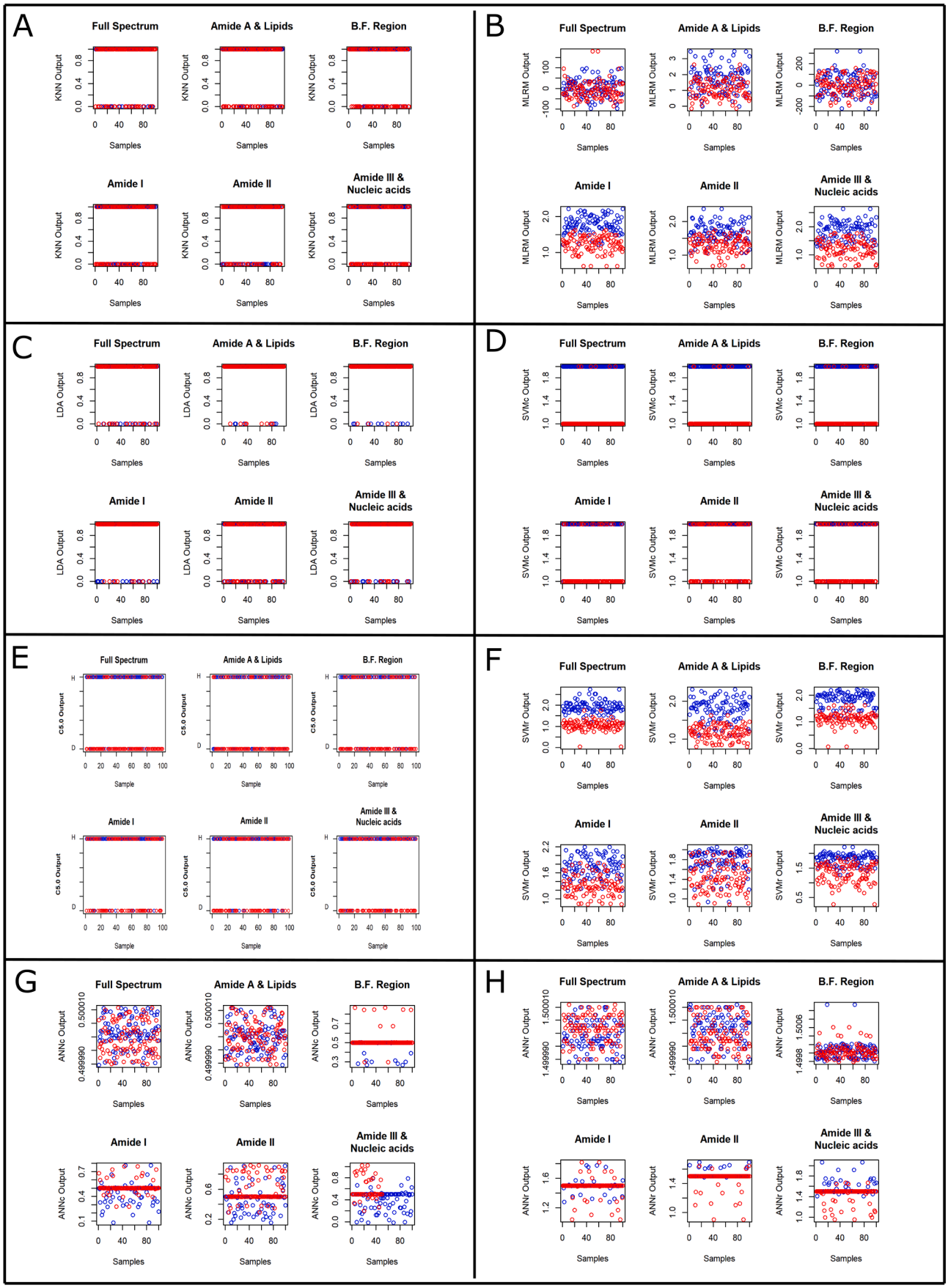

Fig. 11. Evaluation of classifiers using the Hold Out cross validation methodology.

Table 3
Metrics obtained from the evaluation of the seven classifiers using Hold Out.

