| VOLUMEN 44 | NUMBER 4 | **SPECIAL ISSUE 2023**

ib

Una publicación de:
SOMIB
Sociedad Mexicana
de Ingeniería Biomédica

Revista Mexicana de Ingeniería Biomédica

Special Issue

Recent Advances in Artificial Intelligence for Biomedical Engineering

Edición Especial

Avances Recientes en Inteligencia Artificial en la Ingeniería Biomédica

http://www.rmib.mx

ISSN 2395-9126

SOMIB
Sociedad Mexicana
de Ingeniería Biomédica

Sociedad Mexicana de Ingeniería Biomédica

La Mesa Directiva de la Sociedad Mexicana de Ingeniería Biomédica hace una extensa invitación a las personas interesadas en participar, colaborar y pertenecer como Socio Activo de la SOMIB. La SOMIB reúne a profesionistas que se desarrollan en áreas de Ingeniería Biomédica, principalmente ingenieros biomédicos, así como otros profesionistas afines con el desarrollo de tecnología para la salud.

Membresía Estudiante
***$1,400.00* PESOS MXN**
15% de descuento para grupos de 5 o más personas.

Membresía Profesionista
***$2,400.00* PESOS MXN**
15% de descuento para grupos de 5 o más personas.

Membresía Institucional
***$11,600.00* PESOS MXN**
No aplica descuento.

Membresía Empresarial
***$20,000.00* PESOS MXN**
No aplica descuento.

EL PAGO CUBRE UN AÑO DE CUOTA. EN CASO DE REQUERIR FACTURA FAVOR DE SOLICITARLA, ADJUNTANDO COMPROBANTE DE PAGO Y ESPECIFICANDO CONCEPTO, AL CORREO ELECTRÓNICO: ***gerencia@somib.org.mx***

ib Revista Mexicana de Ingeniería Biomédica

AUTORES

Los trabajos a publicar en la RMIB, deben ser originales, inéditos y de excelencia. Los costos de publicación para autores son los siguientes:

NO SOCIOS:	***$200* DÓLARES AMERICANOS**
SOCIOS:	***$150* DÓLARES AMERICANOS**

PUBLICIDAD

A las empresas e instituciones interesadas en publicitar su marca o productos en la RMIB, los costos por número son los siguientes:

MEDIA PLANA:	***$4,999.00* PESOS MXN** (INCLUYE I.V.A.)
UNA PLANA:	***$6,799.00* PESOS MXN** (INCLUYE I.V.A.)
CONTRAPORTADA:	***$7,799.00* PESOS MXN** (INCLUYE I.V.A.)
FORROS INTERIORES:	***$7,799.00* PESOS MXN** (INCLUYE I.V.A.)

DESCUENTO DEL **20%** AL CONTRATAR PUBLICIDAD EN DOS O MÁS NÚMEROS.

Para ser socio

- Realiza el pago de derechos, de acuerdo a la categoría que te corresponde.
- Ingresa a *www.somib.org.mx/membresias* y elige el tipo de membresía por el cual realizaste el pago de derechos.
- Completa el formulario correspondiente y envíalo.
- Se emitirá carta de aceptación y número de socio por parte de la mesa directiva (aprobada la solicitud).
- Para mayor información sobre beneficios, ingresa a *www.somib.org.mx*; o escribe a *gerencia@somib.org.mx.*

Datos bancarios

- **Beneficiario:** *Sociedad Mexicana de Ingeniería Biomédica A. C.*
- **Banco:** *Scotiabank*
- **Referencia:** *1000000333*
- **Cuenta:** *11006665861*
- **CLABE Interbancaria:** *044770110066658614*

INFORMES

Juan Vázquez de Mella #481,
Polanco I Sección,
Alc. Miguel Hidalgo, C. P. 11510,
Ciudad de México, México,
(555) 574-4505
rib.somib@gmail.com

ib Revista Mexicana de
Ingeniería Biomédica

Fundador
Dr. Carlos García Moreira

COMITÉ EDITORIAL

Editora en Jefe
Dra. Dora-Luz Flores
UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Editores Asociados Nacionales

Dr. Christian Chapa González
UNIVERSIDAD AUTÓNOMA DE CIUDAD JUÁREZ

Dra. en C. Citlalli Jessica Trujillo Romero
DIVISIÓN DE INVESTIGACIÓN EN INGENIERÍA MÉDICA
INSTITUTO NACIONAL DE REHABILITACIÓN "LUIS GUILLERMO IBARRA IBARRA"

Dr. Rafael Eliecer González Landaeta
UUNIVERSIDAD AUTÓNOMA DE CIUDAD JUÁREZ

Dra. Rebeca Romo Vázquez
UNIVERSIDAD DE GUADALAJARA

Dra. Isela Bonilla Gutiérrez
UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ

Comité Editorial Internacional

Dr. Leonel Sebastián Malacrida Rodríguez
UNIVERSIDAD DE LA REPÚBLICA, URUGUAY

Dra. Elisa Scalco
INSTITUTE OF BIOMEDICAL TECHNOLOGY
ITALIAN NATIONAL RESEARCH COUNCIL, MILAN, ITALY

Dra. Natali Olaya Mira
INSTITUTO TECNOLÓGICO METROPOLITANO
ITM, MEDELLÍN, COLOMBIA

Índices
La Revista Mexicana de Ingeniería Biomédica aparece en los siguientes índices científicos:
Sistema de Clasificación de Revistas Científicas y Tecnologías del CONACYT - Q4, SCOPUS, SciELO, EBSCO, LATINDEX, Medigraphic Literatura Biomedica, Sociedad Iberoamericana de Información Científica - SIIC.

www.rmib.mx
ISSN 2395-9126

Asistente Editorial
Carla Ivonne Guerrero Robles

Editor Técnico y en Internet
Sandra Sánchez Jáuregui

Se autoriza la reproducción parcial o total de cualquier artículo a condición de hacer referencia bibliográfica a la Revista Mexicana de Ingeniería Biomédica y enviar una copia a la redacción de la misma.

SOMIB
Sociedad Mexicana
de Ingeniería Biomédica

Sociedad Mexicana de Ingeniería Biomédica
Juan Vázquez de Mella #481, Polanco I Sección, Alc. Miguel Hidalgo, C. P. 11510, Ciudad de México, México, (555) 574-4505

SOMIB
Sociedad Mexicana
de Ingeniería Biomédica

Afiliada a:
International Federation of Medical and Biological Engineering (IFMB-IUPSM-ICSU)
Federación de Sociedades Científicas de México, A.C. (FESOCIME)
Consejo Regional de Ingeniería Biomédica para América Latina (CORAL)

SOMIB
Juan Vázquez de Mella #481, Polanco I Sección, Alc. Miguel Hidalgo, C. P. 11510, Ciudad de México, México (555) 574-4505
www.somib.org.mx

REVISTA MEXICANA DE INGENIERÍA BIOMÉDICA, Vol. 44, No. 4, Special Issue 2023, es una publicación cuatrimestral editada por la Sociedad Mexicana de Ingeniería Biomédica A.C., Juan Vázquez de Mella #481, Polanco I Sección, Alc. Miguel Hidalgo, C. P. 11510, Ciudad de México, México, (555) 574-4505, www.somib.org.mx, rib.somib@gmail.com. Editora responsable: Dra. Dora-Luz Flores. Reserva de Derechos al Uso Exclusivo No. 04-2015-041310063800-203, ISSN (impreso) 0188-9532; ISSN (electrónico) 2395-9126, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este número: Lic. Sandra Sánchez Jáuregui, Juan Vázquez de Mella #481, Polanco I Sección, Alc. Miguel Hidalgo, C. P. 11510, Ciudad de México, México, (555) 574-4505, fecha de última modificación, 31 de agosto del 2023.

Disponible en línea:
www.rmib.mx

CONTENTS - *CONTENIDO*

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 6-22

Revista Mexicana de Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.44.4.1

E-LOCATION ID: **1348**

Comparison of Spectral and Sparse Feature Extraction Methods for Heart Sounds Classification

Comparación de Métodos de Extracción de Características Espectrales y Dispersas para Clasificación de Sonidos Cardíacos

Roilhi Frajo Ibarra-Hernández[1], *Miguel Ángel Alonso-Arévalo*[2], *Eloísa del Carmen García-Canseco*[3]

[1] Universidad de Ensenada, Ensenada - México

[2] Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada, Baja California - México

[3] Universidad Autónoma de Baja California, Ensenada, Baja California - México

ABSTRACT

Cardiovascular diseases (CVDs) remain the leading cause of morbidity worldwide. The heart sound signal or phonocardiogram (PCG) is the most simple, low-cost, and effective tool to assist physicians in diagnosing CVDs. Advances in signal processing and machine learning have motivated the design of computer-aided systems for heart illness detection based only on the PCG. The objective of this work is to compare the effects of using spectral and sparse features for a classification scheme to detect the presence/absence of a pathological state in a heart sound signal, more specifically, sparse representations using Matching Pursuit with multiscale Gabor time-frequency dictionaries, linear prediction coding, and Mel-frequency cepstral coefficients. This work compares the performance of PCGs classification applying features as a result of averaging the samples or the features for each PCG sound event when feeding a random forest (RF) classifier. For data balancing, random under-sampling and synthetic minority oversampling (SMOTE) methods were applied. Furthermore, we compare the Correlation Feature Selection (CFS) and Information Gain (IG) for the dimensionality reduction. The findings show a SE=93.17 %, SP=84.32 % and ACC=85.9 % when joining MP+LPC+MFCC features set with an AUC=0.969 showing that these features are promising to be used in heart sounds anomaly detection schemes.

KEYWORDS: classification, heart sounds, matching pursuit, spectral features, time-frequency representation

RESUMEN

Las enfermedades cardiovasculares (ECVs) han persistido como la principal causa de mortalidad en el mundo. La señal de audio cardiaco o fonocardiograma (FCG) es la herramienta más simple, efectiva y de bajo costo para auxiliar a especialistas diagnosticando ECVs. Los avances en el procesamiento de señales y aprendizaje máquina han motivado el diseño de auscultación y detección computarizada. El objetivo de este trabajo es comparar el uso de características espectrales y *dispersas* para un sistema de clasificación que detecte la presencia/ausencia de una patología en un audio cardiaco mediante representaciones dispersas usando *Matching Pursuit* con diccionarios de Gabor tiempo-frequencia, predicción lineal y coeficientes cepstrales Mel. Se crearon 5 conjuntos de características como resultado de combinar las características para cada FCG y se examinó su desempeño usando un clasificador de bosque aleatorio (RF). Se aplicaron métodos de balanceo de muestras basados en sobremuestreo (SMOTE) y submuestreo aleatorio. Se compararon métodos de selección de características por correlación (CFS) y ganancia de información (IG) para *reducir la dimensionalidad* del conjunto. Los resultados muestran métricas de SE=93.17 %, SP=84.32 % y ACC=85.9 % al juntar los parámetros MP+LPC+MFCC además de una AUC=0.969. El trabajo muestra el potencial de las características espectrales y escasas para la detección de patologías en señales de audio cardiaco.

PALABRAS CLAVE: características espectrales, clasificación, matching pursuit, representación tiempo-frecuencia, sonidos cardiacos

Corresponding author

TO: **Roilhi Frajo Ibarra-Hernández**
INSTITUTION: **Universidad de Ensenada**
ADDRESS: **Mar 198, Tercer Ayuntamiento, CP 22830. Ensenada BC.**
CORREO ELECTRÓNICO:
roilhi.ibarra@universidaddeensenada.edu.mx

Received:

27 March 2023

Accepted:

21 June 2023

INTRODUCTION

Heart diseases remain the leading cause of death worldwide, according to a report from the World Health Organization [1]. An effective method that leads to the primary diagnosis of heart illness is automatic abnormal heart sound detection, which aims to identify the presence of a cardiac malfunction. This area has raised interest among researchers with the introduction of electronic stethoscopes and the advances in signal processing. In general, the methods for diagnosing pathological states of heart sounds consist of two stages: firstly, the feature extraction process to obtain the most representative parameters of cardiac sound, and secondly, the classification, which predicts the patient's condition from the patterns found in the extracted features. In healthy individuals (adults), the heart sound signal, also known as phonocardiogram (PCG), comprises two main components called *fundamental heart sounds* (FHS), which are denoted as s_1 and s_2. Usually, a typical time duration and low-frequency spectral content characterize each FHS. For instance, the s_1 components dominate the region from 10 Hz to 140 Hz, while the s_2 components usually concentrate their energy around the 10 Hz to 200 Hz band [2].

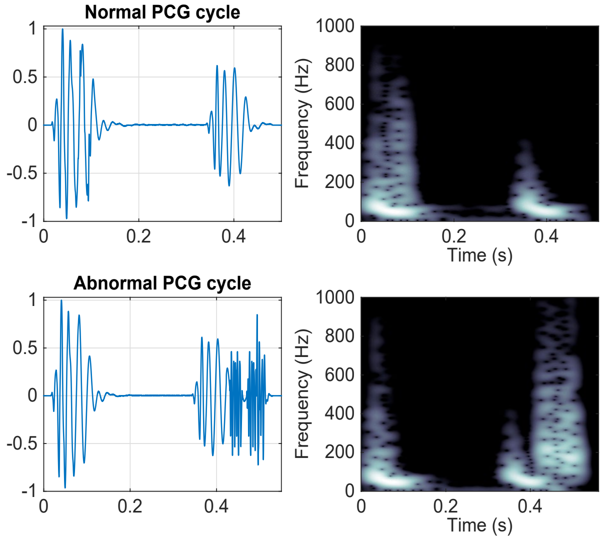


FIGURE 1. Top: time waveform and spectrogram of a normal PCG signal. Bottom: time waveform and spectrogram of an abnormal (pathological) PCG signal.

In pathological conditions, sounds named *murmurs* appear. Murmurs are sounds stemming from a *turbulent blood flow* due to a valve malfunction or an obstruction, denoting a pathological or abnormal state. The energy distribution of murmurs in frequency varies widely and, depending on their nature, can go above 800 Hz. Unfortunately, the frequency content of murmurs can overlap with the distribution of s_1 and s_2, and thus, the correct identification of the sound is a difficult task that requires sophisticated methods to determine the type of sound. Figure 1 illustrates the waveform and the time-frequency content representative spectrogram of a PCG cardiac cycle in both normal and pathological states.

Review of PGCs classification schemes

A thorough review of existing methods to classify heart sounds is out of the scope of this work. However, the 2016 PhysioNet/Computing in Cardiology Challenge (CinC) [3] and the release of one of the more extensive public databases of PCG recordings are a milestone in the field. To provide a literature review for PCGs classification algorithms, we can organize them into two main categories according to a) the feature extraction methods and b) the classification schemes used by each research paper.

For the first category, the feature extraction methods aim to represent the cardiac sound signals in different domains (time, frequency, and joint time-frequency, mainly), revealing the main physiological and pathological PCG attributes to allow an effective feature extraction. Since the PCG signal is quasi-stationary, the features provided would be able to capture concurrent variations and the structural components in time, frequency, and joint time-frequency domains. For these reasons, selecting an adequate feature extraction method is crucial for classifying heart sound signals. For instance, in the time-frequency domain representation of the PCG, researchers have chosen the short-time Fourier Transform (STFT) [4][5][6], Wigner-Ville distributions [7][8], the empirical [9], discrete [10], and continuous Wavelet transform [11][12][13][14]. Among other frequency domain features utilized for PCGs classification, the Mel-Frequency Cepstral Coefficients (MFCCs) have been widely used as classification input features [15][16][17][18][19][20][21] since these parameters are the most popular to characterize the envelope information for audio signals successfully. The Linear Predictive Coefficients (LPCs) [22] have also been used to capture PCG signal spectrum patterns. The second category comprises the selection of a classification scheme which is essential since it is the final step of a PCGs murmur detection algorithm. The classifier takes the extracted features and interprets them by extracting and recognizing the functional patterns to efficiently represent the murmurs associated with diseases of a PCG signal. In the state-of-the-art, the reported classification schemes used are Support Vector Machines [8][16][23][24][25][26][27][28][29], k-Nearest Neighbors [16][30][31][32][33][34], and Random Forests techniques [14][35][36][37], in terms of conventional Machine Learning techniques. On the other hand, reported Deep learning-based methods for PCGs classification are comprised of ensembles of neural networks [15][17][38][39][40], convolutional neural networks (CNN) [6][13][21][41][42][43][44][45][46], long short-term memory networks (LSTM) [47][48][49], and recurrent neural networks (RNN) [50]. Although deep learning has emerged as a powerful approach that has shown promising advances in PCGs classification, there are still limitations due to the lack of data, carrying out training inefficiency, and insufficiently robust models [51][52][53]. Deep learning algorithms require significantly more computational resources and may not be feasible for machines with embedded or limited hardware capabilities. Deep learning algorithms might also present a limitation called the exploding and vanishing gradient descent problem, which causes the classification error rate to increase after attaining a minimum value. This deficiency is also known to cause model overfitting. Another limitation of deep learning is the lack of interpretability of the features to the point where it is impossible to discern what they are and have no physical meaning. While that may be a reasonable price for the theoretical performance gain in some applications, we consider it vital to understand the physiological phenomena in PCG analysis.

This work aims to leverage sparse representations to classify heart sounds. More specifically, Matching Pursuit (MP) coefficients combined with LPCs and MFCCs as features feed our proposed high-performance scheme that detects PCG abnormalities. We selected the Random Forest classifier as a classification algorithm due to its simplicity, low computational requirements, and excellent performance. The RF classifier is still used among researchers to detect pathological states from heart sounds. However, it is noteworthy that deep learning algorithms have significantly improved in recent years and are now the default go-to choice for many problems, especially in computer vision and natural language processing fields.

On the other hand, we used the Synthetic Minority Oversampling technique (SMOTE) to address the problem of unbalancing during the classification by creating synthetic samples for the minority class (abnormal or pathological PCG sound signals). The classification scheme's performance has been analyzed when the inputs are noisy PCG recordings. Finally, the work compared the performance of two feature selection techniques.

This paper presents our study's methodology, results, and conclusions, which aim to investigate the effectiveness of using sparse and spectral features for PCG signal classification. The methodology section describes the methods we used to conduct our experiments, including the selection of datasets, the choice of algorithms, and the evaluation metrics. The results section presents the findings of our investigations, including the performance of different algorithms and their comparison. Finally, in the conclusion section, we summarize our study's key insights and implications, and the limitations and future research directions.

MATERIALS AND METHODS

The main goal of this research is to evaluate the classification performance of different sets of features as input parameters of a classifier to accurately detect pathological states in PCG signals. The Physionet/CinC 2016 is the largest database of PCG signals publicly available to the scientific community [54] in order to evaluate algorithms to segment and classify PCGs. It comprises the merge of six different research groups of recordings from subjects under normal and various pathological cardiac conditions. Specifically, the database includes 3,153 sounds recorded with a 2,000 Hz sampling frequency. Moreover, 2,488 samples come from cardiac sounds of subjects under normal conditions, while 665 represent an abnormal category.

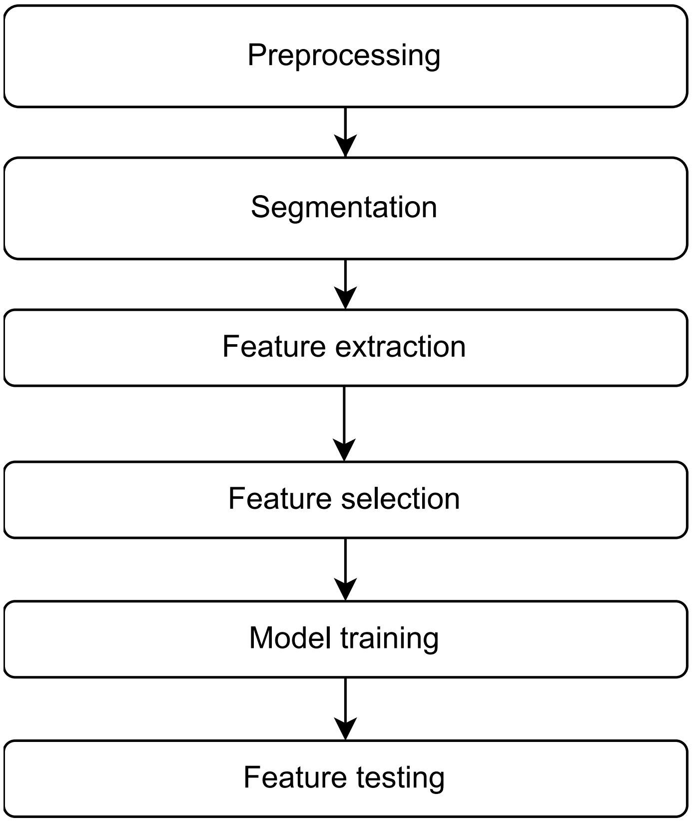


FIGURE 2. **Block diagram to describe the sequence of methods to conduct the experiment of this paper.**

In this paper, we conducted the methodology shown in the block diagram of Figure 2. For the preprocessing stage, the PCG signals were band-pass filtered between 25-600 Hz using a sixth-order Butterworth filter; then, we applied a normalization procedure in amplitude, which consists of dividing the recording samples by the maximum value. The second stage comprises the extraction of FHS, since for these events in the following step, different features will be extracted. The feature selection stage consists of reducing the number of features in order to know which of them are the most relevant and have the most information. Finally, in the training stage, we feed a classifier algorithm using the different sets of features to evaluate the classification performance of each one of them.

Matching Pursuit

The Matching Pursuit algorithm (MP), proposed by Mallat [55], is a greedy and iterative method that computes a *sparse* representation of a signal s as a linear combination of M_a *elementary waveforms* called *atoms* with *minimal error*. Each atom $\hat{g}m$ belongs to a redundant set of all possible predefined signals called dictionary **D**. MP selects the best-correlated atom $\hat{g}m$ iteratively to provide a *sparse* decomposition in the following way:

$$\boldsymbol{s} = \sum_{m=1}^{M_a} \alpha_m \cdot \hat{\boldsymbol{g}}_m + \boldsymbol{r} \, , \tag{1}$$

the atom that MP chooses at each iteration is the one that best matches the local signal structure of *s* by calculating the maximum inner product between the signal and the dictionary:

$$\hat{\boldsymbol{g}}_m = \operatorname{argmax}_{\hat{\mathrm{g}}_\mathrm{m} \in \boldsymbol{D}} |\langle \boldsymbol{r}, \hat{\boldsymbol{g}}_m \rangle| \, , \tag{2}$$

the weighting factor α_m is a scalar that comes from the value of the inner product at each iteration:

$$\alpha_m = \langle \boldsymbol{r}, \widehat{\boldsymbol{g}}_m \rangle, \tag{3}$$

r is a signal called the *residual* term. It comes from the difference between the signal and the weighted-selected atom:

$$\boldsymbol{r} = \boldsymbol{r} - \alpha_m \cdot \widehat{\boldsymbol{g}}_m, \tag{4}$$

notice that at the beginning of the algorithm ***r=s.*** MP is called a *greedy* method to reconstruct sparse signals because it stops until a desired number of iterations (or atoms) M_a or the ratio between the original signal energy and the residual has been reached. The dictionary selection is a crucial step for the MP decomposition into atoms. The dictionaries of Gabor functions have been widely used for the reconstruction of PCG signals due to the accurate signal representation in the time-frequency domain [56][57][58][59][60][61]; nonetheless, Gabor atoms are *well-concentrated* waveforms in both time and frequency. In this work, we use as a dictionary a set of predefined multiscale functions, which is a collection $\boldsymbol{D}=\boldsymbol{U}^{J}_{j=1}\boldsymbol{D}_j$ of blocks $\boldsymbol{D}_j$ of time-frequency atoms at different scales. A Gabor atom in a multiscale dictionary is a waveform defined by the modulation, dilation, translation, and sampling of a continuous window $w_j(t)$ as:

$$\boldsymbol{g}_{j,n,k}(m) = w_j\left(mT_s - nT_j\right)\ \exp\left(\frac{2i\pi kmT_s}{K_j}\right) \text{ for } 1 \le m \le M\,, \tag{5}$$

where the time location or window shift is defined as nTj, the window length or scale L_j and is modulated at a frequency k/K_j , where K_j is a predefined number of possible frequencies (according to the FFT size), T_s is the sampling period, and M the number of samples. Figure 3 shows the time waveform of a Gabor atom, which can be seen as a cosine-modulated Gaussian window. At the right panel, a couple of waveforms illustrate the effect of changing the modulation frequency.after the frequency of the signal has been warped into the Mel scale, each C_n MFCC coefficient is calculated as follows:

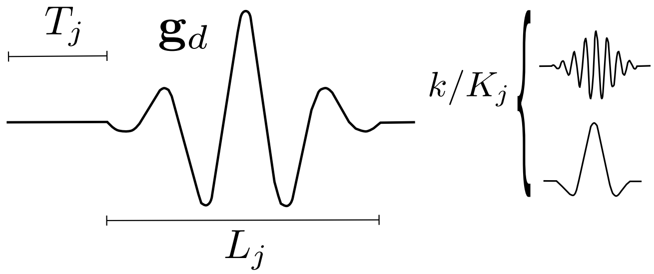


FIGURE 3. Time waveform of a Gabor atom and its defined parameters. In the right panel, the waveforms illustrate the effect of changing the frequency parameter k/Kj.

Linear Predictive Coding

As seen in equation (1), MP decomposes a signal in two main parts, a linear combination of Gabor atoms and a residual. The residual term ***r*** is expected to be lowly correlated with the selected dictionary atoms. Thus, it must be expressed differently to be integrated as a feature representing the PCG signal. For this reason, instead of reconstructing the temporal waveform, we propose to represent ***r*** using the Linear Predictive Coding technique [62], which approximates the signal's spectrum rather than the time domain waveform. The LPC representation is an all-pole filter where the residual ***r*** can be predicted as a linear combination of the previous samples:

$$\boldsymbol{r}_n = -\sum_{i=1}^{\mathrm{p}} h_i\, \boldsymbol{r}_{n-i} + \boldsymbol{e}_n, \tag{6}$$

where n=0,⋯,*N*-1, e_n is the final residual, and *p* is the filter order. Filter coefficients h_i are added to the features set. Published works in the literature review have used the LPC coefficients as features for the automated detection of heart murmurs in PCG signals [22].

Mel-Frequency Cepstral Coefficients (MFCCs)

The Mel-Frecuency Cepstral coefficients (MFCCs) are the predominant features used for speech recognition [63], because they provide a compact and *smooth representation* of the magnitude spectrum. MFCCs are based on the human hearing physiological structure since the human perception of the frequency content of sounds does not follow a linear scale. Thus, having a signal with a fundamental frequency *f* and an estimated pitch should be measured on a ranking called the Mel Scale. The MFCC coefficients are calculated by taking the discrete cosine transform of a logarithmic spectrum after it was warped to the Mel scale as follows:

$$\mathrm{Mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right), \qquad (7)$$

after the frequency of the signal has been warped into the Mel scale, each C_n MFCC coefficient is calculated as follows:

$$C_n = \sum_{m=1}^{M_c} D_m \ cos\left(n\left(\frac{m-0.5}{M_c}\right)\pi\right), \qquad (8)$$

where D_m is the output of the *k*-th triangular filter bank channel and M_c is the number of filter bank channels. In our implementation, we use M_c=14 to cover the range from 20 Hz to 900 Hz. Figure 4 shows the representation in the frequency domain (Hz and Mel scale) of the triangular filter bank used for the MFCC coefficients extraction.

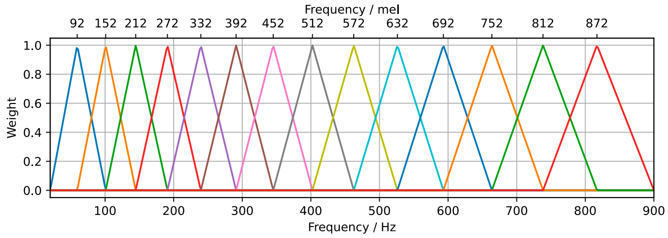


FIGURE 4. **Triangular Mel Filter bank to extract MFCC coefficients used in this work. The frequency is shown at the top in the Mel Frequency and at the bottom in Hertz, respectively.**

Reported research frameworks have used MFCCs as features for PCGs classification [15][16][17][18][19][20][21], since they provide meaningful representations in the spectral envelope rather than time features.

