Performance of Regression Models in the Estimation of Glucose Levels through the Analysis of FTIR Spectra of Saliva Samples

Miguel Sánchez Brito[1], Ricardo Mendoza González[1],
Gustavo J. Vázquez Zapién[2], Francisco J. Luna Rosas[1],
Mónica M. Mata Miranda[2], Julio C. Martínez Romo[1]

[1] Tecnológico Nacional de México,
Instituto Tecnológico de Aguascalientes,
Mexico

[2] Escuela Militar de Medicina,
Centro Militar de Ciencias de la Salud,
Secretaría de la Defensa Nacional,
Mexico

{miguel_sanchezbrito, gus1202}@hotmail.com,
{ricardo.mendoza.gonz, mmcmaribel}@gmail.com,
{fcoluna2000, jucemaro}@yahoo.com

Abstract. According to the World Health Organization (WHO) [1], the diabetes is one of the four main non-communicable diseases that has the greatest impact on the death rate worldwide, with around 1.6 million deaths attributed to it. The American Diabetes Association (ADA) has stated that type 2 diabetes is the most common type of this condition [2]. Diabetes is a degenerative disease that has no cure, however, it is possible to adopt a set of actions that allow minimizing its effects on daily life, such as physical activities, proper nutrition, adequate medication and constant monitoring of glucose levels in order to prevent hyper/hypo glycemia. Currently, there are various methodologies that allow glucose monitoring through blood analysis, however, in this research, we present a non-invasive methodology to estimate glucose levels from the analysis of the molecular changes visible in saliva that produce the different glucose concentrations [3] by means of Fourier Transform Infrared (FTIR) spectroscopy and Artificial Neural Networks (ANN). After correctly characterizing the samples of 540 people, we infer that the proposed methodology would have a good performance to carry out this analysis.

Keywords: Saliva, Fourier transform infrared spectroscopy, artificial neural networks, type 2 diabetes, glucose monitoring.

1 Introduction

The Fourier Transform Infrared (FTIR) spectroscopy involves the study of the interaction of radiation with molecular vibrations [4]. The aforementioned interaction

Table 1. Population information.

Gender		Age (average)
Male	312	61±11
Female	228	60±12

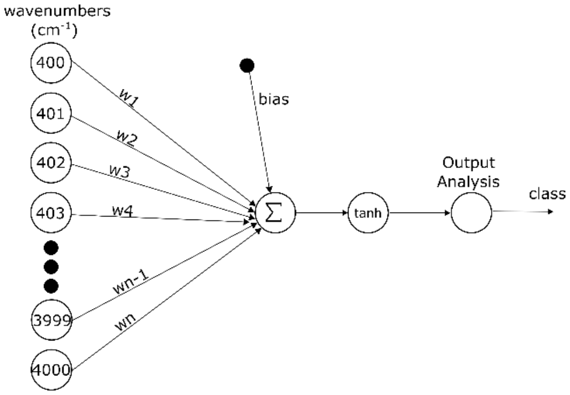

Fig. 1. ANN configuration.

refers to the vibration produced by the molecular bonds that make up a sample when impacted by an electromagnetic wave with a specific frequency, the vibration of the link is stored in a vector known as the FTIR spectrum; as might be expected, for the case of FTIR spectroscopy, these frequencies belong to the region of the infrared spectrum, from 0.11992 to 419.70944 THz approximately.

Based on the frequencies of the electromagnetic wave, the infrared spectrum is divided into three regions: near, medium and far; the middle region (11.9 to 119.9 THz approximately) being the most suitable for analyzing organic samples due to the links that make it up [5].

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys, and nerves. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin [1]; while some people can control their blood sugar levels with healthy eating and exercise, others may need medication or insulin to help manage it. One of the main techniques for monitoring glucose levels according to the ADA is the A1C test, which focuses on the analysis of glycosylated hemoglobin.

Hemoglobin, is a protein that links up with glucose, is found inside red blood cells, its job is to carry oxygen from the lungs to all the cells of the body. Glucose enters your red blood cells and links up (or glycates) with molecules of hemoglobin. The more glucose in your blood, the more hemoglobin gets glycated. By measuring the percentage of A1C in the blood, you get an overview of your average blood glucose control for the past few months [6].

The largest component of saliva is water (99%); however, it is also possible to find proteins, inorganic ions, and enzyme cofactors metabolites, RNA, and DNA [7].

Although it is not possible to find hemoglobin in saliva, it is possible to find other proteins that can bind with glucose molecules [8].

Through the analysis with regression models of Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Multivariable Linear Regression Models (MLRM) of FTIR spectra of saliva samples from people diagnosed with type 2 diabetes, a non-invasive methodology is proposed to estimate the glucose values of 540 patients, the results indicate that ANN models are the most suitable for estimating since they presented the lowest root-mean-square-error for the validation subset.