KNN						MLRM					
FULL		Amide A and Lipids		B.F. Region		FULL		Amide A and Lipids		B.F. Region	
sensibility	76	sensibility	64	sensibility	56	sensibility	52	sensibility	74	sensibility	52
specificity	80	specificity	80	specificity	86	specificity	62	specificity	64	specificity	48
accuracy	78	accuracy	72	accuracy	71	accuracy	57	accuracy	69	accuracy	50
Amide I		**Amide II**		**Amide III and Nucleic acids**		**Amide I**		**Amide II**		**Amide III and Nucleic acids**	
sensibility	56	sensibility	70	sensibility	50	sensibility	94	sensibility	80	sensibility	82
specificity	66	specificity	64	specificity	86	specificity	90	specificity	76	specificity	84
accuracy	61	accuracy	67	accuracy	68	accuracy	92	accuracy	78	accuracy	83
LDA						**SVMc**					
FULL		**Amide A and Lipids**		**B.F. Region**		**FULL**		**Amide A and Lipids**		**B.F. Region**	
sensibility	82	sensibility	92	sensibility	96	sensibility	94	sensibility	92	sensibility	90
specificity	82	specificity	94	specificity	88	specificity	90	specificity	86	specificity	88
accuracy	82	accuracy	93	accuracy	92	accuracy	92	accuracy	89	accuracy	89
Amide I		**Amide II**		**Amide III and Nucleic acids**		**Amide I**		**Amide II**		**Amide III and Nucleic acids**	
sensibility	90	sensibility	70	sensibility	86	sensibility	80	sensibility	54	sensibility	70
specificity	82	specificity	72	specificity	82	specificity	78	specificity	92	specificity	84
accuracy	86	accuracy	71	accuracy	84	accuracy	79	accuracy	73	accuracy	77
C5.0						**SVMr**					
FULL		**Amide A and Lipids**		**B.F. Region**		**FULL**		**Amide A and Lipids**		**B.F. Region**	
sensibility	74	sensibility	68	sensibility	64	sensibility	99	sensibility	72	sensibility	84
specificity	84	specificity	62	specificity	70	specificity	96	specificity	90	specificity	94
accuracy	79	accuracy	65	accuracy	67	accuracy	95	accuracy	81	accuracy	89
Amide I		**Amide II**		**Amide III and Nucleic acids**		**Amide I**		**Amide II**		**Amide III and Nucleic acids**	
sensibility	48	sensibility	48	sensibility	58	sensibility	84	sensibility	76	sensibility	88
specificity	90	specificity	82	specificity	84	specificity	78	specificity	75	specificity	80
accuracy	69	accuracy	65	accuracy	71	accuracy	62	accuracy	75	accuracy	84
ANNc						**ANNr**					
FULL		**Amide A and Lipids**		**B.F. Region**		**FULL**		**Amide A and Lipids**		**B.F. Region**	
sensibility	68	sensibility	56	sensibility	44	sensibility	70	sensibility	64	sensibility	44
specificity	60	specificity	62	specificity	60	specificity	66	specificity	50	specificity	52
accuracy	64	accuracy	59	accuracy	52	accuracy	68	accuracy	57	accuracy	48
Amide I		**Amide II**		**Amide III and Nucleic acids**		**Amide I**		**Amide II**		**Amide III and Nucleic acids**	
sensibility	35	sensibility	55	sensibility	46	sensibility	62	sensibility	84	sensibility	78
specificity	90	specificity	45	specificity	54	specificity	44	specificity	24	specificity	23
accuracy	62	accuracy	50	accuracy	50	accuracy	53	accuracy	54	accuracy	50

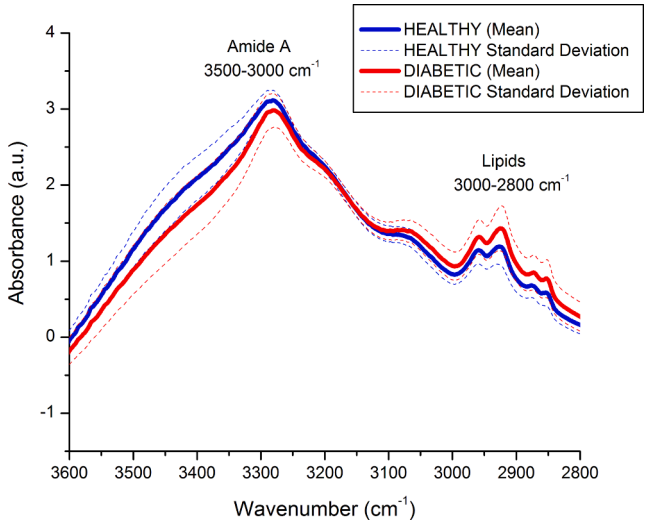

Fig. 12. Average behavior of the populations in the region of amide A and lipids.