Random Forest Classifier

The random forest classifier (RF) comes from combining two or more *decision tree* classifiers. Each classifier uses a random vector sampled independently from the input vector and casts a *unit vote* for the most popular class. The features used are randomly selected to grow a tree. RF uses a *bagging* method to randomly replace the *N* examples of the original training set [64].

Let Θ be a random vector that chooses a random subset x from the training set *X*. Let N_T be the number of decision trees; each one has an additional parameter Θ_t and the ensemble of trees consists of the set $\{f_1(x,\Theta_1), f_2(x,\Theta_2),\ldots,f_{NT}(x,\Theta_{NT})\}$. The RF algorithm attempts to reduce the variance of the model by averaging many trees estimates as follows:

$$f_{RF}(x) = \frac{1}{N_T}\sum_{t=1}^{N_T} \alpha_t\, f_t(\boldsymbol{x}, \Theta_t), \qquad (9)$$

Where α_t represents an associated weight. Because of its simplicity and promising results, the RF classifier has been widely used for PCG signals classification [14] [65][66][67][68][69]. It is still a valuable method to detect heart murmurs accurately. For the experiments conducted in this research, we choose as hyperparameters a number of estimators N_e=100, and the Gini criterion to measure the quality of the splits.

The Synthetic minority oversampling technique (SMOTE)

Most PCG datasets around the reported research works contain more recordings from healthy people (commonly labeled as *normal sounds*) than people with a heart pathology (commonly labeled as *abnormal sounds*). The training stage will be affected due to this unbalancing between class samples, causing overfitting and highly biased results. The Synthetic minority oversampling technique is an algorithm that addresses the unbalancing problem by creating synthetic samples of the minority class. These synthetic samples are generated over the feature space rather than the data space. Each minority sample is created by taking the difference between the feature vector (input sample) and its nearest neighbor. The difference is then multiplied by a random number between 0 and 1 and added to the feature vector under consideration. The SMOTE approach effectively forces the decision region of the minority class to be more general, causing a better performance in a classification that uses decision trees. It has been shown that SMOTE technique performs better in accuracy than under-sampling methods [70].

Feature extraction

We have previously evaluated several time-frequency dictionaries to decompose the PCG, showing that Gabor wavelets accurately represent this signal [71]. For the experiments conducted in this research, the selected number of atoms was M_a=15 in order to reach almost 99 % of the energy to reconstruct a PCG cycle. For the LPC analysis, the number of coefficients was p=15. For the MFCCs, we followed the suggestion proposed by some methods in the context of the Physionet Challenge [15][17][38], setting the number of coefficients as M_c=14. Figure 5 provides a block diagram which describes the feature sets used in our experiments. We generated five feature sets labeled as follows by combining the MP, MFCC, and LPC approaches. Set A contains 90 features (i.e., columns of A data frame are 90) by merging the MP+LPC parameters, while set B contains the same extracted features as set A; however, they are extracted after performing cycle averaging. Set C consists of 146 features after combining MP+MFCC+LPC, set D contains 131 features after joining MP+MFCC. Set E contains 56 features by considering only MFCC. The procedure of feature extraction was conducted in MATLAB ©.

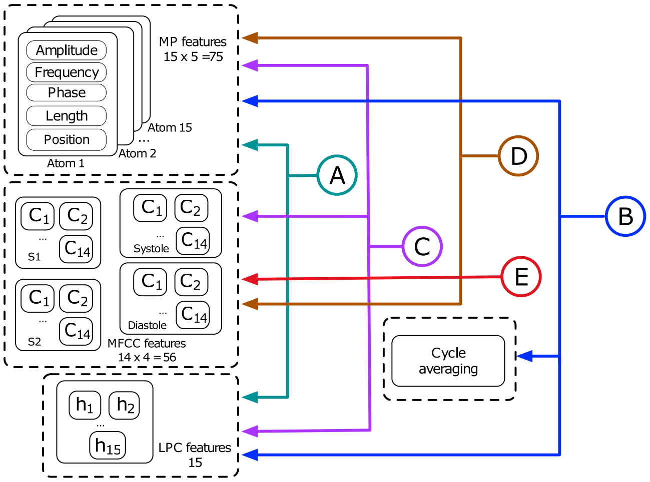


FIGURE 5. Block diagram of feature sets used in this work. In the first stage, we extracted M_a=15 atoms, with five parameters per atom yielding a total of 75 MP features. Using M_c=14 for each of the four states in the PCG cycle yields a total of 56 MFCC features. We defined p=15 as the number of LPC features.

Processing of the low-quality recordings

The database described at the beginning of this section comes from the Physionet/CinC 2016 challenge. Data includes the PCG recordings, the FHS time segmentation boundaries, a label indicating the pathological condition (normal/abnormal), and a quality reference of the PCG signal. According to the noise level in the sound samples, the database is divided into High-Quality Recordings (HQR) and Low-Quality Recordings (LQR). Due to their highly noisy condition, there are 279 signals labeled as LQR, and their FHS time segmentation boundaries are not provided. Since most of the features are calculated per cardiac cycle, in the case of LQR the computation of the features was conducted by splitting the PCG in segments of $T\mu + \sigma_T$=1.15 s, where Tμ is the average cardiac cycle duration for the recordings labeled as abnormal and σ_T is the respective standard deviation. However, to compute the MFCCs each PCG segment was sliced into 4 windows according to the average duration of the FHS [2].

RF classifier settings

We changed the number of estimators for the RF method to 100, as recommended in the presence of unbalanced datasets [72][73]. Specific details and parameter settings used during the evaluation are provided in previous work [74], where the RF classifier outperformed the others. This evaluation and all the classification tests were conducted using the scikit-learn toolbox under Python [75]. The experiments presented in this paper were conducted on a workstation with an Intel i7-9750H processor (2.60 GHz) and NVIDIA GPU GTX 1660Ti.

The confusion matrix is a well-known method used to evaluate the performance of ML classification schemes. In our case, for a binary class problem (having *normal* and *abnormal* labels), the confusion matrix has four values:

-True positives (TP): number of correctly identified PCGs with a pathological condition.

-True negatives (TN): number of correctly classified PCGs that do not have a pathology.

-False positives (FP): number of PCG signals labeled as abnormal but classified as *normal*.

-False positives (FN): number of PCG signals labeled as normal but classified as *abnormal*.

For the experiments conducted in this research, we considered these quantities in order to calculate the following classification metrics:

- Accuracy (ACC)

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \times 100 \in [0\ 100]. \qquad (10)$$

- Sensitivity (SE)

$$SE = \frac{TP}{TN + FP} \times 100 \in [0\ 100]. \qquad (11)$$

- Specificity (SP)

$$SP = \frac{TN}{TN + FP} \times 100 \in [0\ 100]. \qquad (12)$$

- Matthews Correlation Coefficient (MCC) (13)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FP)(TP + FN)(TN + FP)}} \in [-1,1].$$

Nonetheless, to evaluate the classifier performance in terms of adding more data to the training set, we calculate learning curves using a size from 0 to 2500. We only made that computing from feature set C, since it has all types of parameters extracted. Figure 6 illustrates such a calculation.

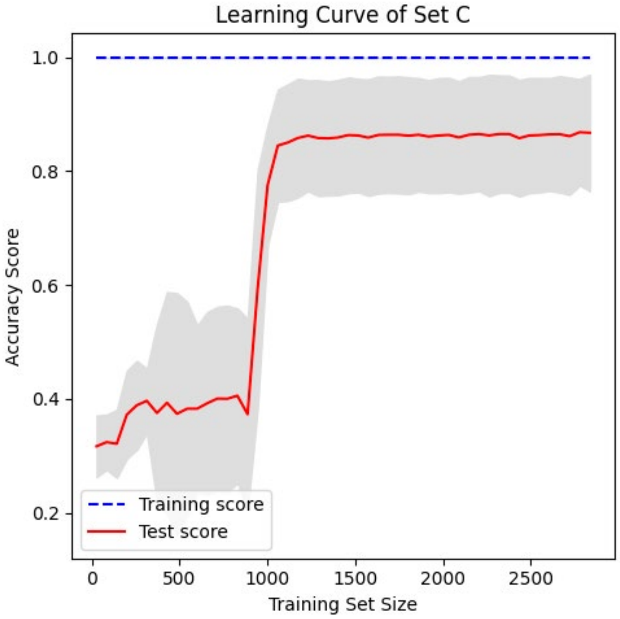


FIGURE 6. **Learning curves generated to evaluate the performance of classification by RF and all parameters extracted (Feature set C).**

RESULTS AND DISCUSSION

Under-sampling vs. oversampling

Since we have an unbalanced set of samples (78.9 % labeled as normal while 21.1 % as abnormal) it was necessary to implement a strategy to equalize the number of samples (rows of a data frame) for each class (i.e., to have the same number of samples labeled as normal or abnormal in each data frame). Most of the algorithms tackle this issue by randomly under-sampling the majority class; however, the main drawback of this technique is that potentially useful information contained in the ignored samples is neglected. We address the unbalanced classes problem by adopting two strategies: dropping entities of the majority class and oversampling of the minority class using SMOTE [70]. In heart sound classification, obtaining new recordings labeled as abnormal is not a simple task (there will always be more healthy people). SMOTE allows us to use the already acquired data and create "new" samples in the feature space as if it were possible to access more pathological heart sounds. Table 1, section I shows the classification scores when using the input features from data frames A-E and comparing the abovementioned balancing techniques. When oversampling is applied, the *SP, ACC,* and *MCC* all increased, while *SE* has decreased.

TABLE 1. **Results from the PCG sounds classification after splitting the recordings in High Quality (HQR) and Low Quality (LQR) labels.**

Dataset	Balancing	SE	SP	ACC	MCC	Section
A	undersampling	88.06	79.03	81.3	0.61	I: HQR+LQR
B		74.22	70.13	71.16	0.4	
C		93.72	83.27	85.9	0.7	
D		94.34	82	84.32	0.67	
E		91.2	84.12	86.69	0.72	
A	oversampling	71.07	94.07	88.28	0.68	
B		27.05	95.56	78.29	0.33	
C		81.14	95.98	92.24	0.8	
D		79.25	95.56	91.45	0.77	
E		77.99	92.59	88.91	0.71	
A	undersampling	76.99	82.47	81.39	0.52	II: HQR
B		65.49	72.51	71.13	0.32	
C		87.61	82.25	83.3	0.6	
D		61.95	92.86	86.78	0.57	
E		87.61	82.25	83.3	0.6	
A	oversampling	61.95	92.86	86.78	0.57	
B		28.32	95.89	82.61	0.34	
C		76.11	93.94	90.43	0.7	
D		78.76	94.16	91.13	0.72	
E		78.76	91.13	88.7	0.66	
A	undersampling	70	69.44	69.64	0.38	III: LQR
B		60	58.33	58.93	0.18	
C		80	86.11	83.93	0.65	
D		80	86.11	83.93	0.65	
E		85	75	78.57	0.58	
A	oversampling	55	75	67.86	0.3	
B		45	83.33	69.94	0.31	
C		75	94.44	87.5	0.72	
D		75	91.67	85.71	0.68	
E		70	83.33	78.57	0.53	

Effects of signal quality on performance

We analyzed the influence of signal quality in the algorithm performance. According to the noise condition labels mentioned in the low-quality recordings subsection, we evaluated the signals tagged as HQR (2,874) and LQR (279) separately. The oversampling and under-sampling balancing procedures were also considered. Table 1 also presents the results for the HQR and LQR, respectively. As expected, the scores are generally higher for the HQR compared to the highly noisy recordings. A more detailed analysis of the results is provided in the following section.

Training time evaluation

To assess the practicality and efficiency of the proposed schemes, we evaluated the training time of the algorithm. The results are shown in Figure 7; it can be seen that for 2,500 features in the training set, the computation time is below 4 seconds. This result is unsurprising since the amount of data is relatively small to use more sophisticated and computationally expensive classification algorithms, such as Deep Learning methods.

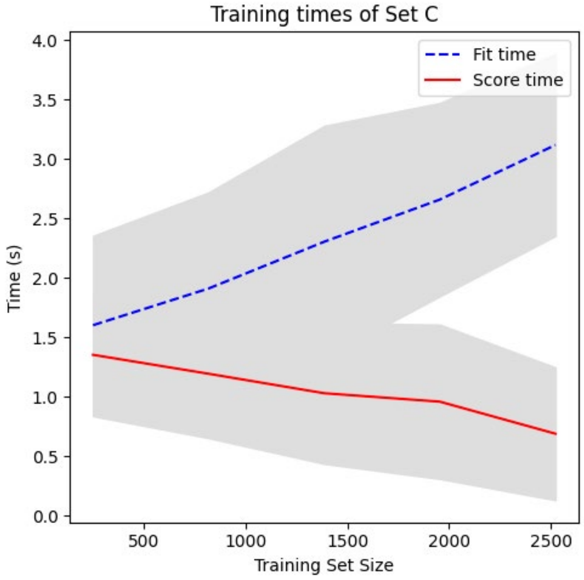


FIGURE 7. **Time to fit/train the proposed algorithm and score or evaluate new instances. Results are shown for set C, since it contains the highest number of features.**

Feature selection

In order to keep the best parameters for classification and reduce overfitting and computational complexity in the proposed classification scheme, we implemented feature selection. There are features or variables in our data that are the most relevant, i.e., those that contribute the most to the output prediction. In the present work, we applied the Correlation Feature Selection (CFS) [76] and Information Gain (IG) methods [77] for this task. The reduced subset was constructed for the IG method, neglecting features that provided a null information gain (zero). Table 2 shows the feature selection results, presenting the number of features before and after. The CFS method works significantly better regarding dimensionality reduction, keeping only between 12 and 30 features, while IG varies between 50 and 116 attributes.

TABLE 2. **Number of features in the datasets A-E originally produced, then number of reduced features after applying CFS and IG feature selection.**

Set	Original	Reduced CFS	Reduced IG
A	90	22	66
B	90	20	65
C	146	30	116
D	131	22	101
E	56	12	50

This section presents the classification performance evaluation when using feature selection by comparing the Receiver Operating Characteristics (ROC) curves and the calculation of the Area Under the Curve (AUC) for each feature set. Figure 8 shows the ROC curves for feature sets A-E. In this case, features set D exhibits the best performance since it has an AUC=0.97, the highest score obtained. However, set C is relatively close, showing AUC=0.969. This result suggests improving classification performance when adding MP and LPC parameters rather than only using MFCC or MP+LPC features. For cycle averaging, set B got the worst AUC score (0.79).

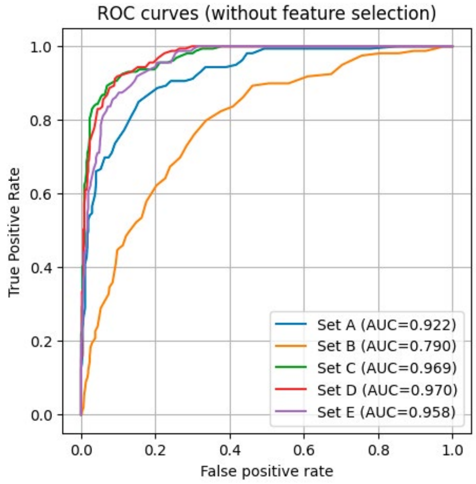


FIGURE 8. **ROC curves and AUC for feature sets A-D without feature selection.**

In another experiment, the ROC curves and AUC calculation was conducted when applying the CFS feature selection algorithm for each feature set, see Figure 9. There is an improvement in ACU scores since this metric increases for all feature sets. However, features set E now shows the best performance with an AUC=0.967. Feature sets C and D present a close AUC score of 0.961 and 0.967, respectively. On the other hand, although there is an increase from 0.76 to 0.867 for the AUC score of set B it is still the lowest obtained.

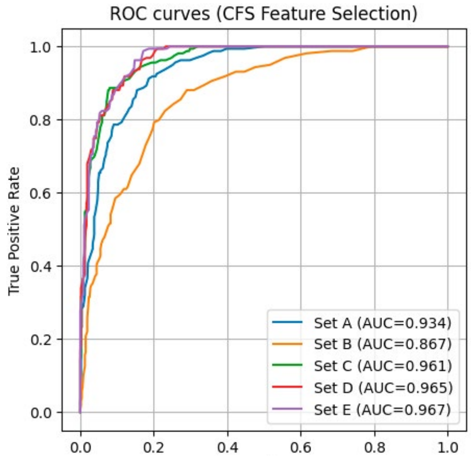


FIGURE 9. **ROC curves and AUC for feature sets A-D after applying Correlation Feature Selection (CFS).**

Finally, the same experiment was conducted but now using the features selected by the IG algorithm. Figure 10 shows the result. There is an improvement compared with results shown in Figure 9, when using CFS. However, feature sets C and D now show the best AUC score of 0.971, while in set D the AUC score is close (0.969). Set B is still presenting the worst ACU score (0.971).

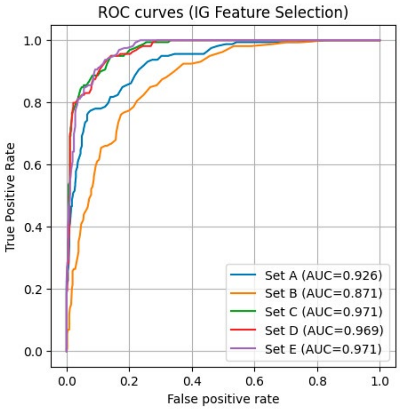


FIGURE 10. ROC curves and AUC for feature sets A-D after applying Information Gain Feature Selection (IG).

Discussion

This work aimed to compare different feature extraction schemes based on spectral and sparse representations for the automated classification of heart sounds. A low-cost system with accurate automatic analysis could prove very useful in assisting early diagnosis and improving the prognosis of patients with cardiovascular diseases. The Physionet/CinC 2016 Challenge provides the research community with the largest open database of annotated heart sounds; the research presented in this paper employed this dataset. An algorithm performance comparison with a universally standardized database contributes to promoting advances in the field of automated heart sound analysis. Our key objective is to evaluate the performance of different feature extraction, balancing, and feature selection techniques that can be relevant for effectively detecting heart murmurs. After thoroughly examining various classification schemes, we selected the RF method for its simplicity and high performance [69]. The main reason to use MP+LPC as features stems from the heart sound reconstruction model we have previously proposed [71]. This sparse time-frequency model accurately represents the non-stationary behavior of the FHS and murmurs. The *de facto standard* feature for sound recognition are the MFCCs; they have been extensively used in PCG classification. In this work, we analyzed the classification performance for MP+LPC+MFCC. Comparing datasets A and B, all the output scores for set A are always higher than B's. In our tests, the feature averaging approach outperforms cardiac cycle averaging. We suppose that this fact results from the higher diversity produced when taking the mean value of the features rather than the direct calculation of the features from a single averaged cycle. The atomic decompositions in MP were performed using MPTK, the Matching Pursuit Toolkit [78].

Although dataset A displayed a good score (with best sensitivity *SE*=88.5 %) using undersampling, dataset E presents better results (best sensitivity *SE*=91.19 %). That is, classification based only on MFCC features outperforms the MP+LPC approach. Nevertheless, a combination of features improves the results as shown by the scores of datasets C and D. The merger of MP+LPC+MFCC ranked second in sensitivity *SE*=93.17 %, and dataset D (MP+MFCC) obtained the highest score (*SE*=94.34 %), both in case of using undersampling. In terms of feature selection, performance improved in terms of the AUC and ROC curve scores, the best obtained when applying the IG method. Feature sets D and E got an AUC=0.971, while set C obtained an AUC of 0.969, which is closer than the other sets. Although CFS produces a higher reduction in the number of features than the IG method and better results, the AUC scores are lower than when applying IG.

CONCLUSIONS

In this work, we analyzed the effects of detecting cardiac murmurs when applying the SMOTE oversampling and random-under sampling methods for class balancing. Classification algorithms work best in cases where the number of samples is balanced. The reason is that they are designed to maximize accuracy and reduce error. SMOTE creates new abnormal PCG synthetic instances using the existing (real) ones. Although highly desirable, increasing the number of abnormal PCG sounds is a challenging task.

For this reason, we opted to use SMOTE. However, it is essential to be aware that this procedure might increase the likelihood of overfitting since it replicates the minority class events. The obtained SE scores were higher for the case of undersampling. However, the remaining scores *SP, ACC*, and *MCC* were improved when applying SMOTE.

The selection of sparse and spectral features helps classify PCG recordings under a high noise level without using FHS segmentation. For the group of LQR samples, the algorithm reached a *SE*=85% in feature set E applying under-sampling. On the other hand, the SMOTE oversampling effects produced lower scores in LQR when applying cycle averaging (data frame B). The best results were achieved when using all LQR+HQR samples of the dataset.

Finally, we compared the effectiveness of our method when using information gain (IG) and correlation feature selection (CFS). The results after applying dimensionality reduction were slightly higher using less features than when using the whole set of attributes. The highest sensitivity obtained was *SE*=96.23 % when using under-sampling and feature set C for both CFS and IG, and for feature set D when using CFS. After inspecting the discarded features, we noticed that the phase of the time-frequency atoms selected by MP is not a relevant feature. This parameter was never chosen when using CFS, and it obtained a value of *IG*=0. This result is not surprising since, in general, the phase is considered a random variable uniformly distributed in the $[-\pi,\pi]$ interval. In contrast, most of the MFCC were selected by both IG and CFS, in this case the first coefficients were ranked higher than the last. Regarding the other time-frequency atom parameters, the frequency, length, and position play a significant role in the classification. On the other hand, the amplitude has a low relevance. Only about a third of the coefficients were selected for the LPC without apparent order.

This research provides a complete assessment of feature selection methods for a classification algorithm to detect a pathological state from heart sounds. Different methods were evaluated, such as the balancing of samples, the comparison of MP+LPC vs. MFCC, and feature averaging vs. cycle time averaging as feature extraction methods. We also analyzed the effect of PCG signal quality and feature selection on classification performance. We selected the Random Forest technique algorithm to generate the classification model for PCG signals because the amount of data available is still small. Classical machine learning algorithms can often perform better than deep learning algorithms since they require a large amount of data for training. Nonetheless, classical machine learning algorithms are preferred when the interpretability of the model is essential since they are simpler and easier to understand [79].

The source code to reproduce the results of this paper can be downloaded free from; https://github.com/roilhi/ABMEPaperPCGClassif.git/.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

AUTHOR CONTRIBUTIONS

R.F.I.H. conceptualized the project, performed data curation, contributed to the research and methodology, participated in the use of specialized software, and oversaw the project, obtained resources, and participated in the writing of the original draft of the manuscript. M.A.A.A. performed formal analyses, validated analyses, visualized results and supervised the development of the project, participated in the writing review and the editing of the manuscript. E.C.G.C. Obtained funding and economic resources, visualized results and supervised the development of the project, participated in the writing review and the editing of the manuscript. All authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] D. I. Macht, “On the absorption of drugs and poisons through the [1] World Health Organization (WHO), “Cardiovascular diseases (CVDs),” WHO. Available: http://www.who.int/mediacentre/factsheets/fs317/en/ (accessed 2022).

[2] A. K. Abbas, R. Bassam, “Phonocardiography signal processing,” Springer Cham, 2009, pp. 194. [Online]. Available: https://doi.org/10.2200/S00187ED1V01Y200904BME031

[3] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, R. G. Mark, “Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016”, in: 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 609-612. [Online]. Available: https://ieeexplore.ieee.org/document/7868816

[4] W. Zhang, J. Han, S. Deng, “Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation,” Biomed. Signal Process. Control, vol. 53, art. no. 101560, Aug. 2019, doi: https://doi.org/10.1016/j.bspc.2019.101560

[5] Y. Soeta, Y. Bito, Y. “Detection of features of prosthetic cardiac valve sound by spectrogram analysis,” Appl. Acoust., vol. 89, pp. 28-33, Mar. 2015, doi: https://doi.org/10.1016/j.apacoust.2014.09.003

[6] L. Orozco-Reyes, M.-Á. A. Arévalo, E. García-Canseco, R. F. Ibarra-Hernández, “Clasificación de la senal de audio cardiaco mediante la transformada de Fourier de tiempo corto y aprendizaje profundo,” Res. Comput. Sci., vol. 151, no. 7, pp. 141-155, 2022. [Online]. Available: http://148.204.65.169/2022_151_7/Clasificacion%20de%20la%20senal%20de%20audio%20cardiaco%20mediante%20la%20transformada%20de%20Fourier%20de%20tiempo%20corto%20y.pdf

[7] M. S. Obaidat, “Phonocardiogram signal analysis: techniques and performance comparison,” J. Med. Eng. Technol., vol. 17 no. 6, pp. 221-227, 1993, doi: https://doi.org/10.3109/03091909309006329

[8] N. E., Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Ten, C. C. Tung, H. H. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” Proc. R. Soc. Lond., vol. 454, no. 1971, pp. 903-995, Mar. 1998, doi: https://doi.org/10.1098/rspa.1998.0193

[9] V. Nivitha Varghees, K. I. Ramachandran, “Effective heart sound segmentation and murmur classification using empirical Wavelet transform and instantaneous phase for electronic stethoscope,” IEEE Sens. J., vol. 17, no. 12, pp. 3861-3872, Jun. 2017, doi: https://doi.org/10.1109/JSEN.2017.2694970

[10] S. Yuenyong, A. Nishihara, W. Kongprawechnon, K. Tungpimolrut, “A framework for automatic heart sound analysis without segmentation,” BioMed. Eng. OnLine, no. 10, art. no. 13, Feb. 2011, doi: https://doi.org/10.1186/1475-925X-10-13

[11] S. M. Debbal, F. Bereksi-Reguig, “Detection of Differences of the Phonocardiogram Signals by Using the Continuous Wavelet Transform Method,” Int. J. Biomedical Soft Computing Hum. Sci., vol. 18 no.2, pp. 73-81, 2013, doi: https://doi.org/10.24466/ijbschs.18.2_73

[12] B. Ergen, Y. Tatar, H. O. Gulcur, “Time-frequency analysis of phonocardiogram signals using Wavelet transform: a comparative study,” Comput. Methods Biomech. Biomed. Eng., vol. 15, no. 4, pp. 371-381, Jan. 2011 doi: https://doi.org/10.1080/10255842.2010.538386

[13] S. A. Singh, T. G. Meitei, S. Majumder, “Short PCG classification based on deep learning,” in Deep Learning Techniques for Biomedical and Health Informatics, B. Agarwal, V. E. Balas, L. C. Jain, R. C. Poonia, Manisha, Eds. London, United Kingdom: Academic Press, 2020, ch. 6, pp. 141-164. [Online]. Doi: https://doi.org/10.1016/B978-0-12-819061-6.00006-9

[14] P. Qiao, Z. Yu, Z. Jingyi, C. Zhuo, “A method for diagnosing heart sounds in adolescents based on wavelet analysis and random forest,” in 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Fuzhou, China, 2020, pp. 69-74, doi: https://doi.org/10.1109/ICBAIE49996.2020.00021

[15] M. Abdollahpur, S. Ghiasi, M. J. Mollakazemi, A. Ghaffari, “Cycle selection and neuro-voting system for classifying heart sound recordings,” in 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 1-4. [Online] Available: https://ieeexplore.ieee.org/document/7868814

[16] P. Lubaib, K.V. Ahammed Muneer, “The Heart Defect Analysis Based on PCG Signals Using Pattern Recognition Techniques,” Proc. Technol., vol. 24, pp. 1024-1031, 2016, doi: https://doi.org/10.1016/j.protcy.2016.05.225

[17] M. Zabihi, A. B. Rad, S. Kiranyaz, M. Gabbouj, A. K. Katsaggelos, Heart sound anomaly and quality detection using ensemble of neural networks without segmentation, in 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 613-616. [Online]. Available: https://ieeexplore.ieee.org/document/7868817

[18] P. Wang, C. S. Lim, S. Chauhan, J. Y. A. Foo, V. Anantharaman, “Phonocardiographic signal analysis method using a modified hidden Markov model,” Ann. Biomed. Eng., vol. 35, pp. 367-374, 2007, doi: https://doi.org/10.1007/s10439-006-9232-3