2 Materials and Methods

With the consent of the patients and with the approval of the research protocol 001/2019 by the ethics committee of the Unidad de Especialidades Médicas (UEM) of the Secretaría de la Defensa Nacional (SEDENA), approximately 1ml of saliva from patients previously diagnosed with type 2 diabetes was collected in the laboratory area after taking a blood sample for their routine examination's glucose control. Our database was made up as indicated in Table 1.

The glucose values recorded by the UEM obtained range between 47 and 503 mg / dL. By pipetting, 3 µl of saliva was deposited in a Jasco FTIR-6600 spectrometer. After drying the sample using an incandescent lamp, the spectrum was captured using a resolution of 4 cm^{-1} and 120 scans as is suggested by [9] for liquid samples. Once all the spectra were obtained, they were normalized by Standard Normal Variate (SNV) (1):

$$z = \frac{(\mathrm{x} - \mu)}{\sigma}. \tag{1}$$

The ANN's initial configuration is a single hidden layer with a neuron using the hyperbolic tangent as the activation function.

SVM (2) was initially configured using a grade 2 polynomial kernel *d* in (3), the initial cost of constraints violation was 1, and the epsilon in the insensitive-loss function 0.1:

$$\min 0.5 \left\| \vec{\mathrm{w}} \right\|^2 + C\,\Sigma_{i=1}^{n} \xi_i \tag{2}$$

$$\text{s.t. } \mathrm{y_i}(\vec{\mathrm{w}}{*}\vec{\mathrm{x}}\text{-b}) \geq 1\text{-}\xi_i,\ \forall \vec{\mathrm{x}}_i,\ \xi_i \geq 0\ ,$$

$$\mathrm{K}(\vec{x_i}, \vec{x_j}) = (\vec{x_i} * \vec{x_j} + 1)^{\mathrm{d}}\ . \tag{3}$$

3 Results

After drying the spectra, it is possible to appreciate the three main macromolecular groups mentioned by [10]: Lipids 3000-2800 cm^{-1}, Proteins in the range of 1700-1600 and 1560-1500 cm^{-1} (Amide I and Amide II respectively), and nucleic acids in the region of 1250-1000 cm^{-1}. In Fig. 2, the morphological changes of the FTIR spectrum of the sample derived from the drying process can be appreciated, in the same way, the

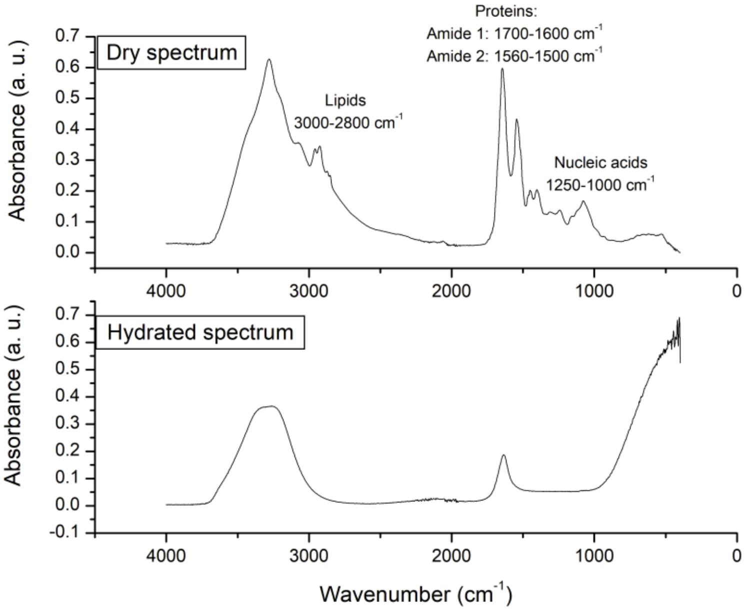

Fig. 2. Morphological changes in the spectrum due to the drying process. In the dry spectrum, the three main macromolecular groups are indicated: lipids, proteins, and nucleic acids.

main macromolecular groups are indicated. Once the spectra of the samples that make up our database were captured, it was normalized according to SNV.

In Fig. 3, the frequencies of the glucose values recorded by the patients contemplated in the present study are presented. We use the Leave One Out Cross Validation (LOOCV) methodology to estimate the glucose value from a certain spectrum, this means that we use n-1 samples to train the ANN, SVM, and MLRM models and we use this model to estimate the glucose value of the spectrum that it was omitted in the training process [11].

It is possible to think about the idea that by capturing an FTIR spectrum of glucose it could indicate the optimal region to analyze without the need for complex methodologies such as ANN or SVM. In Fig 4 a), we present the full FTIR spectra of injectable water solutions with different percentages of glucose. In their research [12] indicates the region 1180-950 cm^{-1} as the optimal region to detect vibrations associated with glucose in saliva, this region is presented in Fig. 4 b).