absorbance in the lipid region than people without this disease, while in the protein region, the absorbances of the spectra of the people in the control group are higher. In works such as that of [29], it has been reported that people with diabetes have uncontrolled lipid and lipoprotein levels, characterized by high triglyceride levels. Additionally, [30–33] have shown that, in general, people with diabetes present a more significant process of protein metabolism during a period of fasting than people who do not have this disease. The results showed in Fig. 11 present a behavior similar to that indicated by the works [29] and [30–33], and according to the results obtained by ANNr, it is essential to identify both populations reliably.

The individual analysis of particular subregions of the spectrum of a biological sample has been suggested in different works to determine the main molecular groups affected by a specific pathology [23,34–36], highlighting biological fingerprint analysis (from 1800 to 900 cm^{-1}), but also suggesting to consider the analysis of the region comprised by wavenumbers 3700–2800 cm^{-1} where vibrations associated with lipids and proteins have been reported. Although is impossible to isolate a molecular group individually without subjecting the saliva sample to a pre-processing with different reagents, it is possible to bombard the sample with electromagnetic frequencies mainly associated with a specific molecular group. The results obtained with the present work suggest that the main differences between people with and without diabetes are found between wavenumbers 3600–2800 cm^{-1}. This range of wavenumbers involves frequencies between 107.9 and 83.9 THz approximately. Together with the machine learning method selected in the present work, these frequencies could be helpful to build a device that allows preliminary assistance in diagnosing diabetes in a more agile economic way (since no reagents were used in this work).

Based on the disparity in the methods that allow the best performance to be obtained through LOOCV and hold out, it is inferred that the best way to assist in the diagnosis of diabetes through the analysis of the FTIR spectra of saliva samples is through the implementation of ANNr since this methodology allows to obtain the greatest separability between the populations (Fig. 10); however, the performance of this methodology considering an 80–20 segmentation could be interpreted as the need for more samples to be able to think about the use of the technique proposed here in a real environment.

5. Conclusions

The early diagnosis of diabetes is indispensable to maintain the quality of life; herein, we present a novel technique to identify people with diabetes by analyzing FTIR spectra of saliva samples. The obtained spectra from 1000 participants (500 previously diagnosed with type 2 diabetes) indicate that the N–H vibrations of amide A and those associated with lipids are essential to differentiate the populations.

Although the traditional tests for the diagnosis of diabetes are relatively short (between 6 and 9 min), the need for screening diabetes without the discomfort that a blood sample represents is essential to seek alternative techniques that assist in diagnosing this disease in a better way. The obtained results by ANNr suggest that this is a reliable and stable technique in contrast to other classification methodologies. Moreover, it is a low-cost and agile technique used as a prelude to gold-standard tests.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgements

The authors wish to thank the people who kindly agreed to participate in the project and provided a saliva sample. To the Medical Specialties Unit (UEM) of the National Defense Secretariat (SEDENA) for the facilities granted for selecting candidates for sampling and the facilities granted to carry out the collection, specifically in the laboratory area. To the Military School of Medicine (EMM) for the facilities granted for collecting and processing samples. To the Technological Institute of Aguascalientes (ITA) and the National Council of Science and Technology (CONACyT) of Mexico for providing academic advice and allowing access to economic scholarships that allowed the research to be carried out. Finally, to the Medical Interns Second Lieutenants: Leonardo Alfredo Pedraza Zuñiga and Cristian Martínez Pérez for their availability to support collecting and analyzing clinical information from patients participating in the project.

Author contribution

Miguel Sanchez-Brito wrote, analysed data, design and perform the experiments and edited the manuscript. Francisco J. Luna-Rosas designed and performed the experiments for data analysis, wrote, analysed data and edited the manuscript. Ricardo Mendoza-Gonzalez wrote, designed the experiments, analysed data and edited the manuscript. Gustavo J. Vazquez-Zapien wrote, designed the experiments, analysed data and edited the manuscript. Julio C. Martinez-Romo wrote, designed the experiments, analysed data and edited the manuscript. Monica M. Mata-Miranda conceived and designed the study, designed the experiments, wrote, and edited the manuscript. All authors read the manuscript and agreed to its contents.