[19] Y. Zheng, X. Guo, X. Ding, “A novel hybrid energy fraction and entropy-based approach for systolic heart murmurs identification,” Expert Syst. Appl., vol. 42, no. 5, pp. 2710-2721, Apr. 2015, doi: https://doi.org/10.1016/j.eswa.2014.10.051

[20] S. Chauhan, P. Wang, C. Sing Lim, V. Anantharaman, “A computer-aided MFCC-based HMM system for automatic auscultation,” Comput. Biol. Med., vol. 38, no. 2, pp. 221-233, Feb. 2008, doi: https://doi.org/10.1016/j.compbiomed.2007.10.006

[21] V. Maknickas, A. Maknickas, “Recognition of normal-abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients,” Physiol. Meas., vol. 38, no. 8, art. no. 1671, Jul. 2017, doi: https://doi.org/10.1088/1361-6579/aa7841

[22] G. Redlarski, D. Gradolewski, A. Palkowski, “A System for Heart Sounds Classification,” PLoS One, vol. 9, no. 11, art. no. e112673, Nov. 2014, doi: https://doi.org/10.1371/journal.pone.0112673

[23] B. M. Whitaker, P. B. Suresha, C. Liu, G. D. Clifford, D. V. Anderson, “Combining sparse coding and time-domain features for heart sound classification,” Physiol. Meas., vol. 38, no. 8, art. no. 1701, Jul. 2017, doi: https://doi.org/10.1088/1361-6579/aa7623

[24] I. J. Diaz Bobillo, “A tensor approach to heart sound classification,” in 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 629-632. [Online]. Available: https://ieeexplore.ieee.org/document/7868821

[25] W. Zhang, J. Han, S. Deng, “Heart sound classification based on scaled spectrogram and tensor decomposition,” Expert Syst. Appl., vol. 84, pp. 220-231, Oct. 2017, doi: https://doi.org/10.1016/j.eswa.2017.05.014

[26] W. Zhang, J. Han, S. Deng, "Heart sound classification based on scaled spectrogram and partial least squares regression," Biomed. Signal Process. Control, vol. 32, pp. 20-28, Feb. 2017, doi: https://doi.org/10.1016/j.bspc.2016.10.004

[27] P. Banerjee, A. Mondal, "An Irregularity Measurement Based Cardiac Status Recognition Using Support Vector Machine," J. Med. Eng., vol. 2015, art. no. 327534, Oct. 2015 doi: http://dx.doi.org/10.1155/2015/327534

[28] Y. Zheng, X. Guo, J. Qin, S. Xiao, "Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics," Comput. Methods Programs Biomed., vol. 122, no. 3, pp. 372-383, Dec. 2015, doi: https://doi.org/10.1016/j.cmpb.2015.09.001

[29] I. Maglogiannis, E. Loukis, E. Zafiropoulos, A. Stasis, "Support vectors machine-based identification of heart valve diseases using heart sounds," Comput. Methods Programs Biomed., vol. 95, no. 1, pp. 47-61, Jul. 2009, doi: https://doi.org/10.1016/j.cmpb.2009.01.003

[30] A. F. Quiceno-Manrique, J. I. Godino-Llorente, M. Blanco-Velasco, G. Castellanos-Dominguez, "Selection of Dynamic Features Based on Time–Frequency Representations for Heart Murmur Detection from Phonocardiographic Signals," Ann. Biomed. Eng., vol. 38, pp. 118-137, 2010, doi: https://doi.org/10.1007/s10439-009-9838-3

[31] L. D. Avendaño-Valencia, J. I. Godino-Llorente, M. Blanco-Velasco, G. Castellanos-Dominguez, "Feature Extraction From Parametric Time–Frequency Representations for Heart Murmur Detection," Ann. Biomed. Eng., vol. 38, pp. 2716-2732, Jun. 2010, doi: https://doi.org/10.1007/s10439-010-0077-4

[32] Singh, S. A., & Majumder, S. (2019). Classification of unsegmented heart sound recording using KNN classifier," J. Mech. Med. Biol., vol. 19, no. 04, art. no. 1950025, doi: https://doi.org/10.1142/S0219519419500258

[33] A. Sofwan, I. Santoso, H. Pradipta, M. Arfan A. A. Zahra M, "Normal and Murmur Heart Sound Classification Using Linear Predictive Coding and k-Nearest Neighbor Methods," in 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2019, pp. 1-5, doi: https://doi.org/10.1109/ICICoS48119.2019.8982393

[34] P. Narváez, S. Gutierrez, W. S. Percybrooks, "Automatic Segmentation and Classification of Heart Sounds Using Modified Empirical Wavelet Transform and Power Features," Appl. Sci., vol. 10 no.14, art. no. 4791, doi: https://doi.org/10.3390/app10144791

[35] M. N. Homsi, N. Medina, M. Hernandez, N. Quintero, G. Perpiñan, A. Quintana, P. Warrick, "Automatic heart sound recording classification using a nested set of ensemble algorithms," in 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 817-820. [Online]. Available: https://ieeexplore.ieee.org/document/7868868

[36] C. H. Antink, J. Becker, S. Leonhardt, M. Walter, "Nonnegative matrix factorization and random forest for classification of heart sound recordings in the spectral domain," in 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 809-812. [Online]. Available: https://ieeexplore.ieee.org/document/7868866

[37] N. E. Singh-Miller, N. Singh-Miller, "Using spectral acoustic features to identify abnormal heart sounds," in 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 557-560. [Online]. Available: https://ieeexplore.ieee.org/document/7868803

[38] C. Potes, S. Parvaneh, A. Rahman, B. Conroy, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 621-624. [Online]. Available: https://ieeexplore.ieee.org/document/7868819

[39] T.-E. Chen, S.-I Yang, L.-T. Ho, K.-H. Tsai, Y.-H. Chen, Y.-F. Chang, Y.-H. Lai, S.-S. Wang, Y. Tsao, C.-C. Wu, "S1 and S2 Heart Sound Recognition Using Deep Neural Networks," IEEE Trans. Biomed. Eng., vol. 64, no. 2, pp. 372-380, Feb. 2017, doi: https://doi.org/10.1109/TBME.2016.2559800

[40] F. Beritelli, G. Capizzi, G. Lo Sciuto, C. Napoli, F. Scaglione, "Automatic heart activity diagnosis based on Gram polynomials and probabilistic neural networks," Biomed. Eng. Lett., no. 8, pp. 77-85, Feb. 2018, doi: https://doi.org/10.1007/s13534-017-0046-z

[41] F. Li, H. Tang, S. Shang, K. Mathiak, F. Cong, "Classification of Heart Sounds Using Convolutional Neural Network," Appl. Sci., vol. 10, no. 11, art. no. 3956, Jun. 2020, doi: https://doi.org/10.3390/app10113956

[42] F. Demir, A. Şengür, V. Bajaj, K. Polat, "Towards the classification of heart sounds based on convolutional deep neural network," Health Inf. Sci. Syst., vol. 7, art. no. 16, Aug. 2019, doi: https://doi.org/10.1007/s13755-019-0078-0

[43] Y. Chen, S. Wei, Y. Zhang, "Classification of heart sounds based on the combination of the modified frequency wavelet transform and convolutional neural network," Med. Biol. Eng. Comput., vol. 58, pp. 2039-2047, Jul. 2020, doi: https://doi.org/10.1007/s11517-020-02218-5

[44] F. Renna, J. Oliveira, M. T. Coimbra, "Deep Convolutional Neural Networks for Heart Sound Segmentation," IEEE J. Biomed. Health Inform., vol. 23, no. 6, pp. 2435-2445, Nov. 2019, doi: https://doi.org/10.1109/JBHI.2019.2894222

[45] F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, X. Liu, M. Hui, "Feature extraction and classification of heart sound using 1D convolutional neural networks," EURASIP J. Adv. Signal Process., vol. 2019, art. no. 59, Dec. 2019, doi: https://doi.org/10.1186/s13634-019-0651-3

[46] B. Xiao, Y. Xu, X. Bi, J. Zhang, X. Ma, "Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption," Neurocomputing, vol. 392, pp. 153-159, Jun. 2020, doi: https://doi.org/10.1016/j.neucom.2018.09.101

[47] S. Latif, M. Usman, R. Rana, J. Qadir, "Phonocardiographic Sensing Using Deep Learning for Abnormal Heartbeat Detection," IEEE Sens. J., vol. 18, no. 22, pp. 9393-9400, Nov. 2018, doi: https://doi.org/10.1109/JSEN.2018.2870759

[48] F. A. Khan, A. Abid, M. S. Khan, "Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features," Physiol. Meas., vol. 41, no. 5, art. no. 055006, Jun. 2020, doi: https://doi.org/10.1088/1361-6579/ab8770

[49] A. Raza, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, B.-W. On, "Heartbeat Sound Signal Classification Using Deep Learning," Sensors, vol. 19, no. 21, art. no. 4819, Nov. 2019, doi: https://doi.org/10.3390/s19214819

[50] E. Messner, M. Zöhrer, F. Pernkopf, "Heart Sound Segmentation—An Event Detection Approach Using Deep Recurrent Neural Networks," IEEE Trans. Biomed. Eng., vol. 65, no. 9, pp. 1964-1974, Sep. 2018, doi: https://doi.org/10.1109/TBME.2018.2843258

[51] A. K. Dwivedi, S. A. Imtiaz, E. Rodriguez-Villegas, "Algorithms for Automatic Analysis and Classification of Heart Sounds-A Systematic Review," IEEE Access, vol. 7, pp. 8316-8345, 2019, doi: https://doi.org/10.1109/ACCESS.2018.2889437

[52] S. Ismail, I. Siddiqi, U. Akram, "Localization and classification of heart beats in phonocardiography signals —a comprehensive review," EURASIP J. Adv. Signal Process., vol. 2018, art. no. 26, 2018, doi: https://doi.org/10.1186/s13634-018-0545-9

[53] W. Chen, Q. Sun, X. Chen, G. Xie, H. Wu, C. Xu, "Deep Learning Methods for Heart Sounds Classification: A Systematic Review," Entropy, vol. 23, no. 6, art. no. 667, 2021, doi: https://doi.org/10.3390/e23060667

[54] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. W. Johnson, et al., "An open access database for the evaluation of heart sound algorithms," Physiol. Meas., vol. 37, no. 12, art. no. 2181, Nov. 2016 doi: https://doi.org/10.1088/0967-3334/37/12/2181

[55] S. G. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries," IEEE Trans. Signal Process., vol. 41, no. 12, pp. 3397-3415, Dec. 1993, doi: https://doi.org/10.1109/78.258082

[56] R. F. Ibarra, M. A. Alonso, S. Villarreal, C. I. Nieblas, "A parametric model for heart sounds," in 2015 49th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2015, pp. 765-769, doi: https://doi.org/10.1109/ACSSC.2015.7421237

[57] C.I. Nieblas, M. A. Alonso, R. Conte, S. Villarreal, "High performance heart sound segmentation algorithm based on Matching Pursuit," 2013 IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), Napa, CA, USA, 2013, pp. 96-100, doi: https://doi.org/10.1109/DSP-SPE.2013.6642572

[58] W. Wang, Z. Guo, J. Yang, Y. Zhang, L.-G. Durand, M. Loew, "Analysis of the first heart sound using the matching pursuit method," Med. Biol. Eng. Comput., vol. 39, pp. 644-648, Nov. 2001, doi: https://doi.org/10.1007/BF02345436

[59] Xuan Zhang, L. Durand, L. Senhadji, H. C. Lee, J.-L. Coatrieux, "Analysis-synthesis of the phonocardiogram based on the matching pursuit method," IEEE Trans. Biomed. Eng., vol. 45, no. 8, pp. 962-971, Aug. 1998, doi: https://doi.org/10.1109/10.704865

[60] S. Jabbari, H. Ghassemian, "Modeling of heart systolic murmurs based on multivariate matching pursuit for diagnosis of valvular disorders," Comput. Biol. Med., vol. 41, no. 9, pp. 802-811, Sep. 2011, doi: https://doi.org/10.1016/j.compbiomed.2011.06.016

[61] W. Wang, J. Pan, H. Lian, "Decomposition and analysis of the second heart sound based on the matching pursuit method," in Proceedings 7th International Conference on Signal Processing, 2004, Beijing, China, 2004, pp. 2229-2232, vol. 3, doi: https://doi.org/10.1109/ICOSP.2004.1442222

[62] P. P. Vaidyanathan, "The Theory of Linear Prediction," in Synthesis Lectures on Signal Processing. Switzerland: Springer Nature, 2022, pp. 1-184, doi: https://doi.org/10.2200/S00086ED1V01Y200712SPR003

[63] T. Virtanen, M. D. Plumbley, D. Ellis, Computational analysis of sound scenes and events. Springer Cham, 2018, pp. 1-422, doi: https://doi.org/10.1007/978-3-319-63450-0

[64] L. Breiman, "Random Forests," Mach. Learn., vol. 45, pp. 5-32, 2001, doi: https://doi.org/10.1023/A:1010933404324

[65] R. Gonzalez-Landaeta, B. Ramirez, J. Mejia, "Estimation of systolic blood pressure by Random Forest using heart sounds and a ballistocardiogram," Sci. Rep., vol. 12, art. no. 17196, doi: https://doi.org/10.1038/s41598-022-22205-0

[66] C. C. Balili, M. C. C. Sobrepena, P. C. Naval, "Classification of heart sounds using discrete and continuous wavelet transform and random forests," in 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 2015, pp. 655-659, doi: https://doi.org/10.1109/ACPR.2015.7486584

[67] W. Xu, K. Yu, J. Ye, H. Li, J. Chen, F. Yin, et al., "Automatic pediatric congenital heart disease classification based on heart sound signal," Artif. Intell. Med., vol. 126, art. no. 102257, Apr. 2022, doi: https://doi.org/10.1016/j.artmed.2022.102257

[68] J. Oliveira, D. Nogueira, C. Ferreira, A. M. Jorge, M. Coimbra, "The robustness of Random Forest and Support Vector Machine Algorithms to a Faulty Heart Sound Segmentation," in 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 1989-1992, doi: https://doi.org/10.1109/EMBC48229.2022.9871111

[69] M. Y. Esmail, D. H. Ahmed, M. Eltayeb, "Classification System for Heart Sounds Based on Random Forests," J. Clin. Eng., vol. 44, no. 2, pp. 76-80, 2019, doi: https://doi.org/10.1097/JCE.0000000000000335

[70] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "Smote: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, Jun. 2002, doi: https://doi.org/10.1613/jair.953

[71] R. F. Ibarra-Hernández, M. A. Alonso-Arévalo, A. Cruz-Gutiérrez, A. L. Licona-Chávez, S. Villarreal-Reyes, "Design and evaluation of a parametric model for cardiac sounds," Comput. Biol. Med., vol. 89, pp. 170-180, Oct. 2017, doi: https://doi.org/10.1016/j.compbiomed.2017.08.007

[72] L. L. Vercio, M. Del Fresno, I. Larrabide, "Detection of morphological structures for vessel wall segmentation in ivus using random forests," in 12th International Symposium on Medical Information Processing and Analysis, Tandil, Argentina, 2017, art. no. 1016012, doi: https://doi.org/10.1117/12.2255748

[73] A. Mellor, S. Boukir, A. Haywood, S. Jones, "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin," ISPRS J. Photogramm. Remote Sens., vol. 105, pp. 155-168, Jul. 2015, doi: https://doi.org/10.1016/j.isprsjprs.2015.03.014

[74] R. F. Ibarra-Hernández, N. Bertin, M. A. Alonso-Arévalo, H. A. Guillén-Ramírez, "A benchmark of heart sound classification systems based on sparse decompositions," in 14th International Symposium on Medical Information Processing and Analysis, Maztlán, México, 2018, art. no. 1097505, doi: https://doi.org/10.1117/12.2506758

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011. [Online] Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/

[76] M. A. Hall, "Correlation-based feature selection for machine learning," PhD dissertation, University of Waikato, Hamilton, New Zeland, 1999. [Online]. Available: https://hdl.handle.net/10289/15043

[77] B. Azhagusundari, A. S. Thanamani, "Feature selection based on information gain," IJITEE, vol. 2, no. 2, pp. 18-21, Jan. 2013. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e17df473c25cccd8435839c9b6150ee61bec146a

[78] S. Krstulovic, R. Gribonval, "Mptk: Matching Pursuit Made Tractable," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, Francia, 2006, pp. III-III, doi: https://doi.org/10.1109/ICASSP.2006.1660699

[79] G. D. Clifford, C. Liu, B. Moody, J. Millet, S. Schmidt, Q. Li, et al., "Recent advances in heart sound analysis," Physiol. Meas., vol. 38, no. 8, art. no. E10, Aug. 2017, doi: https://doi.org/10.1088/1361-6579/aa7ec8

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 23-37

Revista Mexicana de Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.44.4.2

E-LOCATION ID: **1353**

Using Machine Learning Algorithms on Electroencephalographic Signals to Assess Engineering Students' Focus While Solving Math Exercises

Uso de Algoritmos de Aprendizaje Automático en Señales Electroencefalográficas para Evaluar la Concentración de Estudiantes de Ingeniería al Resolver Ejercicios Matemáticos

José Jaime Esqueda Elizondo [1], *Laura Jiménez Beristáin* [1], *Alejandra Serrano Trujillo* [1], *Margarita Zavala Arce* [2], *Diego Armando Trujillo Toledo* [1], *Óscar Roberto López Bonilla* [3], *Gilberto Manuel Galindo Aldana* [4], *J. Reyes Juárez Ramírez* [1], *Andrea López Rivas* [1], *Annette Sofia Martinez Verdin* [1], *María Marlene Muñoz López* [5], *Itzel Arely Romano Pérez* [5]

[1] Facultad de Ciencias Químicas e Ingeniería, Universidad Autónoma de Baja California - México

[2] Instituto Tecnológico de Ciudad Madero, Tecnológico Nacional de México - México

[3] Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California - México

[4] Facultad de Ingeniería y Negocios, Universidad Autónoma de Baja California - México

[5] Facultad de Ciencias de la Electrónica, Benemérita Universidad Autónoma de Puebla - México

ABSTRACT

In this paper, we present an attention classification method using Machine-Learning Algorithms. The EEG signals were recorded from ten engineering students with an EPOC+BCI using the electrodes F3, F4, P7, and P8 while solving some mathematical operations. The recording time for these activities is around 20 minutes. Next, a similar time EEG register is obtained while doing non-academic activities, such as chattering with the staff, checking cell phones, or playing a video game. With these EEG registers, we obtained a set of features to train and evaluate attention using Machine Learning algorithms. This research shows how engineering students interact with math topics in solving mental operations and complex reasoning by increasing brain domain and knowledge for mathematical reasoning-related processes, such as sustained and shifting attention and logical constructions for object interaction during operations resolution. The Random Forest algorithm (RF) obtained the highest accuracy with 0.7392, an F1 Score of 0.7430, and the highest Specificity/Accuracy with 0.7261.

KEYWORDS: attention measurement, brain-computer interface, classification, electroencephalographic signals, machine learning, math

RESUMEN

Se presenta un método de clasificación de la atención utilizando algoritmos de aprendizaje automático. Con las señales EEG de diez estudiantes de ingeniería adquiridas utilizando los electrodos F3, F4, P7 y P8 de una BCI EPOC+ mientras resuelven productos escalares, multiplicaciones algebraicas simples, simplificaciones e integrales por aproximadamente 20 minutos. Posteriormente, se obtiene un registro EEG de tiempo similar mientras se realizan actividades no académicas, como charlar con el personal, consultar el móvil o jugar a un videojuego. Se obtienen algunas características/parámetros, se entrenan y evalúan varios algoritmos de aprendizaje automático para la clasificación de la atención. Los resultados de esta investigación pueden mejorar la forma en que los estudiantes de ingeniería interactúan con los temas matemáticos en la resolución de operaciones mentales y razonamientos complejos, aumentando el dominio y el conocimiento cerebral para los procesos relacionados con el razonamiento matemático, como la atención sostenida y cambiante y las construcciones lógicas para la interacción con objetos durante la resolución de operaciones. El clasificador Random Forest obtuvo la mayor precisión con 0.7392, una puntuación F1 de 0.7430 y la mayor especificidad/precisión con 0.7261.

PALABRAS CLAVE: aprendizaje automático, clasificación, interfaz cerebro-computadora, matemáticas, medición de la atención, señales electroencefalográficas

Corresponding author

TO: **José Jaime Esqueda Elizondo**

INSTITUTION: **Universidad Autónoma de Baja California**

ADDRESS: **Facultad de Ciencias Químicas e Ingeniería,**
Campus Tijuana UABC
Calzada Universidad #14418, Parque Industrial
Internacional Tijuana C.P. 22390

CORREO ELECTRÓNICO:
jjesqueda@uabc.edu.mx

Received:

27 April 2023

Accepted:

2 August 2023

INTRODUCTION

The cognitive process known as attention is made possible by selecting, zooming, and maintaining the processing of certain pieces of information [1]. Whatever is now being processed by sensorial or information systems or reaction possibilities generated by ongoing cognitive activity might serve as objects of attention. It helps us to focus on the information that is relevant to us and discriminate against the rest. In addition, it is essential to have it before beginning the learning process, as it serves as its foundation. There have been a variety of methods of measurement, including recording response times or clicks made on specific software, measuring eye contact time from movies, doing Magnetic resonance imaging (MRI) or Functional magnetic resonance imaging (fMRI) research, and so on.

Attention is fundamental for university students, so measuring it is essential. Some research works are related, like [2][3], where primary school students are observed doing math and language tests using electroencephalogram (EEG). Figure 1 shows the framework followed for the preprocessing and classification stages in [2][3]. They use 64-channel EEG data and process only the signal from frontal (6, 12, 60) and parietal (28, 34, 42) electrodes. They use the average Welch's Power Spectrum Density (PSD) for the 21-45 Hz band and the K-Nearest Neighbor as a feature and classifier for these EEG channels. The preprocessing for the EEG signals was performed with the Python MNE library, and the PSD was obtained with the SciPy library. Five-and-a-half percent did poorly on the arithmetic test, while seventy-four percent did poorly on the language test. Based on the results of the tests, the average sensitivity and specificity for each fold were satisfactory. Similar sensitivity was found when EEG data was combined with socio-demographic and home environment characteristics. The results obtained were: Math = 58.7 %, Language = 66.3 %, but more specificity was found Math = 43.4 % to 50.6 %, Language = 32 % to 60 %.

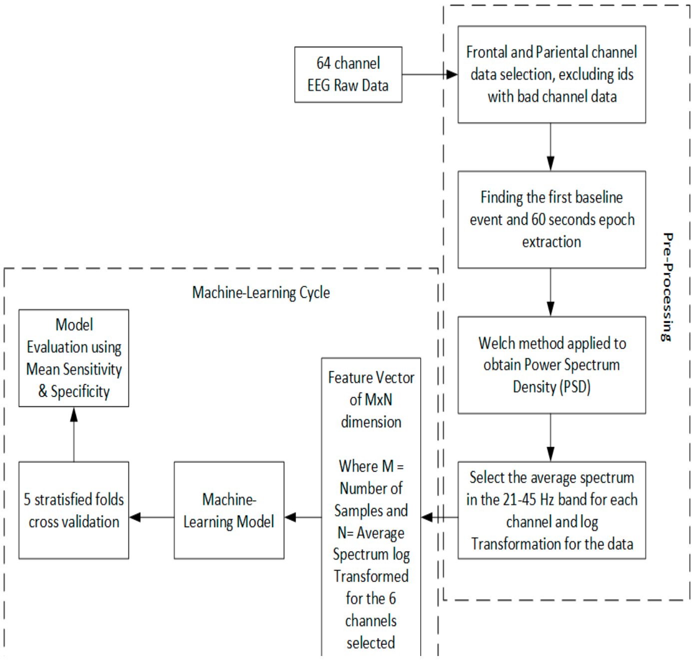


FIGURE 1. **Framework for the preprocessing and classification stages followed by researchers in [2][3]. Taken from [2][3].**

In [4], the authors examine the effects of exercise on the math test scores and anxiety levels of 68 sixth-graders from two primary schools in New South Wales, Australia. In this study, trait anxiety (low vs. high) and condition (activity break vs. control) are the between-subjects factors in a 22-between-subjects design. The dependent variables were math exam scores, mental effort expended, task perceived difficulty, and a three-time state anxiety measure. IBM SPSS Statistics 25 was used to do an ANOVA and ANCOVA on the collected data. As a result, authors found that physical activity break before a test examination does not deteriorate test anxiety and math test performance.

In [5], the authors consider the eye movements of 30 participants' suggesting that they unconsciously focus on the numbers (operands, solution) they are now processing. Faster performance before fixing on the relevant numbers and making fewer trips back to the first operand in the computation resulted in shorter latencies. These telltale signs of superior task performance were most obvious for addition and visually ordered numbers and for subtraction and visually ordered numbers in the opposite direction. In this case, the eye movements were done with EyeLink 1000, a video-based eye tracker, and the acquired information was processed with SR Research Data Viewer software. 2 + 5, 5 + 2, 3 + 5, 5 + 3, 2 + 6, 6 + 2, 3 +6, 6 + 3, 3 +8, 8 + 3, 4 + 8, 8 + 4, 3 + 9, 9 + 3, and 9 + 4 were the addition problems utilized in this study. Each addition problem was transformed into a corresponding subtraction problem with the answer to the addition problem as the first operand (7 + 5 = 7 became 7 - 5 = 2), and so on. Four white numbers, set against a black background, were displayed horizontally in ascending or descending numerical sequence (see Figure 2). The numerals were 18 points tall, had a visual angle of 0.5 degrees, and were set in Times New Roman font type. On an invisible 44 grid (shown as a dotted line in Figure 2), each number may take one of sixteen possible placements, with a single digit occupying each column and row. Two numerical images corresponded to the orally provided operands of the arithmetic problem; another was a distractor operand, and the final image was the answer (correct or incorrect). Eye tracking data showed that people naturally gaze at the numbers they are currently processing (operands, solution). Performance improvements were seen in shorter latency before fixating the relevant integers and fewer returns to the first operand during solution computation. In particular, these hallmarks of high-quality task performance were more prominent for addition and visually-arranged numbers in ascending order and for subtraction and visually-arranged numbers in descending order (compared to the opposite pairings). Our findings demonstrate that the "visual number world"-paradigm provides real-time insight into mental arithmetic, is sensitive to visual layout modifications that are not reflected in response time measures and can capture variability in arithmetic performance.

Authors found that comparing the proportions of fixating the first operand between small-and large-operand-first problems in the computational phase (paired test) revealed a higher proportion for large-operand-first problems, $t(28) = -2.256$, $p = .033$. The fixation proportion of the first operand was 7.2% increased for these problems (SEM = 1.6). Thus, eye movements do not reflect a possible mental rearrangement of operands from a small-large into a large-small order but, rather, the activation of the larger number to which the smaller number is added.

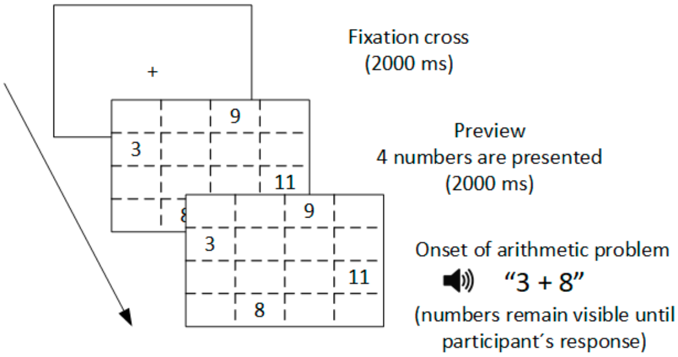


FIGURE 2. **An example of a trial sequence is used in the research, taken from [5].**

Research [6] presents a proposed methodology that evaluates "Attention" and "Meditation" levels in children in virtual classes with a Mindwave BCI, a brainwave-reading EEG headset. They obtain a level of "Attention" and "Meditation" from 0 to 100 % using the software provided with the BCI. Their methodology was evaluated in two situations: with a 6-year-old child and a 15-year-old during virtual classes. This one-channel BCI uses the FP1 electrode, and the software measures the Attention and Meditation levels during a session and displays their average, as is shown in Figure 3. At the beginning of the class, the 6-year-old child's Attention average was 10 % and a Meditation average of 40 %. When the child had to solve an exercise and turn it in, "Attention" rose to 72 %, and "Meditation" a 50 %. For the 15-year-old student, in the beginning, the "Attention" average was 0 %, and the "Meditation" average of almost 100 %. His "Attention" average in Math class was 60 %, and the "Meditation" average was nearly 40 %. While he was analyzing the math exercises, the "Attention" average was 90 %, and the "Meditation" average was 15 %. They conclude that the two students' attention levels are different when performing similar tasks, with the 6-year-old student showing a maximum attention level of 72 % reaching 100 % due to the student at that age being very distracted, and the 15-year-old student showing a maximum attention level of 100 % when performing class activities. They concluded that taking a break to exercise before an exam did not lower performance on arithmetic tests or increase test-related anxiety.