Considering what is indicated by [12], we selected the wavenumber 1075 cm^{-1} as an indicator of the glucose level. From Fig. 4 b), we can see an inverse behavior of the absorbance to the glucose concentration, so, it would be natural to expect to see similar behavior in saliva spectra.

However, in Fig. 5, we present the distribution of the absorbances of the FTIR spectra at wavenumber 1075 cm^{-1}, after calculating the Pearson correlation coefficient (r: 0.056), we determined that the relationship between absorbance and glucose level was null [13].

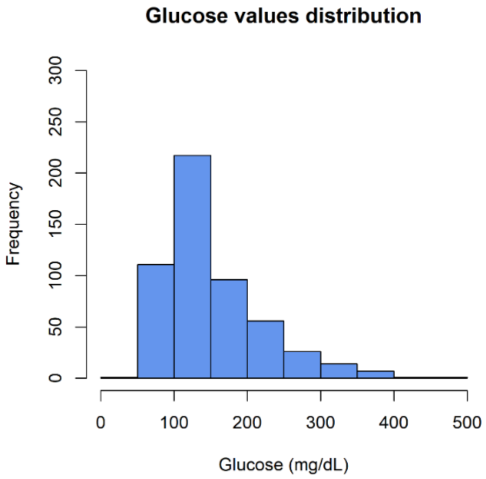

Fig. 3. Distribution of the glucose values recorded.

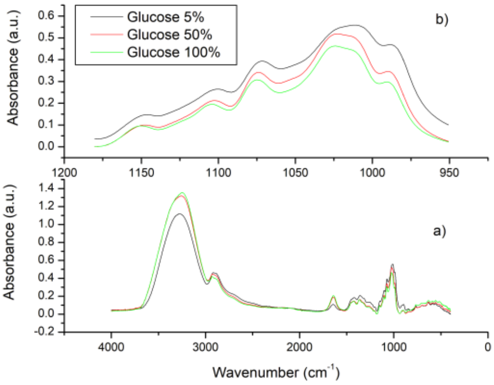

Fig. 4. Variations in the FTIR spectrum of glucose and injectable water solution.

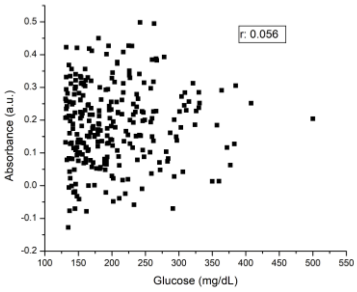

Fig. 5. Correlation between glucose level and absorbance of the FTIR spectra.

Table 2. Calculation of the RMSE-V for the methodologies analyzed.

Methodology	r	RMSE-V
ANN	-0.99	67.68
SVM	0.170	91.28
MLRM	-0.043	21646.21

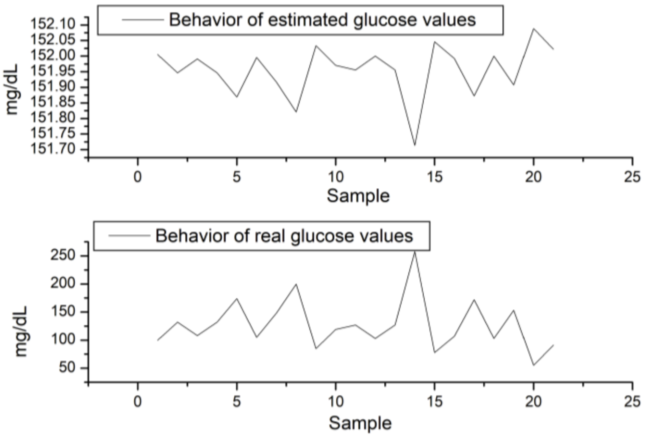

Fig. 6. The behavior of the glucose values of the first 20 samples in the database.

The null relationship presented in Fig. 5 is due to the complexity in the saliva constitution, while in Fig. 4, the spectrum involves only water and glucose, saliva includes several other components [14], this is the main obstacle in the use of FTIR spectra: the complexity of the samples [9]: "*Infrared spectroscopy works best on pure substances since all bands can be assigned to a single molecular structure. If a sample's composition is complex, its spectrum will be complex and it will be hard to know which infrared bands are due to which molecules.*", therefore, the use of techniques such as ANN and SVM are of great interest for the analysis of FTIR spectra.