References

[1] "Diabetes." https://www.who.int/health-topics/diabetes#tab=tab_1 (accessed Jan. 29, 2021).
[2] "Diagnosis | ADA." https://www.diabetes.org/a1c/diagnosis (accessed Jan. 29, 2021).
[3] "A1C and eAG | ADA." https://www.diabetes.org/diabetes/a1c-test-meaning/a1c-and-eag (accessed Jan. 29, 2021).
[4] D.J. Chiego, Essentials of Oral Histology and Embryology - Mena Adapted Reprint E-Book, Elsevier Health Sciences (2017).
[5] J. Faintuch, S. Faintuch, Precision Medicine for Investigators, Practitioners and Providers, Academic Press, 2019.
[6] R.L. Witt, Surgery of the Salivary Glands E-Book, Elsevier Health Sciences (2019).
[7] L. M. Miller, M. W. Bourassa, and R. J. Smith, "FTIR spectroscopic imaging of protein aggregation in living cells," Biochimica et Biophysica Acta (BBA) - Biomembranes, vol. 1828, no. 10, pp. 2339–2346, Oct. 2013, doi: 10.1016/j.bbamem.2013.01.014.
[8] L. V. Bel'skaya, E. A. Sarf, N. A. Makarova, "Use of Fourier Transform IR Spectroscopy for the Study of Saliva Composition," J. Appl. Spectrosc., vol. 85, no. 3, pp. 445–451, Jul. 2018, doi: 10.1007/s10812-018-0670-0.
[9] C. Paluszkiewicz, et al., Saliva as a first-line diagnostic tool: A spectral challenge for identification of cancer biomarkers, J. Mol. Liquids 307 (Jun. 2020), 112961, https://doi.org/10.1016/j.molliq.2020.112961.
[10] S. Yoshida, M. Yoshida, M. Yamamoto, J. Takeda, Optical screening of diabetes mellitus using non-invasive Fourier-transform infrared spectroscopy technique for human lip, J. Pharmaceut. Biomed. Analysis 76 (Mar. 2013) 169–176, https://doi.org/10.1016/j.jpba.2012.12.009.
[11] P. Guang et al., "Blood-based FTIR-ATR spectroscopy coupled with extreme gradient boosting for the diagnosis of type 2 diabetes: A STARD compliant diagnosis research.," Medicine (Baltimore), pp. e19657–e19657, 2020, Accessed: Jan. 29, 2021. [Online]. Available: https://dx.doi.org/10.1097/MD.0000000000019657.
[12] X. Yang, et al., Pre-diabetes diagnosis based on ATR-FTIR spectroscopy combined with CART and XGBoots, Optik 180 (Feb. 2019) 189–198, https://doi.org/10.1016/j.ijleo.2018.11.059.
[13] E. Bernardes-Oliveira et al., "Spectrochemical differentiation in gestational diabetes mellitus based on attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy and multivariate analysis," Scientific Reports, vol. 10, no. 1, Art. no. 1, Nov. 2020, doi: 10.1038/s41598-020-75539-y.
[14] S. Ye, P. Ruan, J. Yong, H. Shen, Z. Liao, and X. Dong, "The impact of the HbA1c level of type 2 diabetics on the structure of haemoglobin," Scient. Rep., vol. 6, no. 1, Art. no. 1, Sep. 2016, doi: 10.1038/srep33352.
[15] P.-T. Dong, H. Lin, K.-C. Huang, and J.-X. Cheng, "Label-free quantitation of glycated hemoglobin in single red blood cells by transient absorption microscopy and phasor analysis," Sci. Adv., vol. 5, no. 5, p. eaav0561, May 2019, doi: 10.1126/sciadv.aav0561.
[16] A.G. Mazumder, S. Banerjee, F. Zevictovich, S. Ghosh, A. Mukherjee, J. Chatterjee, Fourier-transform-infrared-spectroscopy based metabolomic spectral biomarker selection towards optimal diagnostic differentiation of diabetes with and without retinopathy, Spectrosc. Lett. 51 (7) (Aug. 2018) 340–349, https://doi.org/10.1080/00387010.2018.1471510.
[17] R. Coopman, et al., Glycation in human fingernail clippings using ATR-FTIR spectrometry, a new marker for the diagnosis and monitoring of diabetes mellitus, Clin. Biochem. 50 (1) (Jan. 2017) 62–67, https://doi.org/10.1016/j.clinbiochem.2016.09.001.