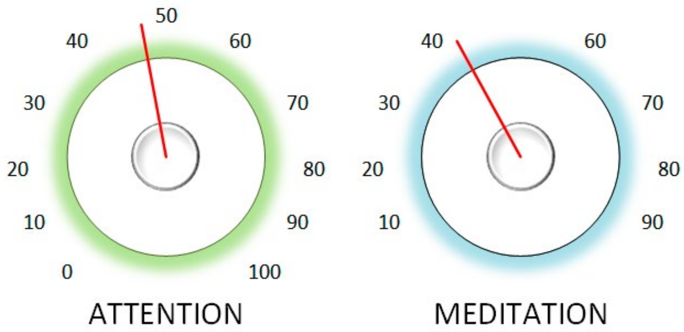


FIGURE 3. **Attention and Meditation averages displayed by the Mindwave BCI software. The illustrative Figure is taken from: https://apkpure.com/brainwave-visualizer/com.neurosky.unitythinkgear.**

Also, Machine-Learning is used to diagnose and study different neurodevelopmental and neurodegenerative disorders. In [7], they use the retina's electrical response (electroretinogram, ERG) to light for autism spectrum disorder (ASD) detection. They collected and analyzed ERG signals from 47 control and 96 ASD individuals. As features, they use four well-known time-domain indices, specifically the amplitude of a "Va" wave, the timing of its "Ta" peak, the amplitude of a "Vb" wave, and the timing of its "Tb" peak. These time domain and statistical features are used for this research. Then they obtain the Power Spectrum Density (PSD), and the Discrete Wavelet Transform of the ERG signals is used with the Least Absolute Shrinkage and Selection Operator (LASSO) regression technique for feature extraction. The machine learning models used included a Decision Tree (DT), a Support Vector Machine (SVM), a Gradient Boosting (GB), and Random Forest (RF). They detected ASD using Random Forest, which obtained the best classification accuracy of 86 % and 98 % sensitivity with time domain and spectral features.

In reference [8], authors suggest using deep convolutional neural networks (CNNs), a sparse coding-based feature mapping methodology, and the Douglas-Peucker (DP) algorithm to spot ASD. This research used the King Abdulaziz University Hospital, Jeddah, Saudi Arabia dataset. This dataset contains 20 children with ASD (ages 6-20) and a healthy group with nine children without neurological conditions. Figure 4 shows the block diagram of the proposed method. They employ a g.tec EEG cap with Ag/AgCl electrodes, G.tech USB amplifiers, and BCI2000 software to acquire signals without artifacts. Using the worldwide 10-20 system with AFz as GND and the right ear lobe, they work with 16 channels (FP1, FP2, F7, F3, Fz, F4, F8, T3, C4, Cz, C3, T5, Pz, O1, Oz, and O2). A popular method for simplifying lines, the Douglas-Peucker (DP) algorithm can reduce curve complexity and storage requirements by eliminating unimportant nodes and isolating key nodes The goal of the DP method is to create a new data series with fewer and more significant points while keeping the original data series within an acceptable range of variance. Using the DP algorithm, they reduce from 47088 original samples to only 2462 with a compression parameter of 20. Next, EEG rhythms are extracted with a Wavelet Transform using Daubechies 4th order decomposition to obtain Gamma, Beta, Alpha, Theta, and Delta band power and coded with sparse representation. All rhythms are neatly lined up, and the sparse presentation is encoded with the help of a histogram. Histograms of the sparsely coded rhythms for each EEG channel are added to form a matrix representation of the input EEG information. Before applying the dB power scale (20log()), the matrix data must be normalized into the [0,1] interval. Scaled color representations of the matrix are generated using the *ima-*

gesc command in Matlab and then saved as images. Autoencoders (AE) based on extreme learning machines (ELM) are used during data augmentation. Next, pre-trained deep CNN models are used to classify the EEG signals from people with ASD and those without the disorder. When applied to the automatic diagnosis of ASD, the proposed method achieved a perfect 100 % sensitivity, 96.4 % specificity, and a perfect F1-score of 99.19 %.

Deep-Learning models are also applied to EEG signals to detect schizophrenia, as is presented in reference [9]. In this instance, they utilized information gathered from Warsaw, Poland's Institute of Psychiatry and Neurology. Models from the Machine Learning (M-L) family employed include the Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Extremely Randomized Trees, Random Forest, and Bagging. Long Short-Term Memories (LSTMs), One-Dimensional Convolutional Networks (1D-CNNs), and One-Dimensional Convolutional Networks-LSTMs are just a few examples of the Deep-Learning (DL) models they employ. The initial step in processing is slicing the EEG signals into 25-second chunks. Following this, both DL and FCF are applied concurrently in the feature extraction phase. Using a CNN-LSTM network, the DL model employs functional connectivity methods, including synchronization likelihood (SL), Fuzzy SL (FSL), and simplified interval type-2 FSL (SIT2FLS). Here, they use a concatenate layer to merge the DL features with those of each functional connectivity type before passing the resulting dataset over to a sigmoid activation layer for classification. K-Fold with K = 5 was utilized in the categorization stage to evaluate the outcomes. As a result, they found that the best performance was for the CNN-LSTM model, with an accuracy percentage of 99.25 %.

Also, reference [10] presents a method for identifying overconfidence patterns by analyzing EEG power spectrum bands. Students solve mathematical tasks and receive feedback about their mistakes in the solved exercises. Twenty healthy engineering students (13 males and seven females, with a mean age of 18.73 ± 0.65 years) were monitored by EEG as they performed mathematical calculations. Before, during, and after problem-solving, the subjects' Delta and Theta band activity was evaluated. The graded work included ten multiple-choice exercises on topics including algebraic fraction simplification, factoring, and the usage of radicals, typically covered in the intermediate years of high school. Each exercise has one correct answer and four incorrect ones, as shown in Figure 5. If they answer correctly, the next exercise is presented. Still, if an incorrect answer is selected, feedback is given to the student, including the probable cause of the error and the correct procedure. Next, a second similar exercise is presented, again with one correct answer and three wrong answers.

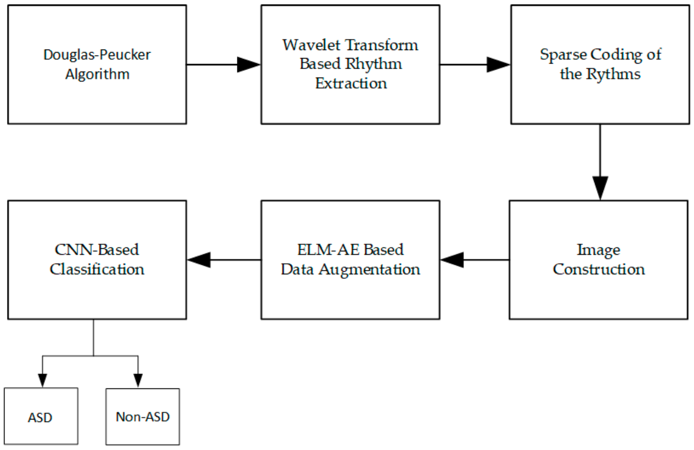


FIGURE 4. Block diagram of the proposed method. Taken from [8].

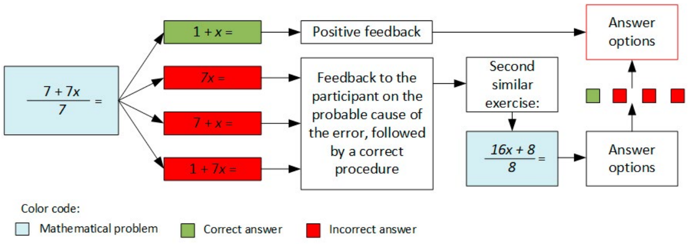


FIGURE 5. The mathematical task presented. Taken from [10].

The authors used an electroencephalograph (EEG) with a 200 Hz sampling rate and 19 electrodes positioned according to the 10-20 International System. This means electrodes Fp1/2, F3/4, F7/8, C3/4, P3/4, T3/4, T5/6, O1/2, Fz, Cz, and Pz are used for EEG data acquisition.

Authors analyzed signals from four discrete times within a single second: (a) just before the math problem was shown (V1), (b) immediately after it was shown (V2), (c) just before the answer was selected (V3), and (d) just as feedback appears in response to the answer chosen (V4). These moments of interest are depicted in Figure 6.

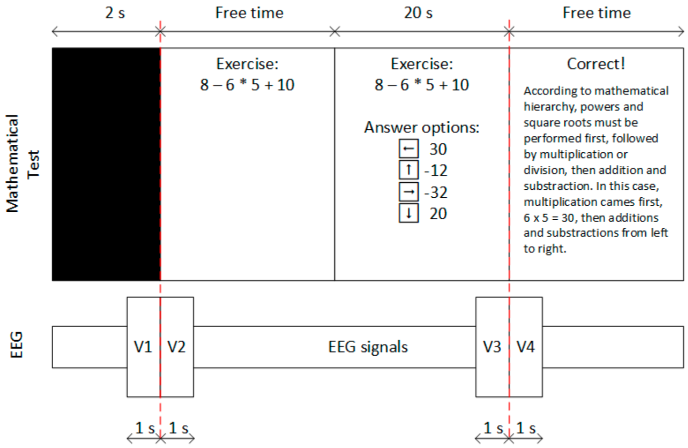
The relative band energy is obtained by estimating the relative energy of each band, and they estimate the average PSD of each window using equation (1).

$$EB_r = \frac{EB}{E\delta \; + \; E\theta \; + E\alpha \; + \; E\beta} \tag{1}$$

Where EBr represents the relative energy of the interesting band, EB is the absolute energy of the interesting band, and Eδ, Eθ, Eα, and Eβ the absolute energies of bands Delta, Theta, Alpha, and Beta, respectively.

Those who got the answers right saw an uptick in Delta band activity when the correct solutions were shown, whereas those who got them wrong saw a decrease. Subjects who got an exercise wrong were given feedback and given a second chance to get it right. Subjects' Theta energy levels increased when they answered correctly and decreased when they answered incorrectly. The authors of this study did not employ any Machine-Learning techniques and found that overconfidence may be quantified by measuring the fluctuations in subject energy during mathematical task errors.

In [11], our previous work uses EEG signals and Machine-Learning Algorithms for attention measurement in an ASD user. For EEG signal acquisition, we used an Epoc+ headset manufactured by Emotiv to acquire EEG signals from a 13 years old boy with ASD diagnosed while he performed learning activities. As features, we use the PSD with two seconds windows. Theta, which goes from 4 to 8 Hz, Alpha, from 8 to 12 Hz; and Beta, from 12 to 30 Hz, are the frequency bands that comprise the Power Spectrum Density. Theta Relative Power, Alpha Relative Power, and Beta Relative Power are all attained and used with the capabilities of this band. Theta-Beta, Theta-Alpha, and Theta/(Alpha + Beta) Ratios are all obtained using the relative powers presented in [12]. These features are obtained for the F3, F4, P7, and P8 channels of the Epoc+ headset. Naive Bayes (NB), Decision trees (DT), k-nearest neighbors (KNN), Support Vector Machine (SVM)-RBF, Stochastic Gradient Descent (SGD), Random Forest (RF), Extra trees (ET), and Multi-Layer Perceptron Neural Network (MLP-NN) from the Scikit-learn library (https://scikit-learn.org/stable/) were among the Machine-Learning models analyzed. An AUC of 0.9299 indicates that the best model is the multi-layer perceptron neural network (MLP-NN). Table 1, compares the state of the art presented in the Introduction section.



According to mathematical hierarchy, powers and square roots must be performed first, followed by multiplication or division, then addition and substraction. In this case, multiplication cames first, 6 x 5 = 30, then additions and substractions from left to right.

FIGURE 6. Example of a mathematical exercise and the structure of the one-second windows for analysis. Taken from [10].

TABLE 1. **State-of-the-art comparative.**

Reference	Year	Dataset	Features	IA model
[2], [3]	2021	Own; 105 children (52 girls)	Average Welch's Power Spectrum Density (PSD) for the 21-45 Hz band	KNN
[4]	2020	Own; 68 sixth-grade children of 11–12	ANOVA and ANCOVA	None
[5]	2018	Own; 30 participants	ANOVA	N/A
[6]	2022	Own; a 6-year-old child and a 15-year-old.	Attention and Meditation averages obtained and displayed by the Mindwave BCI software.	N/A
[7]	2022	Own; an electroretinogram from 47 control and 96 ASD individuals.	PSD and Wavelet Transform of the "Va" wave Amplitude, the timing of its "Ta" peak, the amplitude of a "Vb" wave, and the timing of its "Tb" peak.	Gradient Boosting (GB), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM).
[8]	2022	Dataset from King Abdulaziz University Hospital, Jeddah, Saudi Arabia, with 20 ASD children (ages 6–20) and a healthy group with nine children.	Douglas-Peucker (DP) algorithm, Wavelet Transform using Daubechies 4th order decomposition to obtain Gamma, Beta, Alpha, Theta, and Delta band powers.	Deep CNN
[9]	2022	Data set from the Institute of Psychiatry and Neurology, Warsaw, Poland	25s time frames, normalized by z-score or norm L2	M-L models: Support Vector Machine, K-Nearest Neighbors, Decision Tree, Naïve Bayes, Random Forest, Extremely Randomized Trees, and Bagging. Deep-Learning (DL) models: Long Short-Term Memories (LSTMs), One-Dimensional Convolutional Networks (1D-CNNs), and 1D-CNN-LSTMs.
[10]	2022	Own; EEG signals from 20 engineering students (13 men and seven women with an average age of 18.73 ± 0.65 years)	Average power spectrum of two-second windows to obtain the relative energy of each band with the equation	Statistical analysis.
[11]	2022	Own; one 13 years old boy with ASD diagnosis.	Theta Relative Power, Alpha Relative Power, Beta Relative Power, Theta-Beta Ratio, Theta-Alpha Ratio, and the Theta/(Alpha + Beta) Ratio	Naive Bayes (NB), Stochastic Gradient Descent (SGD), Decision trees (DT), Support Vector Machine (SVM)–RBF, k-nearest neighbors (KNN), Multi-Layer Perceptron Neural Network (MLP-NN), Random Forest (RF), and Extra trees (ET)

Listed below are the sections of this document. The second section, Materials and Methods, presents the proposed approach. Here we present the Activity Sheet, the BCI, the Data Acquisition Process, Signal Processing, Feature Extraction, and the Dataset obtained. Results and Discussion, the third section, summarizes this paper's findings and where the debate takes place. Our findings are summarized in the Conclusions section.

MATERIALS AND METHODS

This study explains how to measure and process the brain's electrical activity and assess attention levels when engaging in cognitive activities and interacting with various software systems or applications. EEG signals from an Epoc+ Brain-Computer Interface (BCI) can be used in this research for detecting when a user has high attention levels while solving mathematical problems, as in a class. The user's "Attention" and "No Attention" states are classified using ML techniques in this paper. This study uses EEG readings and machine learning algorithms to classify the attention of a ten-engineering student sample with an average age of 22.4 years and a standard deviation of 2.2 years, six males and four females. People with Asperger's Syndrome can be diagnosed based on their ability to focus on tasks and interact with computer programs, according to a study published in the Journal of Autism and Developmental Disorders.

This project was registered as POSG/020-1-04 with the University of Baja California's Ethics Committee and Research for Undergraduates and Graduates on October 8, 2020. Data was collected using an Epoc+ BCI connected to the Emotiv Pro platform while every engineering student in the sample solved some math exercises for about 15 to 18 minutes. Then, a similar time EEG register of no attention activities like checking cell phone, playing a videogame, and talking to the staff, among other activities. Matlab 2019a and Emotiv Pro, Student Edition were used for signal processing of the recorded EEG data. Figure 7 shows the Emotiv Epoc+ headset (right) and the electrode placement (left).

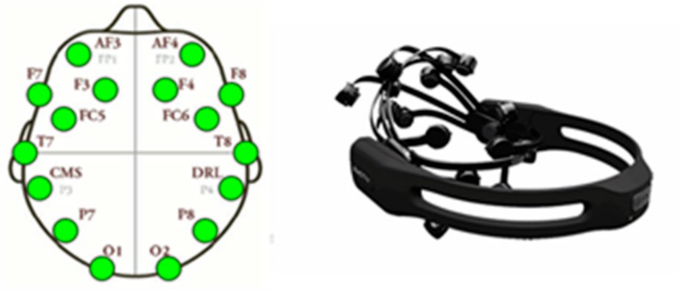


FIGURE 7. **Emotiv Inc.'s Epoc+ headset's electrode placement (left side) (right side), retrieved on December 29, 2021, from the firm's site at https://emotiv.gitbook.io/epoc-user-manual/.**

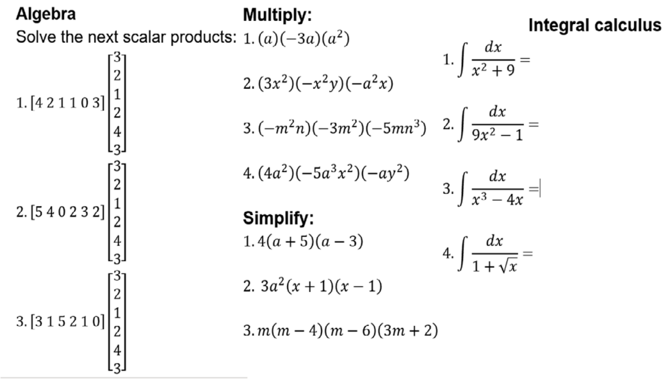
The following is a description of the planned data collection procedure. First, we put the headset on the subject and wet the electrodes. The video recording and EEG data collection can begin at this point. Next, we give the worksheet with the instructions to the test subject. Then, we allow the test participant to begin the exercise, like in a typical math school session. The attention sampling process ends when the subject finishes the exercises or fifteen to eighteen minutes have elapsed.

The Epoc+ BCI uses a 50 Hz/60 Hz dual notch filter and a 64 Hz low-pass filter for data acquisition. Then, the signal was downsampled to 128 Hz before transmission, and next, multiply the signal by $0.51x10^{-6}$ to convert it into a voltage reading.

A new sampling process begins at the same time as the attention recording, but now non-academic activities, like chattering with the staff, checking cell phones, or playing a video game.

Activity sheet

Appendix A shows the activity sheet used for this experiment. It has three scalar vector problems, four algebraic multiplications, three algebraic simplification problems, two synthetic division problems, and five fraction integrals. The idea of these problems is to measure the attention for 15 to 18 minutes while the students solve these problems, whether they were correctly solved. Figure 8 shows how the Activity Sheet looks.



Algebra
Solve the next scalar products:

1. $[4\ 2\ 1\ 1\ 0\ 3]\begin{bmatrix}3\\2\\1\\2\\4\\3\end{bmatrix}$
2. $[5\ 4\ 0\ 2\ 3\ 2]\begin{bmatrix}3\\2\\1\\2\\4\\3\end{bmatrix}$
3. $[3\ 1\ 5\ 2\ 1\ 0]\begin{bmatrix}3\\2\\1\\2\\4\\3\end{bmatrix}$

Multiply:

1. $(a)(-3a)(a^2)$
2. $(3x^2)(-x^2y)(-a^2x)$
3. $(-m^2n)(-3m^2)(-5mn^3)$
4. $(4a^2)(-5a^3x^2)(-ay^2)$

Simplify:

1. $4(a+5)(a-3)$
2. $3a^2(x+1)(x-1)$
3. $m(m-4)(m-6)(3m+2)$

Integral calculus

1. $\int \frac{dx}{x^2+9} =$
2. $\int \frac{dx}{9x^2-1} =$
3. $\int \frac{dx}{x^3-4x} =$
4. $\int \frac{dx}{1+\sqrt{x}} =$

FIGURE 8. The activity sheets are used for measuring attention. Source: self-made.

Brain-Computer interface

Brain-Computer Interfaces are electronic devices that acquire the EEG signals measured from the scalp and transmit the signals to a computer. They get a non-invasive recording of the brain's activity that can be processed and used for different applications, like gaming, controlling devices, and neuromarketing. While designing a variety of educational activities, the EEG signals are collected using an Emotiv Pro platform and an Epoc+ Brain-Computer Interface (BCI) and then analyzed using Matlab 2019a and Emotiv Pro software, Student Version [13][14]. The Emotiv Epoc+ headset (right) and a diagram of its electrode localization (left) are displayed in Figure 7. Electrodes F3, F4, P7, and P8 were chosen based on coherence analysis of attention [15][16]. This selection reduced the amount of data needed to be processed, which decreased processing time.

Data acquisition process

The data acquisition process for the math-solving stage was:

1. First, the individual will be fitted with the headgear containing the hydrated electrodes.
2. Initiate EEG data collection.
3. Give the test subject the worksheet and instructions.
4. Allow the test subject to begin solving math exercises for about 15 to 20 minutes.
5. Stop the data acquisition before 20 minutes of testing.

Figure 9 shows the headset placement, the data acquisition starts, the solving math exercises stage, and the EEG signals acquired. Next, for the non-attention stage, we repeat the same process for the math-solving stages. Still, in step 3, we let the user do any other activity, like talking to the staff, playing video games, checking social networks, and at the same time as the math-solving stage. This EEG data acquisition process was presented in reference [11].

FIGURE 9. EEG acquisition process with the Epoc+ BCI. Source: self-taken.

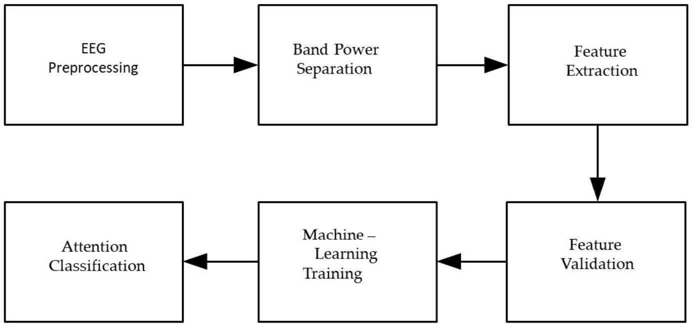


FIGURE 10. The method proposed a block diagram. Source: self-made.

Signal processing

The signal's processing procedure block diagram is shown in Figure 10 and is presented with more details in reference [11]. It begins with preprocessing and then calculating the power density of EEG data to split them into bands.

After the signal is preprocessed, the band power is separated, and features are extracted. After that comes the Attention Quantification result, followed by Validation Process for the Features and the ML Training stage.

Features listed in Section 2.3 will be obtained and verified next and used to teach algorithms for ML. We will go into greater detail about these actions in the following section.

To determine the PSD in absolute values, in V2/Hz, the Emotiv Pro Software estimates them using two-second windows. This 256-sample window spans two seconds [17]. Figure 11 illustrates power bands used in EEG. Delta is 1-4 Hz, Theta is 4-8 Hz, Alpha is 8-12 Hz, Beta is 12-30 Hz, and Ram (also known as Gamma) is 30-50 Hz.

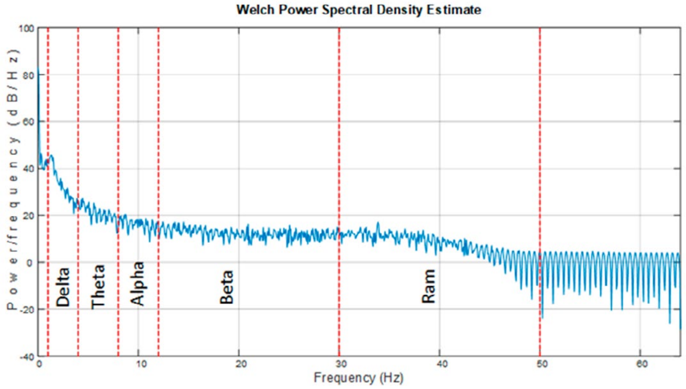


FIGURE 11. Example of band power separation used in EEG signal processing.

Feature extraction

Before the Theta-Beta (TBR) and the Theta-Alpha Ratios (TAR) can be detected, the band PSD of the EEG signal in two-second windows and for each electrode must be calculated. TBR characteristics, as well as the Theta and Beta Relative Powers and the Theta/(Alpha + Beta) relative power, are commonly used as part of attention detection and neurofeedback [11][12]. Table 2 shows the features and their equations.

Dataset

The dataset contains 24 features: six for each one of the four electrodes (F3, F4, P7, and P8) and two for the "Attention" and "No Attention" classes. This dataset is balanced by including exactly 104,244 samples from each category. The dataset depicted in Table 3 has 24 features derived from EEG data analysis. Here we show each electrode's features, minimum and maximum value, and feature type. The user was not paying attention to what he was learning since he was preoccupied with other, more pressing academic matters.

TABLE 2. Features used for attention measurement. Source: Self-made.

Feature	Equation
Theta Relative Power	$\text{RTP} = \frac{\theta}{T}$
Alpha Relative Power	$\text{RAP} = \frac{\alpha}{T}$
Beta Relative Power	$\text{RBP} = \frac{\beta}{T}$
Theta Beta Ratio	$\text{TBR} = \frac{\theta}{\beta}$
Theta Alpha Ratio	$\text{TAR}\,\frac{\theta}{\alpha}$
TBAR	$\text{TBAR} = \frac{\theta}{\beta+\alpha}$

TABLE 3. **Dataset example for the 24 features used. Source: self-made.**

Column name	Minimum	Maximum	Type	Column name	Minimum	Maximum	Type
Rtp_F3	0.052573953	0.986614597	No Attention	**Rtp_F4**	0	1	No Attention
Rap_F3	0.006339848	0.776645802	No Attention	**Rap_F4**	0	0.869707513	No Attention
Rbp_F3	0.004329732	0.889365625	No Attention	**Rbp_F4**	0	1	No Attention
Tbr_F3	0.059417402	203.8173059	No Attention	**Tbr_F4**	0	70.17964679	No Attention
Tar_F3	0.024821496	26.5031529	No Attention	**Tar_F4**	0	30.39827608	No Attention
Tbar_F3	0.055491352	73.70824549	No Attention	**Tbar_F4**	0	22.93691587	No Attention
Rtp_P7	0.012052623	0.953524263	No Attention	**Rtp_P8**	0.032478926	0.931698565	No Attention
Rap_P7	0.02263261	0.80631125	No Attention	**Rap_P8**	0.021125876	0.882221784	No Attention
Rbp_P7	0.013165841	0.926023652	No Attention	**Rbp_P8**	0.011678516	0.926483745	No Attention
Tbr_P7	0.015074921	68.46186848	No Attention	**Tbr_P8**	0.03505612	72.76456274	No Attention
Tar_P7	0.034479116	23.60027814	No Attention	**Tar_P8**	0.034120358	44.40627379	No Attention
Tbar_P7	0.01219966	20.51660325	No Attention	**Tbar_P8**	0.033569218	13.64098078	No Attention

Column name	Minimum	Maximum	Type	Column name	Minimum	Maximum	Type
Rtp_F3	0.016089806	0.954886933	Attention	**Rtp_F4**	0.014039459	0.93383383	Attention
Rap_F3	0.012248581	0.772430196	Attention	**Rap_F4**	0.016629818	0.895142743	Attention
Rbp_F3	0.008185096	0.963361451	Attention	**Rbp_F4**	0.017738286	0.937284135	Attention
Tbr_F3	0.016701733	113.1214596	Attention	**Tbr_F4**	0.031871758	38.06310801	Attention
Tar_F3	0.018291359	13.01625685	Attention	**Tar_F4**	0.020049065	39.94672025	Attention
Tbar_F3	0.016352922	21.16652665	Attention	**Tbar_F4**	0.014239372	14.11346364	Attention
Rtp_P7	0.019977678	0.902437957	Attention	**Rtp_P8**	0.021059371	0.930781794	Attention
Rap_P7	0.009337175	0.938064489	Attention	**Rap_P8**	0.025384797	0.862255367	Attention
Rbp_P7	0.018345381	0.955737888	Attention	**Rbp_P8**	0.01922771	0.867165619	Attention
Tbr_P7	0.027892137	45.20699485	Attention	**Tbr_P8**	0.033203085	46.58142716	Attention
Tar_P7	0.009788323	34.1572323	Attention	**Tar_P8**	0.044500332	25.36725644	Attention
Tbar_P7	0.020384922	9.249887846	Attention	**Tbar_P8**	0.021512409	13.44706619	Attention

RESULTS AND DISCUSSION

After training ML models on the datasets, we evaluate using the parameters listed in Table 4. True Positive values range from 7027 for Naive Bayes to 14968 for Random Forest. True Negative values range from a high of 17219 in Naive Bayes to a low of 12682 in Decision Trees. Again, Naive Bayes had the highest False Positives value at 13818, while Random Forest had the lowest at 5877. False Negatives ranged from a high of 7797 for Decision Trees to a low of 3260 for Naive Bayes. Accuracy was maximized by Random Forest (0.7392) and minimized by Naive Bayes (0.088). Random Forest achieved an F1 Score of 0.7430, while SGD managed only 0.6215 %. Random Forest had the highest Specificity/Accuracy (0.7261), while Naive Bayes (0.5547) had the lowest. Naive Bayes had the greatest Sensitivity/Recall score of 0.8408, while Decision Trees had the lowest at 0.6192.