After implementing regression models for ANN, SVM, and MLRM following the LOOCV methodology, the root-mean-square-error and *r* were calculated for the validation process (RMSE-V), obtaining the following results:

From Table 2, we can see that the best performance is obtained by ANN, however, the RMSE-V obtained is considerable. Analyzing of the amounts of glucose estimated by the ANN models built, we note that they all oscillate between 151 and 153 mg/dL when real values range from 47 to 503 mg/dL.

This allows us to infer that a large number of FTIR spectra of people who presented a glucose value of between 100 and 150 mg / dL as seen in Fig. 3 affects the performance of the ANN models, so it is advisable to have more homogeneous groups in terms of quantity to have a better estimate. The estimates of the model are considerably different, in several cases, from the real value obtained, however, evaluating the behavior of the estimates made, we note that it is very similar to the real one but inverted, this can be seen in Fig. 6.

Table 3. RMSE-V and r obtained through k-fold.

Methodology	r	RMSE-V
ANN	-0.928	78.8

After implementing *r* to evaluate the similarity of the signals, we obtained a value of -0.99, which indicates a strong inverse relationship [13]. Although LOOCV is a commonly used technique for the evaluation of some machine learning model, it is common that the k-fold cross validation (k-fold) methodology is also used to have a better perspective of the performance of the model [15]. The k-fold methodology consists of forming *k* groups from all the samples that make up the database distributed equally and using *k-1* groups for the model training process, which will be used to evaluate the group that did not participate in this process. Considering the results of Table 2, we present in Table 3 the results obtained after evaluating the database with ANN and the k-fold methodology with k = 10.

It is possible to appreciate from the results obtained in Table 3, that the r obtained, despite being reduced, is still a strong inverse correlation.

4 Conclusions

In the present work, the results obtained from the proposal of a model for the estimation of glucose in the blood through the analysis of FTIR spectra of saliva samples using machine learning techniques were presented.

After analyzing more than 500 samples, it is possible to infer that the best technique to estimate the glucose value with the proposed methodology is ANN since it allowed obtaining the lowest RMSE-V.

The RMSE-V obtained could be associated with the number of spectra that correspond to glucose values between 100 and 150 mg / dL, so it is of great interest to continue with the collection of spectra to balance the sample. Additionally, as presented in Fig. 6, probably the output emitted by the ANN model could be entered for one more subsequent process that allows the RMSE-V to be further reduced since, as can be seen in the mentioned figure, the behavior of the ANN estimates is similar but vertically inverted to the behavior of the actual glucose values. Such post-processes could be, in the first instance, the multiplication by a rotational matrix to invert the signal as well as some normalization process to scale the obtained estimates.

References

1. Who: https://who.int/news-room/fact-sheets/detail/noncommunicable-diseases (2011)
2. Diabetes.org: Treatment & Care|ADA. https://diabetes.org/diabetes/treatment-care (2020)
3. Stanford, K.I., Goodyear, L.J.: Exercise and type 2 diabetes: Molecular mechanisms regulating glucose uptake in skeletal muscle. Advances in Physiology Education, 38(4), pp. 308–314 (2014)
4. Larkin, P.: Infrared and Raman Spectroscopy: Principles and spectral interpretation. Elsevier (2011)
5. Sharma, B.K.: Spectroscopy. Krishna Prakashan Media, Goel Publishing House (1981)

6. Diabetes.org: A1C and eAG ADA. https://diabetes.org/diabetes/a1c-test-meaning/a1c-and-eag (2020)
7. Panta, P.: Oral Cancer Detection: Novel strategies and clinical impact. Springer (2019)
8. Shetty, P.K., Pattabiraman, T.N.: Salivary glycoproteins as indicators of oral diseases. Indian Journal of Clinical Biochemistry, 19(1), pp. 97–101 (2004)
9. Smith, B.C.: Fundamentals of Fourier transform infrared spectroscopy. CRC Press (2011)
10. Bel'skaya, L.V., Sarf, E.A., Makarova, N.A.: Use of Fourier transform IR spectroscopy for the study of saliva composition. Journal of Applied Spectroscopy, 85(3), pp. 445–451 (2018)
11. Jarvis, S., Crossley, S.A.: Approaching language transfer through text classification: Explorations in the detectionbased approach. Multilingual Matters (2012)
12. Scott, D.A., Renaud, D.E., Krishnasamy, S., Meriç, P., Buduneli, N., Çetinkalp, Ş., Liu, K.Z.: Diabetes-related molecular signatures in infrared spectra of human saliva. Diabetology and Metabolic Syndrome, 2(1) (2010)
13. PhD, M.S., Dontje, K.J.: Statistics for advanced practice nurses and health professionals. Springer Publishing Company (2014)
14. Wong, D.T.: Salivary diagnostics. John Wiley & Sons (2009)
15. Jarvis, S., Crossley, S.A.: Approaching language transfer through text classification: Explorations in the detection based approach. Multilingual Matters (2012)