[18] D.A. Scott, et al., Diabetes-related molecular signatures in infrared spectra of human saliva, Diabetol. Metabol. Syndr. 2 (1) (Jul. 2010) 48, https://doi.org/10.1186/1758-5996-2-48.
[19] U. Bottoni, et al., Infrared Saliva Analysis of Psoriatic and Diabetic Patients: Similarities in Protein Components, IEEE Trans. Biomed. Eng. 63 (2) (Feb. 2016) 379–384, https://doi.org/10.1109/TBME.2015.2458967.
[20] B.C. Smith, Fundamentals of Fourier Transform Infrared Spectroscopy, CRC Press, 2011.
[21] B.C. Smith, Infrared Spectral Interpretation: A Systematic Approach, CRC Press, 2018.
[22] A. Sala, et al., Biofluid diagnostics by FTIR spectroscopy: A platform technology for cancer detection, Cancer Lett. 477 (May 2020) 122–130, https://doi.org/10.1016/j.canlet.2020.02.020.
[23] C. L. M. Morais, K. M. G. Lima, M. Singh, and F. L. Martin, "Tutorial: multivariate classification for vibrational spectroscopy in biological samples," Nature Protocols, vol. 15, no. 7, Art. no. 7, Jul. 2020, doi: 10.1038/s41596-020-0322-8.
[24] H. Ghimire, M. Venkataramani, Z. Bian, Y. Liu, and A. G. U. Perera, "ATR-FTIR spectral discrimination between normal and tumorous mouse models of lymphoma and melanoma from serum samples," Scientific Reports, vol. 7, no. 1, Art. no. 1, Dec. 2017, doi: 10.1038/s41598-017-17027-4.
[25] B. Lantz, Machine Learning with R. Packt Publishing Ltd, 2013.
[26] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media Inc, 2019.
[27] N. R. PhD, Tietz Fundamentals of Clinical Chemistry and Molecular Diagnostics 8 E; South Asia Edition; e-Book. Elsevier Health Sciences, 2019.
[28] T.S. Raju, Co- and Post-Translational Modifications of Therapeutic Antibodies and Proteins, John Wiley & Sons, 2019.
[29] R.M. Krauss, Lipids and Lipoproteins in Patients With Type 2 Diabetes, Diabetes Care 27 (6) (Jun. 2004) 1496–1504, https://doi.org/10.2337/diacare.27.6.1496.
[30] T.V. Fiorentino, et al., Pioglitazone corrects dysregulation of skeletal muscle mitochondrial proteins involved in ATP synthesis in type 2 diabetes, Metabol. – Clin. Exp. 114 (Jan. 2021), https://doi.org/10.1016/j.metabol.2020.154416.
[31] M. Charlton, K.S. Nair, Protein Metabolism in Insulin-Dependent Diabetes Mellitus, J. Nutri. 128 (2) (Feb. 1998) 323S–327S, https://doi.org/10.1093/jn/128.2.323S.
[32] N. Møller, K.S. Nair, Diabetes and Protein Metabolism, Diabetes 57 (1) (Jan. 2008) 3–4, https://doi.org/10.2337/db07-1581.
[33] P. Felig, J. Wahren, R. Sherwin, G. Palaiologos, Amino Acid and Protein Metabolism in Diabetes Mellitus, Arch. Intern. Med. 137 (4) (Apr. 1977) 507–513, https://doi.org/10.1001/archinte.1977.03630160069014.
[34] K. Naseer, S. Ali, S. Mubarik, I. Hussain, B. Mirza, J. Qazi, FTIR spectroscopy of freeze-dried human sera as a novel approach for dengue diagnosis, Infrar. Phys.

Technol. 102 (Nov. 2019), 102998, https://doi.org/10.1016/j.infrared.2019.102998.

[35] Z. Movasaghi, S. Rehman, D. I. ur Rehman, "Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues," Appl. Spectrosc. Rev., vol. 43, no. 2, pp. 134–179, Feb. 2008, doi: 10.1080/05704920701829043.

[36] R.K. Sahu, S. Mordechai, Spectroscopic techniques in medicine: The future of diagnostics, Appl. Spectrosc. Rev. 51 (6) (Jul. 2016) 484–499, https://doi.org/10.1080/05704928.2016.1157809.