Table 5, shows the results of the parameters of the evaluated ML algorithms. Table 5, also included in, presents the performance metrics obtained with the

ML algorithms evaluated. Random Forest obtains the highest AUC value with 0.7394, and Naïve Bayes has the lowest with 0.5889. The highest Cohen's Kappa coefficient was for Random Forest with 0.4787, while the lowest was 0.1771 for Naïve Bayes. Naïve Bayes had the highest Hamming Loss with 0.41327, and Random Forest had the lowest with 0.2607. Random Forest had Matthew's Correlation Coefficient of 0.4792, while Naive Bayes and Support Vector Machine had values as low as 0.2057.

Table 5 shows how the Area Under the Curve (AUC) of the ML models used in this paper for Attention Classification behaves. It is noticed that Random Forest presents the best performance among the evaluated models with a 0.7394 value. The second best AUC was for Extra Trees, with 0.7335, followed by K-NN, with 0.6917. The worst performance was for Naïve Bayes, with a 0.5889 value. Comparing the results obtained from this research with those obtained in [11], we observe that the ML model performance is lower due to the use of several individuals in the dataset elaboration. Using many test individuals for dataset conformation reduces the performance of the ML models evaluated compared to the same models trained with datasets from one individual.

TABLE 4. **Results of the parameters of the evaluated M-L algorithms. Source: self-made.**

	Machine Learning algorithms							
Results of the evaluated parameters	Naive Bayes	SGD	Decision Trees	(SVM)-rbf	KNN	MLP-NN	Random Forest (RF.)	Extra trees
True positives	7027	12384	13296	12318	13807	14423	14968	14905
True negatives	17219	13049	12682	14758	14767	13675	15582	15400
False positives	13818	8461	7549	8527	7038	6422	5877	5940
False negatives	3260	7430	7797	5721	5712	6804	4897	5079
Accuracy	0.5867	0.6154	0.6286	0.6552	0.6914	0.6804	0.7392	0.7333
F1 Score	0.6684	0.6215	0.6230	0.6744	0.6984	0.6740	0.7430	0.7365
Specificity/Accuracy	0.5547	0.6066	0.6268	0.6337	0.6772	0.6804	0.7261	0.7216
Sensitivity/Recall	0.8408	0.6371	0.6192	0.7206	0.7210	0.668	0.76087	0.7519

TABLE 5. **Performance metrics of the evaluated M-L algorithms. Source: self-made.**

	Performance metrics			
Machine-Learning Algorithm	AUC	Cohen's Kappa coefficient	Hamming loss	Matthew's correlation coefficient
Naive Bayes	0.5889	0.1771	0.41327	0.2057
Stochastic Gradient Descent	0.6156	0.2311	0.3845	0.2314
Decision trees	0.6285	0.2571	0.3713	0.2571
Support Vector Machine (SVM) -rbf	0.6557	0.3111	0.4132	0.2057
KNN	0.6917	0.3832	0.3085	0.3840
Extra trees	0.7335	0.4668	0.2666	0.4672
MLP- NN	0.6798	0.3597	0.3200	0.3597
Random Forest (RF)	0.7394	0.4787	Random	0.4792

CONCLUSIONS

In this research, it is observed that when forming the dataset with ten users, the results are lower than when working with only one user. However, the results obtained are considered acceptable. It is observed that the Random Forest model presented the best performance for the parameters: F1 Score, Accuracy, Area Under the Curve, and Specificity/Precision. In Sensitivity/Recall, Naive Bayes had the best results. We also conclude that using several individuals for dataset conformation reduces the performance compared to a single-user dataset. Increasing the number of test subjects is necessary to increase the dataset. We conclude that this kind of ML system is better if personalized for a specific user than making it general for more users because the ML models tested performance decreases. This performance reduction is caused by the variability of the data obtained from different users. So it is better to obtain personalized features for the training datasets than to obtain generalized features for more than one user.

In future work, it is necessary to perform more tests to compare the performance of the ML models trained with datasets formed with single-user samples or personalized and other datasets made with the samples of more users and compare the performance of the trained models. Also, it is important to program these models on embedded systems and evaluate their performance.

AUTHOR CONTRIBUTIONS

J.J.E.E. conceptualized the experiments, supervised the feature extractions, and wrote the original draft; L.J.B. participated in the methodology and formal analysis. A.S.T. contributed to data curation. M.Z.A. collaborated with the dataset elaboration and validation. D.A.T.T. oversaw the review and editing of the manuscript. O.R.L.B. supervised the data acquisition, G.M.G.A. performed the validation of the process, and the writing and editing of the paper. J.R.J.R. participated in the software validation. A.L.R. and A.S.M.V. realized the EEG data acquisition process and feature extraction. M.M.M.L and I.A.R.P. also participated in signal preprocessing and feature extraction.

REFERENCES

[1] E. Levin, J. Bernier, "Attention Span," in Encyclopedia of Child Behavior and Development, S. Goldstein, J. A. Naglieri, Eds. New York, U. S.: Springer, 2011, pp. 163, doi: https://doi.org/10.1007/978-0-387-79061-9_22

[2] M. A. Rasheed, P. Chad, S. Ahmed, H. Sharif, Z. Hoodbhoy, A. Siddiqui, B. S. Hasan, "Use of artificial intelligence on Electroencephalogram (EEG) waveforms to predict failure in early school grades in children from a rural cohort in Pakistan," PLoS One, vol. 16, no. 2, art. no. e0246236, Feb. 2021, doi: https://doi.org/10.1371/journal.pone.0246236

[3] PLoS One, "Correction: Use of artificial intelligence on Electroencephalogram (EEG) waveforms to predict failure in early school grades in children from a rural cohort in Pakistan," PLoS One, vol. 16, no. 2, art. no. e0247744, Feb. 2021, doi: https://doi.org/10.1371/journal.pone.0247744

[4] M. F. Mavilidi, K. Ouwehand, N. Riley, P. Chandler, F. Paas, "Effects of An Acute Physical Activity Break on Test Anxiety and Math Test Performance," Int. J. Environ. Res. Public Health, vol. 17, no. 5, art. no. 1523, Feb. 2020, doi: https://doi.org/10.3390%2Fijerph17051523

[5] M. Hartmann, J. Laubrock, M. H. Fischer, "The visual number world: A dynamic approach to study the mathematical mind," Q. J. Exp. Psychol., vol. 71, no. 1, Jan. 2018, doi: https://doi.org/10.1080/17470218.2016.1240812

[6] W. Auccahuasi, C. Ovalle, Z. Ayvar, J. Aybar, et al., "Methodology to evaluate the levels of attention and meditation in the development of virtual classes through BCI devices," in Workshop Technol. Innov. Edu. Knowledge Dissemination CEUR-WS, Goa, India, 2021, pp. 51-60. [Online]. Available: https://ceur-ws.org/Vol-2869/PAPER_06.pdf

[7] S. M. Manjur, M.-B. Hossain, P. A. Constable, D. A., Thompson, et al., "Detecting Autism Spectrum Disorder Using Spectral Analysis of Electroretinogram and Machine Learning: Preliminary results," 2022 44th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 3435-3438, doi: https://doi.org/10.1109/EMBC48229.2022.9871173

[8] B. Ari, N. Sobahi, Ö. F. Alçin, A. Sengur, U. R. Acharya, "Accurate detection of autism using Douglas-Peucker algorithm, sparse coding based feature mapping and convolutional neural network techniques with EEG signals," Comput. Biol. Med., vol. 143, art. no. 105311, Apr. 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105311

[9] A. Shoeibi, M. Rezaei, N. Ghassemi, Z. Namadchian, A. Zare, J. M. Gorriz, "Automatic Diagnosis of Schizophrenia in EEG Signals Using Functional Connectivity Features and CNN-LSTM Model," in Artif. Intell. Neurosc. Affective Analysis and Health Applications (IWINAC 2022). J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, H. Adeli, Eds. Puerto de la Cruz, Tenerife, Spain, 2022, pp. 63-73, doi: https://doi.org/10.1007/978-3-031-06242-1_7

[10] F. J. Alvarado-Rodríguez, K. P. Ibarra-González, C. Eccius-Wellmann, H. Vélez-Pérez, R. Romo-Vázquez, "Electrophysiological Brain Response to Error in Solving Mathematical Tasks," Mathematics 2022, vol. 10, no. 18, art. no. 3294, Sep. 2022, doi: https://doi.org/10.3390/math10183294

[11] J. J. Esqueda-Elizondo, R. Juárez-Ramírez, O. R. López-Bonilla, E. E. García-Guerrero, et al., "Attention Measurement of an Autism Spectrum Disorder User Using EEG Signals: A Case Study," Math. Comput. Appl., vol. 27, no. 2, art. no. 21, Mar. 2022, doi: https://doi.org/10.3390/mca27020021

[12] F. Fahimi, C. Guan, W. B. Goh, K. K. Ang, C. G. Lim, T. S. Lee, "Personalized features for attention detection in children with Attention Deficit Hyperactivity Disorder," in 2017 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jeju, Korea (South), Jul. 2017, pp. 414-417, doi: https://doi.org/10.1109/embc.2017.8036850

[13] K. H. Khng, R. Mane, "Beyond BCI—Validating a wireless, consumer-grade EEG headset against a medical-grade system for evaluating EEG effects of a test anxiety intervention in school," Adv. Eng. Inform., vol. 45, art. no. 101106, Aug. 2020, doi: https://doi.org/10.1016/j.aei.2020.101106

[14] I. A. Fouad, "A robust and reliable online P300-based BCI system using Emotiv EPOC+headset," J. Med. Eng. Technol., vol. 45, no. 2, pp. 94-114, Feb. 2021, doi: https://doi.org/10.1080/03091902.2020.1853840

[15] N. V. Dubrovinskaya, R. I. Machinska, Y. V. Kulakovsky, "Brain Organization of Selective Tasks Preceding Attention: Ontogenetic Aspects," in Complex Brain Functions, R. Miller, A. M. Ivantsky, Eds. Boca Ratón, FL, USA: CRC Press, 2000, ch. 9 pp. 169-180. doi: https://doi.org/10.4324/9780203304785

[16] N. V. Dubrovinskaya, R. I. Machinskaya, "Reactivity of θ and α EEG Frequency Bands in Voluntary Attention in Junior Schoolchildren," Hum. Physiol., vol. 28, no. 5, pp. 522-527, Sep. 2002, doi: http://dx.doi.org/10.1023%2FA%3A1020266516114

[17] Emotiv Inc., Frequency Bands - EmotivPRO v3.0, (2020). Accessed: Nov. 2021. [Online]. Available: https://emotiv.gitbook.io/emotiv-pro-v3/data-streams/frequency-bands

RESEARCH ARTICLE
VOL. 44 | NO. 3 | **SPECIAL ISSUE 2023** | PP 38 - 52

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.3

E-LOCATION ID: 1364

Feature Selection of Motor Activity in Intervals of Time with Genetics Algorithms for Depression Detection

Selección de Características de la Actividad Motora en Intervalos de Tiempo con Algoritmos Genéticos para la Detección de Depresión

Carlos H. Espino-Salinas [1], *Carlos E. Galván-Tejada*[1], *Ana G. Sánchez-Reyna*[1], *Huizilopoztli Luna-García*[1], *Hamurabi Gamboa-Rosales*[1], *Jorge A. Morgan-Benita*[1], *José M. Celaya-Padilla*[1], *Jorge I. Galván-Tejada*[1],

[1]Universidad Autónoma de Zacatecas, Zacatecas - México

ABSTRACT

It is estimated that depression affects more than 300 million people in worldwide. Unfortunately, the current method of psychiatric evaluation requires a great effort on the part of clinicians to collect complete information. The aim of this paper is determine the optimal time intervals to detect depression using genetic algorithms and machine learning techniques; from motor activity readings of 55 participants during a week at one-minute intervals. The time intervals with the best performance in detecting depression in individuals were selected by applying Genetic Algorithms (GA). Methodology. 385 observations of the study participants were evaluated, obtaining an accuracy of 83.0 % with Logistic Regression (LR). Conclusion. There is a relationship between motor activity and people with depression since it is possible to detect it using machine learning techniques. However, the changes in the variables of the time intervals could be established as key factors since, at different times, they could give good or bad results because the motor activity in the patients could vary. However, the results present a first approximation for developing tools that help the opportune and objective diagnosis of depression.

KEYWORDS: artificial intelligence, depression, feature selection, genetic algorithm, motor activity

RESUMEN

Se estima que la depresión afecta a más de 300 millones de personas en el mundo. Desafortunadamente, el método de evaluación psiquiátrica actual requiere un gran esfuerzo por parte de los médicos para recopilar información completa. Objetivo. Determinar los intervalos de tiempo óptimos para detectar depresión mediante algoritmos genéticos y técnicas de aprendizaje automático, a partir de las lecturas de actividad motora de 55 sujetos durante una semana en intervalos de un minuto. Los intervalos de tiempo con mejor desempeño en la detección de depresión en individuos fueron seleccionados aplicando algoritmos genéticos. Metodología. Se evaluaron 385 observaciones de los sujetos de estudio, obteniendo una precisión del 83.0 % con Regresión Logística (LR). Conclusión. Existe una relación entre la actividad motora y las personas con depresión ya que es posible detectarla utilizando técnicas de aprendizaje automático. Sin embargo, los cambios en las variables de los intervalos de tiempo podrían establecerse como factores clave ya que en diferentes momentos podrían dar buenos o malos resultados debido a que la actividad motora en los pacientes podría llegar a variar. No obstante, los resultados presentan una primera aproximación para el desarrollo de herramientas que ayuden al diagnóstico oportuno y objetivo de la depresión.

PALABRAS CLAVE: actividad motora, algoritmos genéticos, depresión, inteligencia artificial, selección de características

Corresponding author

TO: **Carlos E. Galván-Tejada**
INSTITUTION: **Universidad Autónoma de Zacatecas**
ADDRESS: **Jdn. Juárez #147, Centro Histórico, 98000 Zacatecas, Zac.**
EMAIL: **ericgalvan@uaz.edu.mx**

Received:

19 July 2023

Accepted:

14 September 2023

INTRODUCTION

Per the World Health Organization (WHO), depression is distinguished from typical mood variations and brief emotional reactions to everyday life challenges. Particularly when it becomes recurrent and exhibits moderate or severe intensity, depression can evolve into a significant health concern [1]. Generally, it emerges early in life, causing a substantial decline in the overall functioning of individuals. The condition tends to recur and imposes notable economic and social burdens, making it a prominent contributor to the list of debilitating illnesses [2]. In its most severe form, depression can tragically result in suicide, with nearly one million people committing suicide each year. Shockingly, it stands as the second leading cause of death among individuals aged 15 to 29, as reported by WHO data. Additionally, in the past, the COVID-19 pandemic has brought forth numerous overwhelming stresses. Some evident factors include job loss, bereavement of family members, friends, or coworkers, financial instability, and social isolation, especially for individuals living alone. When required, healthcare providers must differentiate between demoralization and depression; however, access to in-person consultations with qualified mental health experts may not be easily accessible to everyone in need [3].

Apart from the difficulties posed by the inability to meet patients with depression in person, there are inherent challenges in conducting psychiatric assessments. Such evaluations require considerable effort from specialists to collect objective patient information. Moreover, successful assessment heavily relies on the patient willingness to cooperate and effectively communicate their symptoms and concerns [4]. One of the techniques employed to assess patients' depression is the Montgomery-Asberg Depression Rating Scale (MADRS), designed to gauge the current severity of ongoing depression [5]. Clinicians evaluate ten depression-relevant items through observations and discussions with the patient, and a cumulative score (ranging from 0 to 60) indicates the level of depression. Scores below ten are categorized as having no depressive symptoms [6]. While scores above 30 indicate a severe depressive state [7]. Therefore, an objective detection mechanism based on biological signals is needed to improve timely diagnosis.

The extensive use of wearable devices for monitoring mental and physical health has gained significant popularity. Now, people are consistently collecting data to improve their well-being and monitor fitness advancements. Moreover, the data collected from these devices can hold considerable value from a psychiatric perspective, extending beyond evaluating the overall quality of life. It holds the potential to aid in diagnosing various mental health conditions, including depression [8]. Undoubtedly, motor activity reflects social patterns influenced by cyclical biological rhythms, which are regulated by the 24-hour circadian clock and interwoven with several ultradian rhythmic cycles lasting 2 to 6 hours [9]. Disrupted biological rhythmic patterns have been proposed as significant indicators of mood episodes [10]. Actigraphy serves as a non-intrusive approach to observing human rest and activity patterns. Typically, it involves using a wrist-worn device to record gravitational acceleration units [11].

The actigraph is among the devices frequently employed to collect motor activity data. Numerous research studies have leveraged this data to create models using artificial intelligence techniques for classifying, detecting, and monitoring the illness. García-Ceja *et al.*, [12] utilize machine learning to differentiate between depressed and non-depressed patients. To assess the algorithms' performance, they employ leave-one-patient-out validation. The collective results reveal that sensor data contains valuable information for determining an individual depression status. On the other hand, Zanella-Calzada *et al.*, [13] proposed a novel approach to distinguish depressive participants from control participants using data from their wearable device-recorded motor activity. Statistical features were extracted from the motor activity signals, which were then utilized to train a random forest classifier. Galván-Tejada *et al.*, [14] investigated the accelerometer signal from smart bands to identify depressive states based on patients' activity. A statistical feature extraction technique was devised, focusing on the temporal and spectral evolution of the signal. Furthermore, an intelligent feature selection method utilizing GA was incorporated to optimize the non-invasive diagnostic process efficiently The results demonstrate the potential to distinguish between depressive states using the smart band activity signal, offering a preliminary and automated tool for almost real-time depression diagnosis at a lower computational cost to specialists.

Moreover, researchers have explored the utilization of motor activity in time series data to gather valuable insights into identifying potential cases of depression, among other applications. Frogner *et al.*, [15] Initially employed One-Dimensional Convolutional Neural Networks (1D-CNN) to assess motor activity for depression detection. Subsequently, the study extended its scope to identify three levels of depression (no depression, mild, and severe) using the MADRS scale. The final model successfully predicts the MADRS scores of the participants.

Rodríguez-Ruiz *et al.*,[16] introduced a series of models aimed at classifying depressive and non-depressive episodes throughout different moments of the day (day, night, and full day) based on participants motor activity levels. The Depresjon database, containing activity data from both depression patients and controls, was utilized in the study. Additionally, they proposed a Random Forest Classifier (RFC) model for multiclass classification, distinguishing schizophrenia, depression, and healthy controls using night-time activity data with an impressive 98 % accuracy in detecting all three classes. Experimental results demonstrated the model efficacy in identifying episodes of depression and schizophrenia, as well as healthy controls, surpassing prior studies that employed computationally expensive algorithms such as CNN and Bidirectional Recurrent Neural Networks (BRNN), resulting in a noteworthy boost in accuracy. [17].

Jakobsen *et al.*, [18] conducted a study to investigate the potential of various machine learning algorithms in distinguishing between depressed patients and healthy controls using motor activity time series. Furthermore, their research demonstrated that machine learning capacity to reveal hidden patterns in the data aligns with the conclusions drawn from previous studies utilizing both linear and nonlinear statistical methods in motor activity analysis.

Artificial Intelligence (AI), has proven to be useful tool for detecting cases of depression. Kour *et al.*, [19] propose that unipolar depression and bipolar depression display similar clinical symptom profiles, presenting a considerable challenge in distinguishing between the two depression types. The disruptions in motor activity offer a potential avenue to detect pathological mental states and may prove valuable in addressing this diagnostic challenge. That is why every day, there is more research studying this type of data where different types of algorithms derived from AI have been implemented, as is the case with Pacheco-González *et al.*, [20] conducted a comparative analysis of various classification techniques, including conditional inference trees, random forest, K-Nearest Neighbor, support vector machine, and Naïve Bayes. The study aimed to predict depressive states based on patients' activity, measured using a smart band accelerometer. Conversely, Raihan *et al.*,[21] employed a combination of motor sensor readings and demographic data along with machine learning techniques like Random Forest (RF), AdaBoost, and Artificial Neural Networks (ANN). Finally, as mentioned by Singh *et al.*, [22], identifying disruptions in motoric activity could serve as a valuable approach to detecting pathological mental states. Thus, creating distinct motor activity database (Depresjon) containing data from patients with unipolar depression, bipolar depression, and healthy individuals has demonstrated effectiveness in the timely detection of depression cases.

The objective of this research is to present an approach capable of objectively identifying episodes of depression by employing various AI algorithms in order to explore a wider range of options for detecting depression, this using a limited amount of data intelligently selected with genetics algorithms that reduced the dimensionality and redundancy of the data, it can obtain model capable of training simply with new input data and processing it to generate models adapted to different types of patients, in different environments. It also seeks to show that the behavior of each patient as well as each person is always different, depending on the situation or environment in which they find themselves, therefore, it is important to create adaptive artificial intelligence models for each participants to detect depression in a timely manner and help mental health specialist to develop a treatment for each degree of depression that can be found. Critical analysis of the current relevant literature, statement of the general aims of the work and the importance of the same. In the text of the article, the references must be numbered in the order that they appear.

MATERIALS AND METHODS

The methodology proposed in this work consists of five main stages as shown in Figure 1. Initially, the data are obtained from the Depresjon database. Next, in the second phase of the process, data pre-processing, a comprehensive explanation is given on how the dataset will be organized to extract each minute as a feature from the original data obtained in the initial step. This arrangement enables a subsequent feature selection employing GA. Lastly, to verify the significance of the chosen features, two crucial stages were undertaken: the first involved classification analysis through the implementation of AI algorithms such as Logistic Regression (LR), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Decision Tree (DT). The second stage focused on result validation.

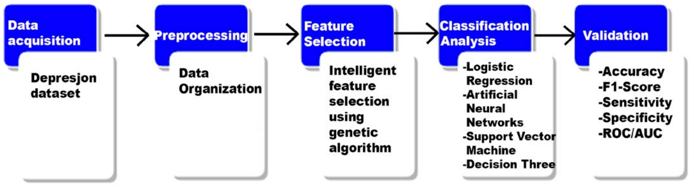


FIGURE 1. **Proposed methodology for intelligent feature selection for objective detection of depressive symptoms.**

Data Acquisition

The motor activity dataset comprises patient data monitored using an actigraph watch on the right wrist this measure activity using a piezoelectric accelerometer. This actigraph watch, named "Actiwatch AW4", measures activity levels at a sampling frequency of 32 Hz, recording movements above 0.05 g. These movements correspond to particular voltage (v) values, which are stored as activity counts in the Actiwatch memory. The count values directly correlate to the intensity of the movements. Continuous recording of total activity counts occurred at one-minute intervals, with the activity counts being recorded accordingly at the same one-minute intervals [1]. Information was gathered from a group of 22 psychotic patients and all used antipsychotic medications who were admitted to Hauklend University Hospital, the mean age to the first time of hospitalization was 24 +/- 9.3. The specialist diagnosed the patients using a semi structured interview based on Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [23]. Within this group, were 3 females and 19 males, with an average age of 42.6 years (ranging from 27 to 69 years). As for the healthy control group included, it included 23 hospital employees, five students, and four individual practitioner clinic [18].

Data Pre-processing

To obtain a dataset consisting of a certain number of observations capable of being processed through AI techniques, a period of a week to maintain data diversity was taken in one minute intervals. we reordered them to identify the number of observations per participant of study as a relation between participant-day, with their respective registered data.

Motor Activity Data (MAD) for each study participant is filed in "C" columns, spanning a full day from 00:00 to 23:59, with each minute corresponding to its respective MAD record. These columns are then transposed and added to the first row of an "A" matrix, accompanied by the corresponding output, where 0 indicates the control group, and 1 indicates the condition group based on the source dataset, successively. The same process is applied for the subsequent participating days until completing a week of information, resulting in a matrix A = [385 x 1441], including its output. The process is explained in more detail below.

$$
C_{11} = \begin{array}{c} 00{:}00 \\ 00{:}01 \\ 00{:}02 \\ 00{:}03 \\ 00{:}04 \\ \vdots \\ 23{:}59 \end{array} \begin{bmatrix} 168\ v \\ 107\ v \\ 550\ v \\ 157\ v \\ 0\ \ v \\ \vdots \\ 208\ v \end{bmatrix}
C_{12} = \begin{array}{c} 00{:}00 \\ 00{:}01 \\ 00{:}02 \\ 00{:}03 \\ 00{:}04 \\ \vdots \\ 23{:}59 \end{array} \begin{bmatrix} 168\ v \\ 17\ v \\ 560\ v \\ 167\ v \\ 23\ \ v \\ \vdots \\ 280\ v \end{bmatrix}
$$

$$
C_{13} = \begin{array}{c} 00{:}00 \\ 00{:}01 \\ 00{:}02 \\ 00{:}03 \\ 00{:}04 \\ \vdots \\ 23{:}59 \end{array} \begin{bmatrix} 118\ v \\ 157\ v \\ 450\ v \\ 157\ v \\ 0\ \ v \\ \vdots \\ 305\ v \end{bmatrix}
C_{14} = \begin{array}{c} 00{:}00 \\ 00{:}01 \\ 00{:}02 \\ 00{:}03 \\ 00{:}04 \\ \vdots \\ 23{:}59 \end{array} \begin{bmatrix} 148\ v \\ 157\ v \\ 120\ v \\ 457\ v \\ 121\ v \\ \vdots \\ 892\ v \end{bmatrix}
$$

$$
C_{15} = \begin{array}{c} 00{:}00 \\ 00{:}01 \\ 00{:}02 \\ 00{:}03 \\ 00{:}04 \\ \vdots \\ 23{:}59 \end{array} \begin{bmatrix} 168\ v \\ 107\ v \\ 550\ v \\ 157\ v \\ 0\ \ v \\ \vdots \\ 208\ v \end{bmatrix}
C_{16} = \begin{array}{c} 00{:}00 \\ 00{:}01 \\ 00{:}02 \\ 00{:}03 \\ 00{:}04 \\ \vdots \\ 23{:}59 \end{array} \begin{bmatrix} 168\ v \\ 17\ v \\ 560\ v \\ 167\ v \\ 23\ \ v \\ \vdots \\ 280\ v \end{bmatrix}
$$

$$
C_{17} = \begin{array}{c} 00{:}00 \\ 00{:}01 \\ 00{:}02 \\ 00{:}03 \\ 00{:}04 \\ \vdots \\ 23{:}59 \end{array} \begin{bmatrix} 118\ v \\ 157\ v \\ 450\ v \\ 157\ v \\ 0\ \ v \\ \vdots \\ 305\ v \end{bmatrix} \tag{1}
$$

Given the columns of the days corresponding to week 1, we would have C_{11}, C_{12}, C_{13}, C_{14}, C_{15}, C_{16}, C_{17}, where the first subscript indicates the week number and the second indicates the day number with the accompanying timestamp of the motor activity record generated during that specific minute as shown in Equation 1. To obtain the total number of time intervals defined as characteristics corresponding to a day, a simple operation was carried out consisting of multiplying the 60 minutes that an hour contains by the number of hours that a day contains to obtain, as a result, a column that contains 1440 rows. These rows represent the minutes of a day with their respective MAD record, thus defining the time range to be used.

Afterward, Following the transposition, the columns are reorganized to create a vector V^T, where each one-minute interval signifies a feature. Each vector corresponds to an observation per participant, culminating in a row with 1440 columns. These vectors are then added to the matrix “A” rows. Each participant, whether depressed or not (labeled as one and zero), has their vector extracted daily, spanning 7 days (one week). Each vector is appended to matrix “A” simultaneously with the assignment of its output label. This process continues for the next participants until all 55 participants are incorporated into the new dataset configuration, as illustrated in Equation (2). This setup is utilized for feature selection to identify the most relevant minutes for depression detection.

$$
\begin{array}{lccccccc}
 & 00{:}00 & 00{:}01 & 00{:}02 & 00{:}03 & 00{:}04 & \dots & 23{:}59 \\
V_{1,1}^{T} = [& 256v & 107\ v & 550\ v & 157\ v & 0\ v & \dots & 208] \\
V_{1,2}^{T} = [& 168\ v & 17\ \ v & 560\ v & 167\ v & 23\ \ v & \dots & 280] \\
V_{1,3}^{T} = [& 118\ v & 157\ v & 450\ v & 157\ v & 0\ v & \dots & 305] \\
V_{1,4}^{T} = [& 148\ v & 157\ v & 120\ v & 457\ v & 121\ v & \dots & 892] \\
V_{55,7}^{T} = [& 563\ v & 124\ v & 456\ v & 321\ v & 265\ v & \dots & 201]
\end{array} \tag{2}
$$

Feature Selection

Bearing in mind that the data collected through a portable device known as an activewatch in intervals of one minute are considered here as characteristics for each observation, in this section, we seek to select the minutes that provide the most significant information to generate a simple, intelligent model capable of detecting depression, considering that the number of features is much higher compared to observations, this process also aims to reduce redundancy and the amount of data to process to improve its performance.

The methodology for intelligent feature selection using Genetic Algorithms (GA) is shown in Figure 2.

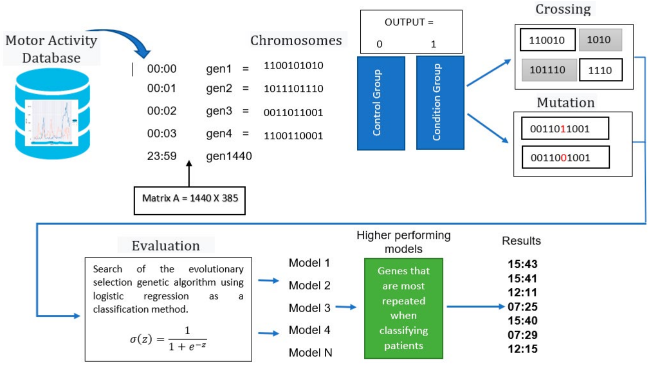


FIGURE 2. Methodology of the GA to select the most important features to classify patients with depression using Logistic Regression (LR).

GA draw inspiration from nature, particularly the process of natural selection. They implement a population-based search approach grounded in the fundamental principle of 'survival of the fittest.' At their core, GAs consist of several essential components: chromosome representation, selection, crossover, mutation, and fitness function computation [23].

In the context of the research, A library of genetic algorithms known as Genetic Algorithms for Multivariate Statistical Models from Large-Scale Functional Genomic Data 1.4 (GALGO) [24]. was implemented. GALGO, designed specifically for the R programming language, serves the purpose of selecting models with high fitness. This process begins by creating a random population of features with a specified size (n). These features are evaluated using a fitness function to assess their ability to classify the dependent variable, typically yielding an accuracy value.

For the specific task of depression detection, we utilized the logistic function as our classification method. In our study, the genetic algorithm iteratively explores and evolves combinations of genes (intervals of time) from a dataset comprising two classes. The goal is to identify features that effectively distinguish between these classes using the logistic regression (LR) method over 500 generations, considering 1000 possible solutions, and aiming for a fitness level of 95 %. As a result, the intervals of time obtained represent the most relevant contributors to the processed dataset. They are consistently favored within the intelligent models generated by the GA, demonstrating superior performance in classifying subjects.

Classification Analysis

Before starting the analysis of the dataset, a z-score normalization process is applied to the MADs to contain outliers as their variation in some time intervals can be significant. This normalization also speeds up the training time for each function within the same scale and is especially beneficial for modeling applications where the inputs often have varying scales. The mean and standard deviations are calculated for each feature [25], as shown in Equation 3. Then generate binary classification models since the data contains two possible classes as output, depressed (represented as 1) and not depressed (represented as 0). The techniques used to develop the models are ANN, LR, DT, and SVM. Finally, a comparison of the performance of these techniques was applied.

$$Z_i = \frac{x_i - \bar{x}}{\sigma} \tag{3}$$

The artificial neural network consists of two multilayer perceptron (4, 2) and an output layer with a logistic activation function. On the other hand, the parameters established for the classification SVM are a linear kernel with a cost of 1.

The process for the development of the models consists in two steps, training, and testing. Therefore, the data were randomly subsampled into two sets, one for each step. The dataset encompassed 80 % of data for training process, while the remaining 20 % was reserved for testing.

Validation

These parameters comprise accuracy, sensitivity, specificity, The Receiver Operating Characteristic Curve with the Area Under the Curve (ROC/AUC), and the F1-Score.

Accuracy (Acc) is a performance criterion that indicates the degree to which the outcome of a calculation aligns with the correct value [13]. as represented in Equation (4).

$$Accuracy(1 - Error) = \frac{TP + TN}{CP + CN} \tag{4}$$

In this context, TP represent True Positives, TNstands for True Negatives, CP denotes Truly Positive, and CN indicates Truly Negative.

The accuracy of a classifier encompasses other crucial aspects like precision, which refers to the number of correctly detected targets among all the targets detected. Additionally, there exists a relationship between the number of correctly detected targets and all known true targets, known as recall, as shown in Equations (5) and (6).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

FP is False Positive, and FN is False Negative.

The measure that takes into account both precision and recall to evaluate the classification capability of an algorithm is referred to as the F1-Score. It is defined as the harmonic mean of precision and recall [26], as shown in Equation (7).

$$F_1 = \frac{precision \cdot recall}{precision + recall} \tag{7}$$

Sensitivity, is the ability to accurately identify data with depressed symptoms, represents the number of condition participants correctly identified [27]. It is calculated using Equation (8).

$$Sensitivity(1 - \beta) = \frac{TP}{CP} \tag{8}$$

Specificity, referring to the ability to identify data without the condition as healthy, measures the proportion of negative samples that are accurately classified as such [28]. It is calculated using the following Equation (9).

$$Specificity(1 - \alpha) = \frac{TN}{CN} \tag{9}$$

The ROC curve is a commonly used method for evaluating machine learning models. It provides a visual representation of the classifier performance, enabling the selection of an appropriate operating point, referred to as the decision threshold, along with the AUC value [29]. The AUC can be calculated through integration, as shown in Equation (10).

$$AUC = \sum_i (1 - \beta_i \cdot \Delta\alpha) + \frac{1}{2}[\Delta(1 - \beta) \cdot \Delta\alpha] \tag{10}$$

RESULTS AND DISCUSSION

In this section, we explain the performed experiments and discuss the results.

Initially, Figure 3 shows the motor activity of depressed and non-depressed people throughout the day, extracted from the research paper entitled: Two-Dimensional Convolutional Neural Network for Depression Episodes Detection in Real Time Using Motor Activity Time Series of Depresjon Dataset [30]. As evident from the data, a significant distinction between the two cases can be observed, with a noticeable decrease in movement observed among depressed individuals. The data was collected for each of the 55 participants under study (32 healthy and 23 depressed) for a week. Each day was treated as an individual observation for each participant, resulting in a total of 385 observations, providing enough data to develop a classification model for episodes of depression. This was achieved by multiplying the number of patients by the number of days in a week (55X 7).

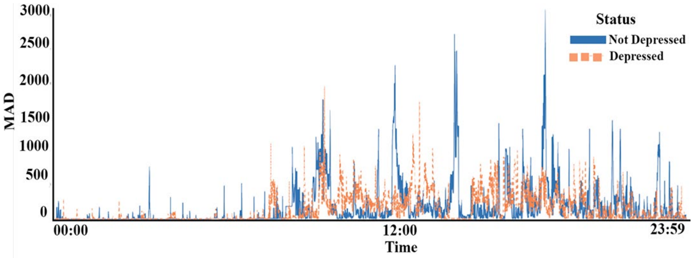


FIGURE 3. Samples collected with Actiwatch from both a non-depressed and a depressed subject of the Depresjon database [30].

The dataset comprises 385 observations, each containing 1440 minutes. Following the feature selection process, the most significant features were chosen, specifically: 15:43, 15:41, 12:11, 7:25, 15:40, 7:29, and 12:15, as depicted in Figure 4. The graph illustrates the frequency of appearance for each feature, with the highest-ranked features displayed in black.

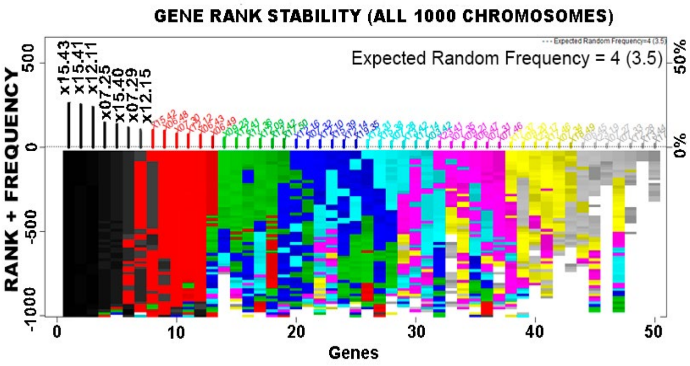


FIGURE 4. Most relevant features obtain out of the Genetic Algorithm.

Different artificial intelligence techniques were applied and validated with different metrics obtaining the following results. The ANN has an accuracy of 0.74, DT has 0.73, SVM obtained a performance of 0.81, and LR 0.83. Considering that logistic regression presents the best level of accuracy, it can be concluded that this technique can correctly classify a greater number of test participants than the others. To know the number of true positives and true negatives, as well as false positives and false negatives that were obtained by the different implemented algorithms, Table 1 shows the results of their confusion matrices which is often applied in machine learning to evaluate or visualize the model behavior in supervised classification scenarios [31]. Additionally, we add the Matthews Correlation Coefficient (MCC) which is a more reliable statistical index that produces a high score only if the prediction obtained good results in the four categories of the confusion matrix [32] and also the Kappa Correlation Coefficient (KCC) which is a metric to summarize the agreement between two nominal classifications, based on the same categories [33].

TABLE 1. Confusion matrix values and coefficient correlation.

ML Algorithm	Results of the algorithms with the test data set					
	FP	FN	TP	TN	MCC	KCC
SVM	5	9	27	36	0.635	0.632
LR	5	8	27	37	0.659	0.660
DT	11	11	21	34	0.411	0.411
ANN	8	10	24	35	0.523	0.523

The Results of the different validation metrics implemented to know the performance of classification models are shown in Table 2, where the values of accuracy, AUC, F1-Score, sensitivity, and specificity are specified.

TABLE 2. Validation Results with AG Features Selection.

Validation Metrics	Artificial Intelligence Algorithms			
	SVM	LR	DT	ANN
Accuracy	0.81	0.83	0.71	0.76
AUC	0.82	0.83	0.70	0.76
F1-Score	0.83	0.85	0.75	0.80
Sensitivity	0.87	0.90	0.75	0.81
Specificity	0.75	0.75	0.65	0.70

As can be seen in the results of the previous table, we can notice that the specificity values are lower in all the models generated to detect cases of depression, this is due to a class balancing problem since there are more data from healthy study participants than with a certain degree of depression, but regardless of this phenomenon and considering the limited number of observations, the algorithms used and the small number of features used, the results are favorable.

On the other hand, the ROC curves that graphically show the relationship between the sensitivity and specificity of the classification models generated when applied to a set of test data. This next image gives us a picture of the extent to which the study participants (healthy and depressed) are classified correctly.

First, in Figure 5 shows the performance obtained by the SVM algorithm to classify the participants, the model was subjected to a detailed analysis through the construction of a Receiver Operating Characteristic (ROC) curve and several key results were obtained indicating its predictive ability. The AUC was calculated to be 0.822, with a 95% confidence interval between 0.722 and 0.921. This AUC value, which is closer to 1 than 0, suggests that the model has a robust ability to discriminate between classes, indicating promising performance.

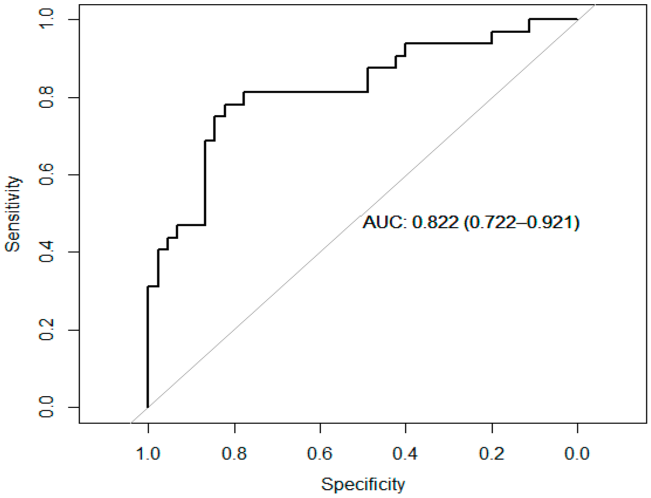


FIGURE 5. ROC curves of the SVM algorithm.

Secondly, A comprehensive analysis of the LR model was also carried out in the same classification context. Figure 6 illustrates the performance achieved by the LR algorithm in classifying the participants.

For the LR model, the AUC was calculated to be 0.838, with a 95 % confidence interval between 0.751 and 0.926. This AUC value, which is even closer to 1 than that obtained with the SVM model, reinforces the ability of the Logistic Regression model to effectively distinguish between the classes of interest. The high AUC is a strong indication of its predictive performance.

Comparing the results of both models, we observed that the Logistic Regression model obtained a slightly higher AUC (0.838) compared to the SVM model (0.822). This difference might suggest that the Logistic Regression model performs marginally better on this particular classification task. However, it is important to note that the choice between these models could depend on other factors, such as interpretability and simplicity of the model.

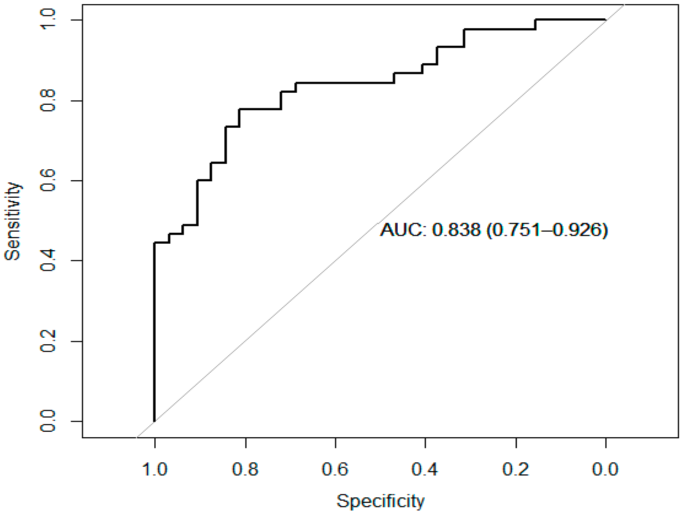


FIGURE 6. ROC curves of the LR algorithm.

Figure 7 depicts the performance attained through the DT algorithm for participant classification. Although the AUC is lower compared to the previous models (SVM and Logistic Regression), it is still in a range that suggests some discriminative ability. However, this AUC value indicates that the Decision Tree model may have more limited predictive performance in this particular task.

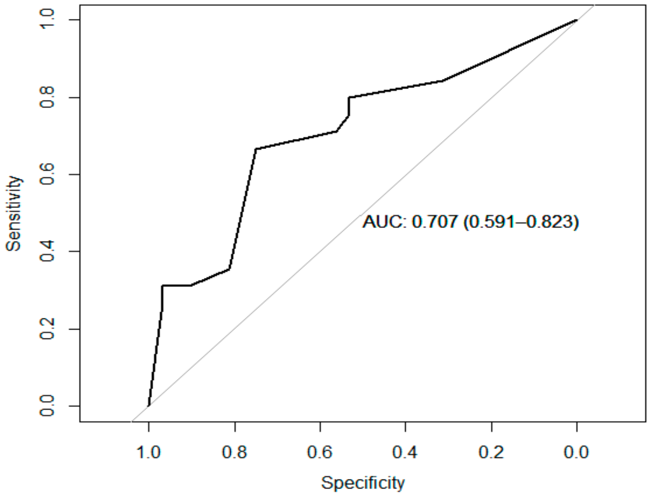


FIGURE 7. **ROC curves of the DT algorithm.**

In the framework of the research, the performance of an Artificial Neural Network (ANN) model as shown in Figure 8 was also evaluated in the classification task, along with SVM, Logistic Regression and Decision Tree models. When we compare the AUC of the ANN model (0.764) with the SVM (0.822) and Logistic Regression (0.838) models, we observe that ANN is in an intermediate position in terms of discrimination ability. Although it does not outperform the previous models in AUC, its performance is competitive and can be considered for applications where interpretability is not the main concern.

In summary, the results of the ROC curves of the models are highly promising. The high AUC, the balance between sensitivity and specificity, and the low false positive and false negative rate suggest that the model is an effective tool in the classification task studied. These results have important implications in the context of developing tools to support the diagnosis of depressive episodes, and could be valuable for clinical applications. However, it is important to consider the limitations of the model and future areas of research, such as the optimization of classification thresholds, to maximize its usefulness in the real world.

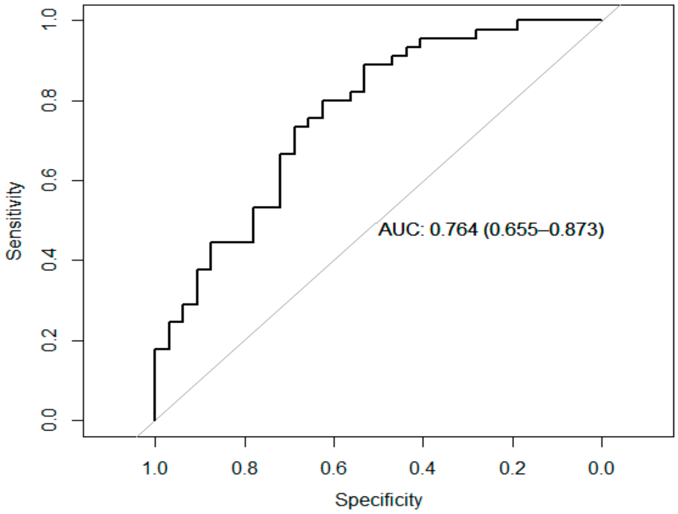


FIGURE 8. **ROC curves of the ANN algorithm.**

One of the assumptions that could not be omitted is where these algorithms were tested with all the features proposed in this research work; that is why we proceeded to a phase where the algorithms used the 1440 features to know their performance; the results are shown in Table 3 and Table 4.

TABLE 3. **Confusion matrix values and coefficient correlation with 1440 Features.**

ML Algorithm	Results of the algorithms with the test data set					
	FP	FN	TP	TN	MCC	KCC
SVM	9	16	23	29	0.358	0.352
LR	16	20	16	25	0.054	0.054
DT	10	18	22	27	0.283	0.277
ANN	10	15	22	30	0.349	0.346

TABLE 3. **Validation Results with 1440 Features**

Validation Metrics	Artificial Intelligence Algorithms			
	SVM	LR	DT	ANN
Accuracy	0.67	0.53	0.63	0.67
AUC	0.68	0.52	0.64	0.67
F1-Score	0.69	0.58	0.65	0.70
Sensitivity	0.76	0.60	0.72	0.75
Specificity	0.58	0.44	0.55	0.59

Results presented in the previous table are not as favorable with those presented in Table 2. This is an indicator of the importance of an intelligent feature selection phase that prevents the process of analysis and processing of the information to have better results, avoiding overfitting, reducing redundancy, and selecting the most significant data.

Although there are currently several investigations that have obtained very important results in the detection of depression, the simplicity of the process developed in this research is a very important aspect to consider since the results obtained through an algorithm such as logistic regression and reduced number of data contrast with the methodological approach used in state of the art. Table 5. shows the results obtained from the most recent research as well as the features used as source data related to the data set used in this work.

TABLE 5. **Comparative performance with state of the art.**

Author	Features	Technique	Acc.
Garcia-Ceja *et al.* [1]	Feature vector	SVM	0.72
Frogner *et al.* [15]	Feature Vector	1D-CNN	0.71
Jakobsen *et al.* [18]	Statistical Features	CNN	0.84
Kumar *et al.* [34]	Statistical Features	CNN	0.85
Rodríguez-Ruiz *et al.* [16]	Statistical Features	RFC	0.98
Ghate *et al.* [35]	Statistical Features	Transfer Learning	0.96
Zakariah *et al.* [36]	Statistical Features	Deep Neural Network	0.99

The table above reveals several noteworthy observations. Firstly, many previous studies employ considerably more intricate methodologies to enhance depression identification through motor activity. These approaches often involve extracting a multitude of statistical features, resulting in a substantial increase in the number of variables for analysis, processing, and the creation of classification models. This research aims to make a significant contribution by introducing a novel feature selection method using Genetic Algorithms (GA), which has not been employed with the Depression Dataset before. The primary objective is to reduce data volume without sacrificing critical information for identifying depressive states while also mitigating data redundancy.

Additionally, the achieved accuracy provides an initial step toward the development of more efficient models in terms of time and computational cost. Lastly, it's important to note that the obtained results hold statistical significance compared to the existing literature, despite the limited amount of processed data and the unique methodology proposed in this study

The proposal offers an innovative optimization approach for future work in developing algorithms, methodologies, and tools to aid in the detection of depression.

CONCLUSIONS

Artificial intelligence algorithms such as LR, supported by an objective feature selection method such as GA, allowed the efficient generation of a model capable of detecting depression with 83.0 % accuracy using only a few time intervals at different times of the day as a data source. The time intervals that showed significant information to generate models capable of detecting depression were: 15:43, 15:41, 12:11, 7:25, 15:40, 7:29, and 12:15 of a 24 Hrs. time system. Based on these findings, it can be inferred that the methodology introduced in this paper facilitates automatic and objective depression detection through various artificial intelligence approaches, achieving noteworthy accuracy with a set of seven features derived from patients' motor activity. As a result, this preliminary development of an assisted diagnostic tool emerges, offering potential assistance in mitigating the elevated error rates associated with diagnosing this condition. However, in this research, we assume that adding other variables such as type of diet, family history, sex, age, place of birth, and habits and implementing these algorithms in a real environment to test their efficiency and improve his learning could help in the future to improve the diagnosis of this mental illness.

As part of future work, there is a suggestion to augment the number of experimental observations to present more robust results with a greater diversity of data, including the increase of variables other than motor activity. In addition, it is proposed to implement deep learning like convolutional neural networks in smart devices capable of monitoring the objects of study in different circumstances or recurrent neural networks to detect depression early and allows a prevention strategy.

AUTHOR CONTRIBUTIONS

C.H.E.S conceptualized the project, designed and developed the methodology, contributed to the writing of the original manuscript, and participated in the programming of the software. C.E.G.T. participated in the data curation and gathering, carried out formal analyses and participated in the development of the software for the data analyses. A.G.S.R. contributed to the writing, editing and reviewing of the manuscript and participated in the data visualization for their correct interpretation. H.L.G. analyzed and validated the results. H.G.R. obtained funding and financial resources and oversaw the project. J.A.M.B provided access to material resources and equipment, participated in the data gathering and performed experiments. J.M.C.P. supervised and guided the general research and oversaw the project. J.I.G.T. conceptualized the project. All authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J. Tørresen, T. Nordgreen, K. J. Oedegaard, O. Bernt Fasmer, "Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients," in MMSys '18: Proceedings of the 9th ACM Multimedia Systems Conference, Amsterdam, Netherlands, 2018, pp. 472-477, doi: https://doi.org/10.1145/3204949.3208125

[2] S. Berenzon, M. A. Lara, R. Robles, and M. E. Medina-Mora, "Depresión: Estado del conocimiento y la necesidad de políticas públicas y planes de acción en méxico," Salud Publica Mex., vol. 55, no. 1, pp. 74-80, Jan.-Feb. 2013, doi: https://doi.org/10.1590/s0036-36342013000100011

[3] R. I. Shader, "COVID-19 and Depression," Clin. Ther., vol. 42, no. 6, pp. 962-963, Jun. 2020, doi: https://doi.org/10.1016/j.clinthera.2020.04.010

[4] S. Khairuddin, S. Ahmad, A. H. Embong, N. N. W. N. Hashim, T. M. K. Altamas, S. N. S. Badaruddin, S. S. Hassan, "Classification of the Correct Quranic Letters Pronunciation of Male and Female Reciters," IOP Conf. Ser.: Mater. Sci., vol. 260, art. no. 012004, 2017, doi: https://doi.org/10.1088/1757-899X/260/1/012004

[5] S. A. Montgomery and M. Asberg, "A new depression scale designed to be sensitive to change," Br. J. Psychiatry, vol. 134, no. 4, pp. 382-389, Apr. 1979, doi: https://doi.org/10.1192/bjp.134.4.382

[6] C. J. Hawley, T. M. Gale, and T. Sivakumaran, "Defining remission by cut off score on the MADRS: Selecting the optimal value," J. Affect. Disord., vol. 72, no. 2, pp. 177-184, Nov. 2002, doi: https://doi.org/10.1016/s0165-0327(01)00451-7

[7] M. J. Müller, H. Himmerich, B. Kienzle, and A. Szegedi, "Differentiating moderate and severe depression using the Montgomery-Åsberg depression rating scale (MADRS)," J. Affect. Disord., vol. 77, no. 3, pp. 255-260, Dec. 2003, doi: https://doi.org/10.1016/s0165-0327(02)00120-9

[8] F. J. Penedo and J. R. Dahn, "Exercise and well-being: A review of mental and physical health benefits associated with physical activity," Curr. Opin. Psychiatry, vol. 18, no. 2, pp. 189-193, Mar. 2005, doi: https://doi.org/10.1097/00001504-200503000-00013

[9] C. Bourguignon and K. F. Storch, "Control of rest: Activity by a dopaminergic ultradian oscillator and the circadian clock," Front. Neurol., vol. 8, art. no. 614, Nov. 2017, doi: https://doi.org/10.3389/fneur.2017.00614

[10] L. B. Alloy, T. H. Ng, M. K. Titone, and E. M. Boland, "Circadian Rhythm Dysregulation in Bipolar Spectrum Disorders," Curr. Psychiatry Rep., vol. 19, no. 4, art. no. 21, Apr. 2017, doi: https://doi.org/10.1007/s11920-017-0772-z

[11] J. O. Berle, E. R. Hauge, K. J. Oedegaard, F. Holsten, O. B. Fasmer, "Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression," BMC Res. Notes, vol. 3, art. no. 149, May 2010, doi: https://doi.org/10.1186%2F1756-0500-3-149

[12] E. Garcia-Ceja et al., "Motor Activity Based Classification of Depression in Unipolar and Bipolar Patients," in 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), Karlstad, Sweden, 2018, pp. 316-321, doi: https://doi.org/10.1109/CBMS.2018.00062

[13] L. A. Zanella-Calzada, C. E. Galván-Tejada, N. M. Cávez-Lamas, M. C. Gracia-Cortés, R. Magallanes-Quintanar, J. M. Celaya-Padilla, J. I. Galván-Tejada, H. Gamboa-Rosales, "Feature extraction in motor activity signal: Towards a depression episodes detection in unipolar and bipolar patients," Diagnostics, vol. 9, no, 1, art. no. 8, Jan. 2019, doi: https://doi.org/10.3390/diagnostics9010008

[14] C. E. Galván-Tejada, L. A. Zanella-Calzada, H. Gamboa-Rosales, J. I. Galván-Tejada, N. M. Chávez-Lamas, M. C. Gracia-Cortés, R. Magallanes-Quintanar, J. M. Celaya-Padilla, "Depression Episodes Detection in Unipolar and Bipolar Patients: A Methodology with Feature Extraction and Feature Selection with Genetic Algorithms Using Activity Motion Signal as Information Source," Mob. Inf. Syst., vol. 2019, art. no. 8269695, 2019, doi: https://doi.org/10.1155/2019/8269695

[15] J. I. Frogner, F. M. Noori, P. Halvorsen, S. A. Hicks, E. Garcia-Ceja, J. Torresen, M. A. Riegler, "One-dimensional convolutional neural networks on motor activity measurements in detection of depression," in HealthMedia '19: Proceedings of the 4th International Workshop on Multimedia for Personal Health & Health Care, Nice, Francia, 2019, pp. 9-15, doi: https://doi.org/10.1145/3347444.3356238

[16] J. G. Rodríguez-Ruiz, C. E. Galván-Tejada, L. A. Zanella-Calzada, J. M. Celaya-Padilla, et al., "Comparison of night, day and 24 h motor activity data for the classification of depressive episodes," Diagnostics, vol. 10, no. 3, art. no. 162, Mar. 2020, doi: https://doi.org/10.3390/diagnostics10030162

[17] J. G. Rodríguez-Ruiz, C. E. Galván-Tejada, H. Luna-García, H. Gamboa-Rosales, J. M. Celaya-Padilla, J. G. Arceo-Olague, J. I. Galván Tejada, "Classification of Depressive and Schizophrenic Episodes Using Night-Time Motor Activity Signal," Healthcare, vol. 10, no. 7, art. no. 1256, Jul. 2022, doi: https://doi.org/10.3390/healthcare10071256

[18] P. Jakobsen, E. Garcia-Ceja, M. Riegler, L. A. Stabell, T. Nordgreen, J. Torresen, O. B. Fasmer, K. J. Oedegaard, "Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls," PLoS One, vol. 15, no. 8, art. no. e0231995, Ago. 2020, doi: https://doi.org/10.1371/journal.pone.0231995

[19] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," Multimed. Tools Appl., vol. 81, no. 17, pp. 23649-23685, 2022, doi: https://doi.org/10.1007/s11042-022-12648-y

[20] S. L. Pacheco-González, L. A. Zanella-Calzada, C. E. Galván-Tejada, N. M. Chávez-Lamas, J. F. Rivera-Gómez, and J. I. Galván-Tejada, "Evaluation of Five Classifiers for Depression Episodes Detection," Res. Comput. Sci., vol. 148, no. 10, pp. 129-138, 2019, doi: https://doi.org/10.13053/rcs-148-10-11

[21] M. Raihan, A. K. Bairagi, and S. Rahman, "A Machine Learning Based Study to Predict Depression with Monitoring Actigraph Watch Data," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-5, doi: https://doi.org/10.1109/ICCCNT51525.2021.9579614

[22] P. M. Singh and P. S. Sathidevi, "Design and Implementation of a Machine Learning-Based Technique to Detect Unipolar and Bipolar Depression Using Motor Activity Data," in Smart Trends in Computing and Communications. Lecture Notes in Networks and Systems, Nevada, USA, 2022, pp. 99-107, doi: https://doi.org/10.1007/978-981-16-4016-2_10

[23] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," Multimed. Tools Appl., vol. 80, no. 5, pp. 8091-8126, 2021, doi: https://doi.org/10.1007/s11042-020-10139-6

[24] V. Trevino and F. Falciani, "GALGO: An R package for multivariate variable selection using genetic algorithms," Bioinformatics, vol. 22, no. 9, pp. 1154-1156, May 2006, doi: https://doi.org/10.1093/bioinformatics/btl074

[25] T. Jayalakshmi and A. Santhakumaran, "Statistical Normalization and Back Propagationfor Classification," Int. J. Comput. Theory Eng., vol. 3, no. 1, pp. 89-93, 2011, doi: https://doi.org/10.7763/IJCTE.2011.V3.288

[26] A. A. AlBeladi and A. H. Muqaibel, "Evaluating compressive sensing algorithms in through-the-wall radar via F1-score," Int. J. Signal Imaging Syst. Eng., vol. 11, no. 3, pp. 164-171, 2018, doi: https://doi.org/10.1504/IJSISE.2018.093268

[27] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet, vol. 1, no. 8476, pp. 307-310, Feb. 1986, doi: https://doi.org/10.1016/S0140-6736%2886%2990837-8

[28] N. Banaei, J. Moshfegh, A. Mohseni-Kabir, J. M. Houghton, Y. Sun, and B. Kim, "Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips," RSC Adv., vol. 9, no. 4, pp. 1859-1868, 2019, doi: https://doi.org/10.1039/C8RA08930B

[29] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognit., vol. 30, no. 7, pp. 1145-1159, Jul.1997, doi: https://doi.org/10.1016/S0031-3203(96)00142-2

[30] C. H. Espino-Salinas, C. E. Galván-Tejada, H. Luna-García, H. Gamboa-Rosales, J. M. Celaya-Padilla, L. A. Zanella-Calzada, and J. I. Galván-Tejada, "Two-Dimensional Convolutional Neural Network for Depression Episodes Detection in Real Time Using Motor Activity Time Series of Depresjon Dataset," Bioengineering, vol. 9, no. 9, art. no. 458, Sep. 2022, doi: https://doi.org/10.3390/bioengineering9090458

[31] O. Caelen, "A Bayesian interpretation of the confusion matrix," Ann. Math. Artif. Intell., vol. 81, pp. 429-450, Sep. 2017, doi: https://doi.org/10.1007/s10472-017-9564-8

[32] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, art. no. 6, Jan. 2020, doi: https://doi.org/10.1186/s12864-019-6413-7

[33] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment," IEEE Access, vol. 9, pp. 78368-78381, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3084050

[34] A. Kumar, S. R. Sangwan, A. Arora, and V. G. Menon, "Depress-DCNF: A deep convolutional neuro-fuzzy model for detection of depression episodes using IoMT," Appl. Soft Comput., vol. 122, art. no. 108863, 2022, doi: https://doi.org/10.1016/j.asoc.2022.108863

[35] R. Ghate, N. Kalnad, R. Walambe, and K. Kotecha, "Transfer Learning for Real-time Deployment of a Screening Tool for Depression Detection Using Actigraphy," in UKSim-AMSS 25th International Conference on Modelling & Simulation, Cambridge, United Kingdom, 2023, pp. 1-5, doi: https://doi.org/10.48550/arXiv.2303.07847

[36] M. Zakariah and Y. A. Alotaibi, "Unipolar and Bipolar Depression Detection and Classification Based on Actigraphic Registration of Motor Activity Using Machine Learning and Uniform Manifold Approximation and Projection Methods," Diagnostics, vol. 13, no. 14, art. no. 2323, Jul. 2023, doi: https://doi.org/10.3390/diagnostics13142323

ARTÍCULO DE INVESTIGACIÓN

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 53 - 69

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.4

E-LOCATION ID: **1378**

Sistema de Cribado Primario para la Sarcopenia en Personas Adultas Mayores Basado en Inteligencia Artificial

Primary Screening System for Sarcopenia in Elderly People Based on Artificial Intelligence

Santiago Arceo-Diaz[1], *Elena Elsa Bricio-Barrios*[2], *Xochitl Angélica Rosio Trujillo-Trujillo*[1],
José Ramón González-Farias[1], *Jaime Alberto Bricio-Barrios*[1], *Mónica Rios-Silva*[1], *Miguel Huerta-Viera*[1]

[1]Universidad de Colima, Colima - México

[2]Tecnológico Nacional de México campus Colima, Colima - México

RESUMEN

Este estudio propone un sistema de cribado primario para diagnosticar la sarcopenia en adultos mayores a través de medidas antropométricas. Esta investigación exploratoria involucró inicialmente a 150 personas de edad avanzada, de las cuales 122 fueron seleccionadas después de un proceso de depuración de datos. Empleando técnicas de aprendizaje automático como el agrupamiento jerárquico y los árboles de decisión, se redujeron las 13 medidas antropométricas originales a cinco características clave. Se crearon tres sistemas de clasificación: el primero basado en parámetros previamente establecidos (masa muscular apendicular, velocidad de marcha y fuerza de agarre); el segundo consideró medidas de las extremidades superiores (masa muscular promedio de ambos brazos, fuerza de agarre, velocidad de marcha y porcentaje de grasa corporal); y el tercero se centró en las medidas de las extremidades inferiores (masa muscular promedio de ambas piernas, fuerza de agarre, velocidad de marcha y porcentaje de grasa corporal). Estos sistemas de clasificación se validaron clínicamente en un grupo de 57 pacientes previamente diagnosticados por especialistas, de los cuales 10 recibieron un diagnóstico positivo de sarcopenia. Los resultados mostraron eficiencias similares en los tres sistemas, con ocho de los diez diagnósticos positivos conocidos clasificados en el mismo grupo. Además, el estudio proporciona puntos de corte específicos para cada sistema, facilitando así el diagnóstico clínico de la sarcopenia por parte de profesionales médicos.

PALABRAS CLAVE: clustering jerárquico, medidas antropométricas, personas adultas mayores, Random Forest, sarcopenia

ABSTRACT

This study proposes a primary screening system for diagnosing sarcopenia in older adults through anthropometric measures. This exploratory research initially involved 150 elderly individuals, of whom 122 were selected after a data purification process. Using machine learning techniques such as hierarchical clustering and decision trees, the original set of 13 anthropometric measures was reduced to five key features. Three classification systems were created: the first based on previously established parameters (appendicular muscle mass, walking speed, and grip strength); the second considered upper limb measures (average muscle mass of both arms, grip strength, walking speed, and body fat percentage); and the third focused on lower limb measures (average muscle mass of both legs, grip strength, walking speed, and body fat percentage). These classification systems were clinically validated in a group of 57 patients previously diagnosed by specialists, of which 10 received a positive sarcopenia diagnosis. The results showed similar efficiencies in all three systems, with eight of the ten known positive diagnoses classified in the same group. Additionally, the study provides specific cut-off points for each system, thus facilitating the clinical diagnosis of sarcopenia by medical professionals.

KEYWORDS: anthropometric measures, elderly people, hierarchical clustering, Random Forest, sarcopenia

Autor de correspondencia

DESTINATARIO: **Xochitl Angélica Rosio Trujillo-Trujillo**
INSTITUCIÓN: **Universidad de Colima**
DOMICILIO: **Centro Universitario de Investigaciones Biomédicas. Avenida 25 de Julio # 965, Colonia Las Víboras, Colima, Col. México**
CORREO ELECTRÓNICO:: **rosio@ucol.mx**

Recibido:

01 Octubre 2023

Aceptado:

22 Noviembre 2023

INTRODUCCIÓN

En la actualidad, las mejoras en la atención de los sistemas de salud, los avances médicos y los programas de planificación familiar han contribuido a incrementar la expectativa de vida en países emergentes como México [1]. Sin embargo, al prolongarse la esperanza de vida, se ha vuelto cada vez más frecuente que los sistemas de salud brinden atención a pacientes que sufren de condiciones asociadas con el envejecimiento natural, las cuales pueden tener un impacto negativo importante en sus vidas [2].

TABLA 1. **Diagnóstico primario de sarcopenia.**

Grupo de trabajo	Masa muscular	Fuerza muscular	Rendimiento físico
Grupo Europeo de Trabajo sobre Sarcopenia en Personas Mayores	MMA: < 20 en hombres y < 15 en mujeres	FPP: ≤ 27 en hombres y < 16 en mujeres	VC < 0.8 en hombres y mujeres
Grupo Asático de Trabajo sobre Sarcopenia en Personas Mayores	IMMA: ≤ 7 en hombres y ≤ 5.4 en mujeres	FPP: < 26 en hombres y < 18 en mujeres	VC ≤ 0.8 en hombres y mujeres
Fundación para el Instituto Nacional de Salud	MA/ IMC: <0.789 en hombres y <0.512 en mujeres	FPP: < 26 en hombres y <16 en mujeres	VC ≤ 0.8 en hombres y mujeres
Grupo con interés especial	≤ 2 desviaciones estándar del promedio del porcentaje de masa muscular	No se considera	VC ≤ 0.8 en hombres y mujeres
Grupo internacional de sarcopenia	IMME < 7.23 en hombres y <5.67 en mujeres	No se considera	VC < 1.0
Sociedad de Sarcopenia y Caquexia	≤ 2 desviaciones estándar del promedio de MA/altura evaluada	No se considera	TC6M < 400

La sarcopenia es una enfermedad progresiva y generalizada del músculo esquelético, caracterizada por una disminución de la fuerza, de la masa muscular y, finalmente, del rendimiento físico y se asocia al envejecimiento [3]. Además, en personas adultas mayores, la sarcopenia está relacionada con la degradación proteica de los músculos, procesos inflamatorios, alteraciones hormonales y comorbilidades [4]. La disminución de la masa muscular suele tener como consecuencia una reducción en la capacidad de las personas para desarrollar actividades de la vida cotidiana, aspectos como la velocidad de marcha o la capacidad para sujetar objetos que dependan de la fuerza de agarre [5]. De acuerdo con el grado de deterioro, la sarcopenia puede clasificarse como pre-sarcopenia, sarcopenia y sarcopenia severa que, en su estadio más severo, las afectaciones producidas por la sarcopenia pueden afectar la vida de las personas o inducir caídas o fracturas, que pueden llevar al fallecimiento prematuro [6].

En años recientes, diversos avances tecnológicos han facilitado el diagnóstico de la sarcopenia. Esta capacidad de diagnóstico se fundamenta en la detección de cambios en la composición corporal mediante técnicas como la medición de la bioimpedancia bioeléctrica, la resonancia magnética o las tomografías computarizadas [7]. Sin embargo, en muchos casos, el costo relativo de los equipos de medición es demasiado alto para la mayoría de los hospitales o clínicas.

Esta problemática ha sido abordada por diversos grupos de trabajo que han identificado características y valores de referencia para el diagnóstico primario de sarcopenia. La Tabla 1 muestra los principales estudios que recopiló Aquiles *et al.* en [8].

En la segunda, tercera y cuarta columna reporta las medidas antropométricas y puntos de corte como: FPP: Fuerza de presión palmar en kg; IMMA: Índice de masa muscular apendicular en kg*m^{-2}; MA/IMC: corresponde a Masa apendicular sobre Índice de masa corporal, adimensional, MMA: Masa muscular apendicular en kg*m^{-2};

TC6M: Test de caminata de 6 minutos; VC: velocidad de caminata en $m*s^{-1}$.

Si bien, la medición de parámetros antropométricos involucra un costo inferior a las técnicas computacionales mencionadas, los puntos de corte de los parámetros antropométricos (los valores críticos que definen los umbrales en los que puede considerarse que una medición es anómala) difieren entre estudios debido a las características de las muestras evaluadas [9]. Esta variabilidad limita el uso de estos puntos de corte a muestras en las que las etnias de los participantes coincidan con la de las personas que se usaron para obtenerlos. Además, dado que la sarcopenia es una enfermedad multifactorial, su diagnóstico puede confundirse con síndromes metabólicos, neuromusculares y reumáticos [10]. En México, Espinel en [11] emplearon una base de datos conformada por 5046 registros de adultos mayores en el 2012 e identificaron, sin puntos de corte, que las mujeres son más propensas de adquirir sarcopenia y este riesgo incrementa con la edad, caídas previas, deterioro cognitivo, obesidad abdominal y alta marginación.

En este contexto, la Inteligencia Artificial (IA) ha emergido como una herramienta para el diseño de sistemas que emulan la inteligencia humana que ayude en la toma de decisiones a los especialistas de la salud, así como la creación de modelos predictivos de diagnósticos clínicos [12]. Dentro de las aplicaciones en el área de ingeniería médica y biomédica, los árboles de decisión permiten crear modelos predictivos a partir de registros históricos de pacientes que previamente fueron diagnosticados con padecimientos o enfermedades [13] como apendicitis aguda, cáncer de mama, hepatopatía crónica [14]. Sin embargo, esta técnica es susceptible a la cantidad de variables evaluada y la variabilidad de cada registro.

Esta problemática fue abordada por Breiman en [15], que propuso la creación y combinación de los mejores árboles de decisión que permiten incrementar la precisión del modelo predictivo, esta técnica es conocida como *Random Forest*.

En el área de geriatría, *Random Forest* ha sido aplicado en beneficio de las personas adultas mayores para identificar la sintomatología y puntos de cohorte de personas con sarcopenia a través de biomarcadores moleculares [16], parámetros bioquímicos [17], daño muscular [18].

Si bien la IA es capaz procesar grandes volúmenes de información y desempeña un papel fundamental en la investigación médica, a menudo nos enfrentamos a un dilema ético y práctico. En muchos casos, los participantes pueden sentirse utilizados o preocupados por la invasión de su privacidad al compartir un gran número de datos personales o médicos y, en consecuencia, se niegan a participar en estos estudios [19].

Dada la importancia de ofrecer calidad de vida a las personas adultas mayores, este trabajo reporta una metodología basada en la depuración de una base de datos, la clasificación y la construcción de un modelo predictivo basado en registros históricos, para el diagnóstico primario de sarcopenia empleando mediciones antropométricas, así como su validación respecto a estrategias reportadas en la literatura abierta. A diferencia de otras propuestas, este trabajo evalúa variables ajenas a la medición en personas adultas mayores que carecen de alguna o algunas extremidades o se encuentran postradas. Asimismo, se proponen los puntos de corte en una muestra de la población de personas adultas mayores del estado de Colima, México.

MATERIALES Y MÉTODOS

Este trabajo se realizó en tres etapas: i) la identificación de los participantes y la recolección de variables antropométricas, ii) la depuración de los datos a partir de herramientas estadísticas y iii) la categorización de los participantes en función de las variables antropométricas. Los detalles de cada etapa se mencionan a continuación:

Identificación y recolección de datos

Durante el 2021, como parte de un estudio longitudinal, profesorado del área de ciencias de la salud del estado de Colima capacitó a un grupo de estudiantes de ambos géneros adscritos a la Facultad de Medicina de la Universidad de Colima, quienes acudieron al Centro de Asistencia Social de Día para Personas Adultas Mayores Zona Oriente del DIF Estatal Colima ubicado en la zona metropolitana Colima-Villa de Álvarez, donde invitaron a personas adultas mayores a participar en el estudio. Dicho estudio tuvo como objetivo comprender el envejecimiento patológico en la población adulta mayor, así como el impacto que tuvo el cese de actividades recreativas tras el periodo de contingencia por la COVID-19; dicha propuesta contó con la aprobación del Comité de Ética en Investigación de una institución de salud que cuenta con registro ante la Comisión Nacional de Bioética (CEICANCL08012021-MARCOGP-01). Los datos generados de dicha investigación se encuentran en un repositorio, el cual puede ser compartido a través de un convenio de colaboración con el grupo de investigación de la Universidad de Colima.

Para ofrecer una propuesta que cumpla con los principios éticos y evitar sesgos en los resultados, se realizó un filtro primario de personas adultas mayores que cumplieran características como edad igual o superior a 60 años cumplidos. En cambio, no se tomaron como participantes aquellos que se desplazaban en sillas de ruedas o tenían prótesis en brazos o piernas.

Con base en lo anterior, 150 participantes cumplieron con los requisitos y firmaron un acuerdo de consentimiento informado que indicaba el objetivo del proyecto, las actividades que se comprometían a realizar, así como los riesgos de su participación. Asimismo, para asegurar la ética del proyecto, la propuesta se sometió bajo los principios éticos de la Declaración de Helsinki y ante la Comisión Nacional de Bioética de México [20], en la que los autores se comprometieron a mantener la confidencialidad y privacidad de los datos, así como la equidad e imparcialidad y la transparencia en la selección de los participantes para evitar sesgos en el análisis y presentación de resultados. Cada persona adulta mayor que cumplió con los requisitos y manifestó su interés de manera voluntaria fue incluida en el estudio.

TABLA 2. **Mediciones registradas para la base de datos.**

Variable	Unidad de medida	Equipo de medición
Sexo	variable categórica	Entrevista
Edad	años	Entrevista
Estatura	m	Estadímetro
Peso medido	kg	Balanza de bioimpedancia
Grasa corporal total	%	Monitor de composición corporal
Masa muscular total (MUS)	kg	Monitor de composición corporal
Masa muscular en pierna izquierda y derecha (MPI y MPD)	kg	Monitor de composición corporal
Masa muscular en brazo izquierdo y derecho (MBI y MBD)	kg	Monitor de composición corporal
Masa muscular en zona central (MC)	kg	Monitor de composición corporal
Circunferencia media de brazo (CMB)	cm	Cinta métrica
Circunferencia media de pantorrilla (CMP)	cm	Cinta métrica
Fuerza de agarre en brazo izquierdo y brazo derecho (FIZ y FDER)	kg fuerza	Dinamómetro
Velocidad de Marcha (Marcha)	$m*s^{-1}$	Cinta métrica Cronómetro

A cada persona adulta mayor que cumplió con los requerimientos se le asignó un número de folio para garantizar la confidencialidad de sus datos y como primera actividad, el personal de apoyo solicitó a los participantes que proporcionaron sus datos demográficos, medidas antropométricas y pruebas de función muscular. La Tabla 2 muestra las variables registradas, unidades de medida y equipo empleado en este estudio.

Las variables de la Tabla 2 se registraron en una hoja de cálculo de Excel. Además de estas variables se calcularon también los índices de masa corporal (a partir del peso total y la altura), de masa muscular en la zona central (definido como la masa muscular del tronco dividido entre la altura al cuadrado) y de masa muscular apendicular (a partir de la suma de masa muscular brazos y piernas dividida entre la altura al cuadrado), todos en unidades de kg*m^{-2}.

Depuración de datos

En esta etapa, se implementó la metodología propuesta por Osborne en [21] para la depuración de datos. Esta metodología destaca la importancia de seguir ciertos pasos para asegurar la calidad de los datos antes de su análisis. El procesamiento de la información se llevó a cabo en *Google Colab*, una plataforma gratuita en línea que permite ejecutar código en la nube.

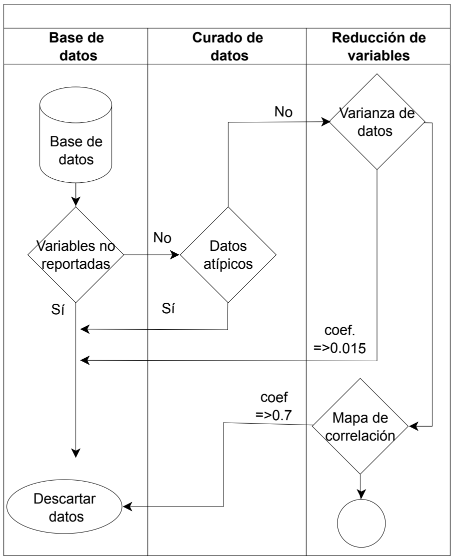


FIGURA 1. Diagrama de flujo de la depuración de datos.

La Figura 1 esclarece el proceso de depuración de datos, que se inició con la identificación y exclusión de participantes cuyos registros carecían de información completa de las variables detalladas en la Tabla 2. Inicialmente, la base de datos original se redujo a 144 participantes una vez que se eliminaron los registros incompletos. Posteriormente, se llevó a cabo la identificación de datos atípicos empleando diagramas de caja y bigotes, herramientas que facilitan la visualización de la distribución de los datos y permiten la detección de valores que se hallan fuera del rango esperado [22]. Aquellos registros que presentaron valores atípicos fueron excluidos, lo que resultó en una base de datos depurada con 122 pacientes (con una edad promedio de 71.8 años y una desviación estándar de 6 años). El personal de salud en atención primaria evaluó los resultados de las pruebas clínicas en un total de 64 pacientes, aunque no fue posible completar el diagnóstico para toda la muestra debido a limitaciones logísticas. La mayoría de los pacientes, que formaban un grupo de 52 individuos de ambos géneros, recibieron un diagnóstico negativo, en contraste con un grupo más reducido de 12 pacientes, en su mayoría mujeres, que obtuvieron un diagnóstico positivo. Para asegurar la integridad de los análisis, se excluyeron a los pacientes con datos faltantes o valores atípicos. Esto resultó en dos grupos finales: 47 pacientes con diagnóstico negativo (compuestos por 33 mujeres y 14 hombres) y 10 pacientes con diagnóstico positivo confirmado (9 mujeres y 1 hombre).

Una vez depurada la base de datos, se procedió a reducir las mediciones registradas para el diagnóstico de sarcopenia, siguiendo la metodología de “reducción de dimensionalidad” [23]. Este proceso, se llevó a cabo en dos fases. En la primera fase se analizó la varianza de cada una de las variables cuantitativas. De acuerdo con la reducción de dimensionalidad, se considera que las variables con las varianzas normalizadas más pequeñas son menos eficientes en la clasificación de pacientes a través de sus diferencias. En el caso de la muestra, se observó que la variable cuantitativa de “Edad” tiene una varianza normalizada mucho menor al resto, por lo que fue eliminada de la lista original de variables. Este resultado es consistente por el reportado por Lera en [24] quienes identificaron mayor prevalencia de sarcopenia en personas adultas mayores a 80 años que radican en Chile.

En la segunda fase, se evaluó la correlación entre las variables de la Tabla 2 utilizando el coeficiente de correlación de Spearman. A partir del mapa de correlación resultante, se eliminaron aquellas variables que tuvieran un índice de correlación mayor 0.7, siguiendo la propuesta de la eficiencia de un proceso de clasificación no se ve comprometido si se eliminan variables que estén altamente correlacionadas con otras [25]. El resto de las variables fueron consideradas como la lista de variables con las que se continua en el proceso de clasificación de la base de datos depurada.

Clasificación de datos y modelo predictivo

Una vez concluido el proceso de depuración de datos, se procedió a la clasificación de estos y a la construcción del modelo predictivo. En particular, los modelos predictivos construidos por IA, han reportado alto desempeño en el análisis de imágenes en el diagnóstico de diferentes tipos de cáncer y prognosis [26], reducen sesgos y subjetividad al evaluar al paciente [22], capacidad de procesamiento de información para la construcción de modelos predictivos cualitativos y cuantitativos basados en registros históricos de pacientes que comparten la misma sintomatología de una enfermedad [23], entre otros.

Como actividad previa a la implementación de la IA como herramienta para apoyar al especialista de la salud en la toma de decisiones se recomienda desarrollar la metodología propuesta por Azar y *El-Metwally* en [27]. La descripción general se reporta en la Figura 2 y los detalles de cada elemento se mencionan a continuación:

La Figura 2, en su primera columna titulada "Agrupamiento de Pacientes", muestra cómo se agruparon a los participantes con características similares. Usamos una técnica llamada clustering jerárquico aglomerativo [28]. El clustering jerárquico aglomerativo es utilizado para agrupar objetos o puntos de datos de manera jerárquica. Inicialmente, cada punto se considera un clúster independiente y, mediante el cálculo de la distancia entre este punto y el resto, todos aquellos puntos que estén a una distancia menor se incluyen como parte del mismo clúster en la siguiente iteración. El clustering jerárquico aglomerativo es utilizado para agrupar objetos o puntos de datos de manera jerárquica. Inicialmente, cada punto se considera un clúster independiente y, mediante el cálculo de la distancia entre este punto y el resto, todos aquellos puntos que estén a una distancia menor se incluyen como parte del mismo clúster en la siguiente iteración. Dos parámetros que deben definirse para aplicar este algoritmo son la distancia y el tipo de enlace. En este contexto se optó por usar la "distancia euclidiana" y el "enlace de Ward". Esta combinación de distancia y métrica es efectiva para formar clústeres densos y cohesivos, lo que facilita la identificación de grupos bien definidos en aplicaciones como el estudio de sarcopenia. Esta técnica se ha empleado en el ámbito médico para diagnosticar enfermedades y comorbilidades en adultos mayores [29]. Tanto el algoritmo de clustering jerárquico como el tipo de distancia y enlace para su creación están contenidos dentro de la librería de Python *"Scikit-learn"* [30]. Esta librería es una herramienta gratuita y abierta para el aprendizaje automático que incluye funciones para tareas como clasificación, regresión, clustering y reducción de dimensionalidad [31].

Para determinar el número ideal de clústeres, se emplearon las técnicas del método del codo y el análisis de inercia, tal como se detallan en el trabajo de Castañeda en [32]. Además, se llevó a cabo un análisis de hipótesis utilizando pruebas *t-Student* [33] para asegurar que no hubiera clústeres redundantes, es decir, aquellos que no pudieran distinguirse significativamente del resto en función de los valores promedio de los parámetros utilizados en la clasificación de los pacientes. Además, se utilizaron las métricas de "Silhouette Score", *"Índice Davies-Bouldin"* y *"Calinski-Harabasz"* para evaluar el desempeño de los algoritmos de agrupamiento. Respectivamente, se encontraron los valores 0.23, 1.12 y 58.4. Estos valores sugieren que los grupos están relativamente bien separados, en relación con la dispersión interna de los datos.

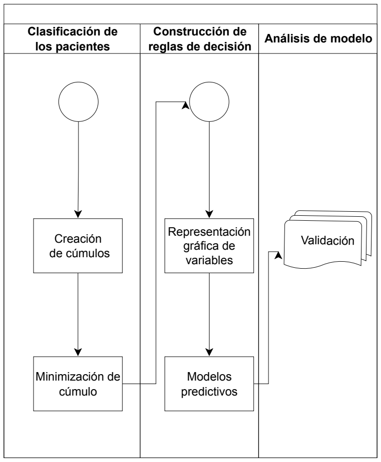


FIGURA 2. Diagrama de flujo de la clasificación de datos.

En la segunda columna del diagrama de flujo de la Figura 2, se muestra el proceso de construcción de las reglas de decisión. Para abordar esta problemática, se utilizó la técnica del *Random Forest*, que permite la creación de múltiples árboles de decisión con el propósito de mejorar la precisión del modelo predictivo. En este estudio, se optó por generar 100 árboles de decisión y seleccionar aquellos con un mejor desempeño para la creación del árbol de decisión final. Este procedimiento conlleva la identificación del árbol individual del *Random Forest* que se considera el óptimo en función de la importancia de las características. Dicha selección se fundamenta en la relevancia de las características del modelo y permite utilizar un árbol específico del conjunto con el fin de comprender con mayor profundidad el proceso de toma de decisiones o llevar a cabo predicciones basadas en ese árbol particular. Este árbol final proporciona un conjunto de reglas, un modelo predictivo, que permiten el diagnóstico de la sarcopenia a partir de la clasificación de los pacientes de acuerdo con los puntos de corte calculados para los parámetros considerados pertinentes [34].

Para ello, se empleó el método de análisis de componentes principales, que permite reducir las variables a sólo dos dimensiones y visualizar gráficamente las agrupaciones de datos que comparten características semejantes. Este método fue propuesto por Bakator y Radosav en [35], pero tiene la limitación de estar restringido a magnitudes cualitativas difusas, como "alto" o "bajo", lo que puede tornar subjetivo el diagnóstico de un paciente.

Luego de obtener un modelo de clasificación basado en *Random Forest*, los datos de pacientes con diagnóstico confirmado por personal médico se utilizaron para validar las reglas de decisión generadas por el árbol final. Para evaluar la coherencia tanto de los agrupamientos como de las reglas de decisión obtenidas, se estableció como criterio esencial que un modelo adecuado debía ser capaz de agrupar a los pacientes con diagnósticos confirmados en grupos específicos sin superposición. En otras palabras, las reglas de agrupamiento propuestas debían permitir la formación de un grupo para los diagnósticos positivos y otro para los negativos sin que se produjeran solapamientos entre ellos. Además, se consideró deseable que se pudieran agrupar todos los pacientes con un diagnóstico compartido en un grupo específico.

Dado que cada árbol final involucra la consideración de múltiples variables para representar a los pacientes, se aplicaron técnicas secuenciales de reducción de dimensionalidad con el propósito de facilitar la identificación de posibles patrones. Se optó por la aplicación de dos técnicas disponibles dentro de la librería *Sci-kit-learn* (t-SNE y PCA).

Inicialmente, se implementó t-SNE *(t-Distributed Stochastic Neighbor Embedding)* para disminuir la complejidad del conjunto de datos y mejorar la identificación de patrones. Está técnica de reducción de dimensionalidad es utilizada principalmente para visualizar datos complejos de alta dimensión de una manera más sencilla y comprensible. Su propósito principal es ayudar a identificar patrones, estructuras y agrupaciones en datos que son difíciles de apreciar en su forma original. t-SNE asigna a cada punto de datos en un espacio de alta dimensión un punto en un espacio de baja dimensión de tal manera que los puntos similares en el espacio de alta dimensión se representan como puntos cercanos en el espacio de baja dimensión.

A continuación, se llevó a cabo el análisis de componentes principales (PCA) para lograr una reducción adicional de la dimensionalidad [36]. El Análisis de Componentes Principales es una técnica que, en esencia, busca transformar un conjunto de variables originales en un nuevo conjunto de variables llamadas "componentes principales". Cada componente principal es una combinación lineal de las variables originales y están ordenadas en función de su capacidad para explicar la variabilidad en los datos. Esta técnica simplifica los datos al resaltar las direcciones en las que los datos varían más y facilita la identificación de las variables originales que contribuyen significativamente a esa variabilidad.

Finalmente, se procedió a la aplicación de t-SNE una vez más en el espacio de menor dimensión resultante con el propósito de perfeccionar la visualización y destacar con precisión los patrones y agrupamientos de interés, así como de identificar clusters específicos o agrupamientos relevantes en los datos. Este enfoque se centra en preservar la estructura local de los datos, lo que facilita la identificación de clusters y patrones específicos.

La última etapa del proceso se muestra en la tercera columna del diagrama de flujo en la Figura 2. En esta etapa, se realizó el análisis del modelo predictivo utilizando la base de datos depurada. Se identificaron a los pacientes con diagnóstico confirmado de sarcopenia y se compararon sus registros con las reglas de decisión creadas por el modelo predictivo. De esta manera, se pudo evaluar si el modelo es capaz de describir el diagnóstico clínico realizado por bioimpedancia eléctrica utilizando medidas antropométricas.

A partir de la lista depurada de parámetros de interés, se formularon tres sistemas de diagnóstico, cada uno basado en combinaciones específicas de parámetros para los cuales se determinaron puntos de corte que se utilizaron para generar un diagnóstico de sarcopenia. En el primer sistema, se incluyó la combinación del Índice de Masa Muscular Apendicular, la velocidad de marcha y la fuerza de agarre. Esta combinación se propuso con el propósito de permitir la comparación de puntos de corte y reglas previamente reportados en investigaciones anteriores [36]. El segundo sistema propone una combinación que incluye la masa muscular promedio de ambos brazos, la fuerza de agarre, la velocidad de marcha y el porcentaje de grasa corporal. Del mismo modo, el tercer sistema se basa en la combinación de la masa muscular promedio de ambas extremidades inferiores, la fuerza de agarre, la velocidad de marcha y el porcentaje de grasa corporal. La lógica subyacente en la elección de estas variables radica en la evaluación de la viabilidad de un sistema de diagnóstico que requiera únicamente la medición de las extremidades superiores o inferiores, en lugar de abordar las cuatro extremidades. Además, se incorporó el porcentaje de grasa corporal con el fin de determinar la capacidad de distinguir casos de sarcopenia primaria de aquellos que puedan estar relacionados con la obesidad sarcopénica. Cada sistema de diagnóstico propuesto fue evaluado en términos de la consistencia de los puntos de corte que sugieren y del grado de coincidencia con los diagnósticos hechos por el personal médico de atención primaria. Además, los tres sistemas fueron integrados en una plataforma de libre acceso que permite el análisis y diagnóstico desde un teléfono celular.

RESULTADOS Y DISCUSIÓN

Sistema: ASMI + FA + Marcha

El primer sistema de clasificación se basa en la evaluación de la fuerza de agarre (FA), la velocidad de marcha y el índice de masa muscular apendicular (ASMI), una combinación de parámetros propuesta por Yuki *et al.* en [36] para detectar sarcopenia en personas adultas mayores que radican en la zona metropolitana de Japón. Es importante mencionar que, debido a posibles diferencias étnicas, los puntos de corte sugeridos por dicho estudio no son directamente aplicables a la población mexicana.

Los diagramas de cajas se utilizan para la comparación de los clústeres generados en base a variables relevantes en el proceso de clasificación. Las Figuras 3, 4 y 5 comparan directamente la distribución de valores para cada parámetro de interés en cada una de las agrupaciones obtenidas por el algoritmo de clustering. Las cinturas de las cajas sirven como una guía visual que permite evaluar si existen diferencias significativas entre las medianas de cada clúster con relación al parámetro de interés. Cuando las cinturas están separadas y no se superponen, indican una clara diferencia entre los clústeres.

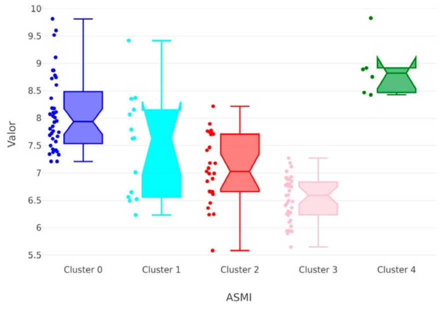


FIGURA 3. Diagrama de cajas para los valores de ASMI en los clústeres del sistema ASMI + FA + Marcha.

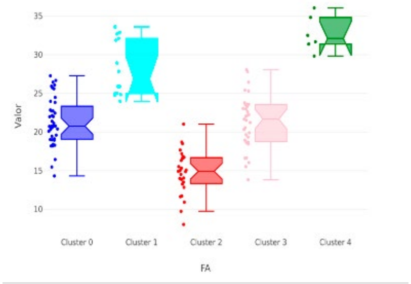


FIGURA 4. Diagrama de cajas para los valores de FA en los clústeres del sistema ASMI + FA + Marcha.

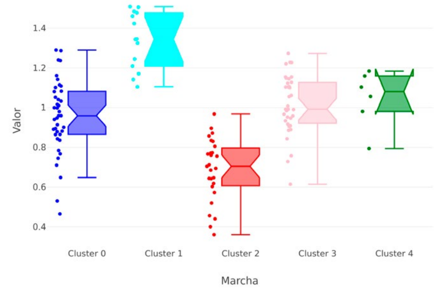


FIGURA 5. Diagrama de cajas para los valores de Marcha en los clústeres del sistema ASMI + FA + Marcha.

De acuerdo con los diagramas de caja, los clústeres 2 y 3 exhiben los valores más bajos en los tres parámetros. El clúster 2 se caracteriza por presentar los valores más bajos en los tres parámetros, lo que lo convierte en un grupo de pacientes con un alto riesgo de sarcopenia. Mientras que, en el clúster 3 también se encuentran pacientes con valores bajos en el índice de masa muscular apendicular (ASMI). Dado que el ASMI depende específicamente de la cantidad de músculo en las extremidades, estos pacientes también son susceptibles a presentar síntomas característicos de la sarcopenia.

En la Figura 6, se muestra el diagrama de árbol generado por el algoritmo *Random Forest*. Este árbol proporciona un conjunto de condiciones que explican la clasificación de pacientes en los clústeres obtenidos mediante el algoritmo de clustering jerárquico aglomerativo. Para simplificar la interpretación del diagrama de árbol, se especificó un número mínimo de 15 muestras por nodo terminal. Esto significa que los modelos de árbol no describen a todos los pacientes de la muestra, pero esta simplificación apenas afecta la precisión de la clasificación [34].

En la Figura 6 se observa que la primera condición se relaciona con la medición de la fuerza promedio de agarre. Si esta medida es menor que 17.9, los participantes de cada clúster siguen un conjunto de pruebas a la izquierda; de lo contrario, se someten a las pruebas indicadas en el lado derecho para este árbol de decisión.

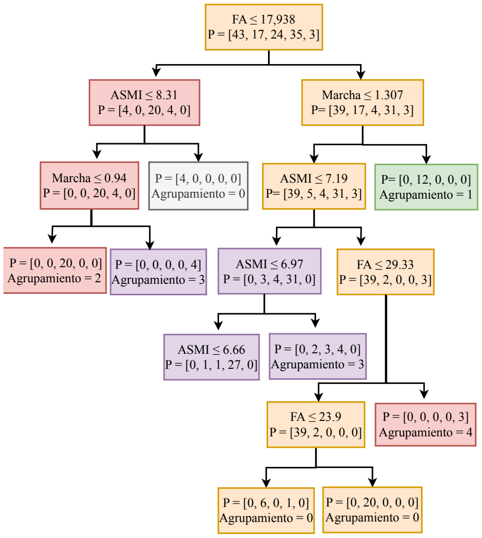


FIGURA 6. **Árbol de decisión para clasificar pacientes de acuerdo con el sistema ASMI + FA + Marcha.**

Con base a los diagramas de cajas (Figuras 3 a 5), los clústeres 2 y 3 contienen a los pacientes que podrían recibir un diagnóstico de sarcopenia, ya que muestran los valores más bajos para ASMI, FA y marcha. El clúster 1 también presenta valores bajos en la marcha, similares a los del clúster 2, pero se caracteriza por valores significativamente más altos en FA. El clúster 2 consta de 24 pacientes, de los cuales 20 presentaron una FA menor a 17.94, ASMI < 8.31 y marcha < 0.94. Estas condiciones permiten describir a 20 de los 24 pacientes en este clúster, considerándose como la rama principal. Los otros 4 pacientes cumplen con las condiciones FA > 17.94, marcha < 1.3 y ASMI < 7.20. Respecto al tercer clúster, de los 35 participantes, 31 cumplen con las condiciones FA > 17.93, marcha < 1.30 y ASMI < 6.97. Estas secuencias de condiciones representan la base de la clasificación de los pacientes en los clústeres de interés.

La siguiente Figura es la representación simplificada de los clústeres proporcionando una visión visual de la distribución de las personas adultas mayores en función de sus características. La Figura 5, se presenta un diagrama bidimensional generado mediante la aplicación secuencial de técnicas de reducción de dimensionalidad, como t-SNE *(t-distributed Stochastic Neighbor Embedding)*, PCA (*Principal Component Analysis)* y t-SNE nuevamente.

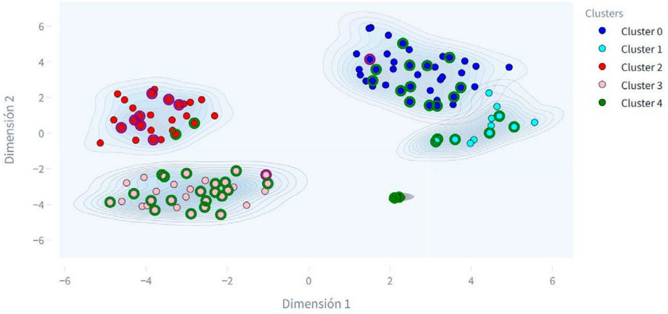


FIGURA 7. **Gráfico bidimensional que muestra la densidad de puntos del sistema ASMI + FA + Marcha.**

La Figura 7 muestra los registros con diagnóstico confirmado de sarcopenia, estos son representados por puntos encerrados en círculos. De los 11 pacientes con diagnóstico confirmado, 10 se localizan dentro del clúster 2, mientras que 1 se encuentra en el tercer clúster. Esta observación sugiere una alta concordancia entre la clasificación de los pacientes con diagnóstico confirmado y su agrupación en los clústeres identificados.

Además, es importante señalar que los pacientes con diagnóstico confirmado que se encuentran dentro de cada clúster cumplen con las condiciones descritas por las ramas principales correspondientes en el diagrama de árbol previamente mencionado. Esta concordancia entre las condiciones establecidas por el algoritmo de árbol y la distribución de los pacientes con diagnóstico confirmado refuerza la utilidad de las condiciones propuestas para la identificación de la sarcopenia en la población estudiada.

Sistema: MP + FA + Marcha + Grasa

Las Figuras 8 a 11 muestran los diagramas de cajas correspondientes al segundo sistema de clasificación, el cual se basa en la fuerza promedio de agarre (FA), la velocidad de marcha, la masa muscular promedio de piernas (MP) y el porcentaje corporal de grasa.

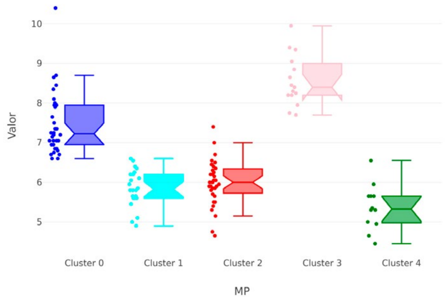


FIGURA 8. **Diagrama de cajas para los valores de MP en los clústeres del sistema MP + FA + Marcha + Grasa.**

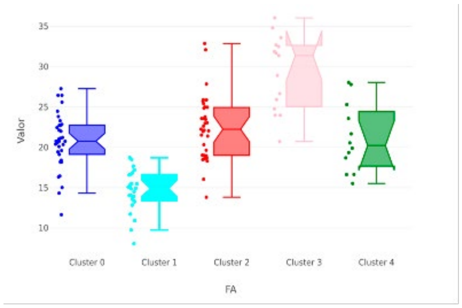


FIGURA 9. **Diagrama de cajas para los valores de FA en los clústeres del MP + FA + Marcha + Grasa.**

FIGURA 10. **Diagrama de cajas para los valores de Marcha en los clústeres del sistema MP + FA + Marcha + Grasa.**

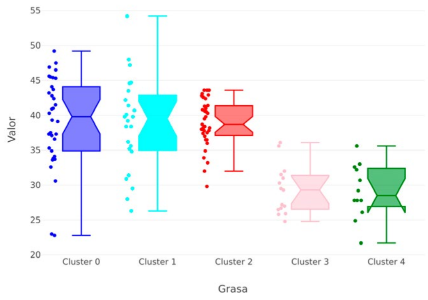


FIGURA 11. **Diagrama de cajas para los valores de Grasa en los clústeres del sistema MP + FA + Marcha + Grasa.**

De la Figura 8 a la 11, los clústeres de interés son el 0 y el 1. En el caso del clúster 0, se destacan por mostrar valores bajos en velocidad de marcha y fuerza promedio de agarre, así como valores altos en el porcentaje de grasa corporal, lo que sugiere una proporción muscular reducida, siendo esta una medida complementaria al porcentaje de masa muscular corporal. No obstante, ambos clústers se pueden diferenciar por los valores promedio de masa muscular en las piernas, siendo menores a 7 kg para más de la mitad de los pacientes en el clúster 0 y menores a 5.8 kg para la mayoría de las personas adultas mayores en el primer clúster y, en consecuencia, presentan poca fuerza de agarre, baja velocidad de marcha y escasa masa muscular en las piernas, características que coinciden con los perfiles asociados a la sarcopenia [4][5][6][7][8].

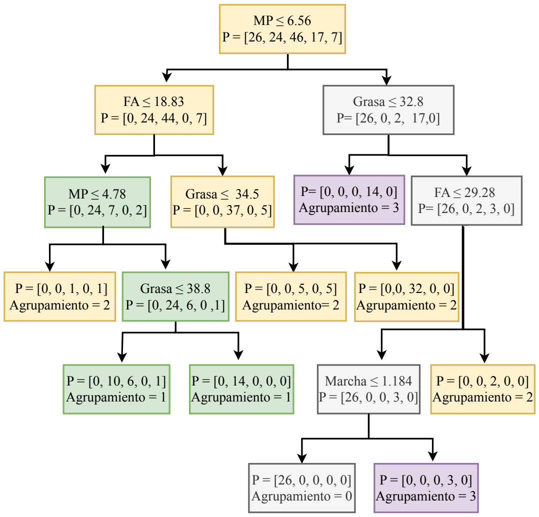


FIGURA 12. **Árbol de decisión para clasificar pacientes de acuerdo con el sistema MP + FA + Marcha + Grasa.**

En lo que respecta al modelo de *Random Forest* (Figura 12), las reglas que definen la rama principal de pacientes para el clúster 1 abarcan a 24 pacientes. La secuencia de condiciones implica que MP debe ser menor o igual a 6.57, FA menor o igual a 18.8, MP menor o igual a 4.77, y luego se bifurca en la última condición relacionada con el porcentaje corporal de grasa, donde 14 de los 24 pacientes presentan un porcentaje de grasa superior al 38.88 %, mientras que los 10 restantes cumplen la condición complementaria.

En cuanto al clúster 0, un total de 26 pacientes fueron clasificados por el modelo de *Random Forest*. Todos estos pacientes cumplen con las condiciones de MP > 6.57, Grasa > 32.8, FA <= 29.27 y Marcha <= 1.18.

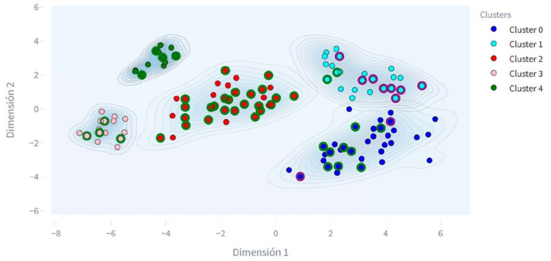


FIGURA 13. Gráfico bidimensional que muestra la densidad de puntos del sistema MP + FA + Marcha + Grasa.

Como se puede observar en la Figura 13, el clúster 1 contiene a 8 de los 10 diagnósticos confirmados, mientras que los otros 2 se clasifican dentro del clúster 0.

Sistema: MB + FA + Marcha + Grasa

Las Figuras 14 a 17 muestran los diagramas de cajas correspondientes al segundo sistema de clasificación, el cual se basa en la masa muscular de brazo (MB), fuerza promedio de agarre (FA), la velocidad de marcha y el porcentaje corporal de grasa.

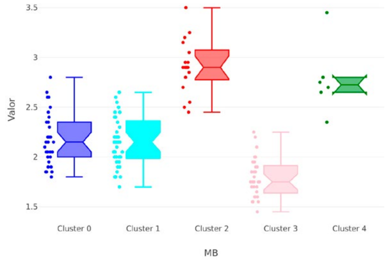


FIGURA 14. Diagrama de cajas para los valores de MB en los clústeres del sistema MB + FA + Marcha + Grasa.

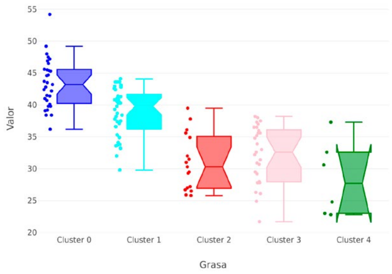


FIGURA 15. Diagrama de cajas para los valores de Grasa en los clústeres del sistema MB + FA + Marcha + Grasa.

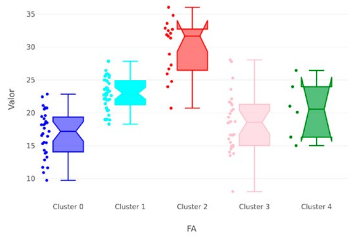


FIGURA 16. Diagrama de cajas para los valores de MB en los clústeres del sistema MB + FA + Marcha + Grasa.

FIGURA 17. Diagrama de cajas para los valores de Marcha en los clústeres del sistema MP + FA + Marcha + Grasa.

Las Figuras 14 a 17 exhiben diagramas de cajas que resaltan las diferencias esenciales entre los clústeres. El clúster 0 se caracteriza por una marcha significativamente más lenta en comparación con los otros grupos. Además, este conjunto de pacientes muestra niveles más elevados de grasa corporal y una fuerza de agarre considerablemente menor en contraste con los otros clústeres, junto con valores intermedios de masa muscular en brazos. Ningún otro cumple con condiciones similares. El clúster 3 se acerca al clúster 0 en términos de los valores más bajos de masa muscular en brazos, fuerza reducida y valores normales de grasa, pero la mayoría de sus pacientes mantienen velocidades promedio de marcha superiores a 1.0 metro por segundo. Estas observaciones enfatizan las características distintivas de cada clúster con respecto a la marcha, la composición corporal y la fuerza muscular. A continuación, se reporta el árbol de decisión para este sistema.

En relación con el modelo de *Random Forest* (Figura 18), las reglas que definen la rama principal de pacientes para el clúster 0 involucran a 31 pacientes. La secuencia de condiciones incluye que la velocidad de marcha sea menor o igual a 0.88, que la fuerza promedio de agarre esté en el rango de 8.9 a 22.95, que la masa muscular promedio en brazos sea menor o igual a 2.58 kg, y que el porcentaje de grasa corporal sea mayor al 36.86 %. Estas reglas describen las condiciones que cumplen 29 de estos 31 participantes. Como se puede apreciar en la Figura 19, el clúster 0 contiene a 8 de los 10 diagnósticos confirmados, mientras que los otros 2 se clasifican en los clústeres 1 y 4.

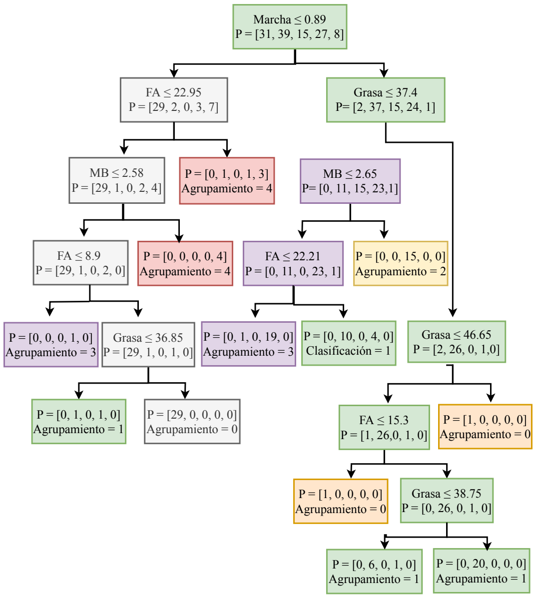


FIGURA 18. **Árbol de decisión para clasificar pacientes de acuerdo con el sistema MB + FA + Marcha + Grasa.**

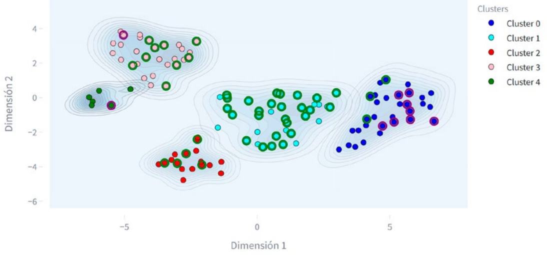


FIGURA 19. **Gráfico bidimensional que muestra la densidad de puntos del sistema MB + FA + Marcha + Grasa.**

CONCLUSIONES

Este estudio explora el uso de mediciones antropométricas en adultos mayores para desarrollar una aplicación que asista en el diagnóstico de sarcopenia. Se identificaron reglas de decisión basadas en estas mediciones para la construcción de la metodología propuesta.

Este estudio inició con la depuración de los participantes eliminando registros repetidos, incompletos y datos atípicos. Se realizó la reducción de los parámetros antropométricos eligiendo la mayor varianza, seguida por una alta correlación de Spearman y se validó con la prueba de *t-Student*, obteniendo una lista de cinco parámetros que son consistentes con los reportados en la sintomatología de sarcopenia.

Para establecer los puntos de corte de los parámetros minimizados se propusieron tres sistemas que fueron validados respecto a diagnósticos confirmados de sarcopenia. Los resultados mostraron que, es posible realizar el diagnóstico primario empleando el modelo japonés. El segundo sistema pudo categorizar la sarcopenia clásica y la relacionada con la obesidad, aunque este diagnóstico requiere evaluación adicional con pruebas específicas. Mientras que, el tercer sistema identificó síntomas consistentes con sarcopenia, como marcha lenta, debilidad en la fuerza de agarre, baja masa muscular en los brazos y exceso de grasa corporal.

Un hallazgo interesante fue que los parámetros que no dependen de medir los brazos o las pantorrillas son más simples y menos intrusivos, lo que los hace más accesibles y cómodos para los pacientes, así como más económicos y rápidos de aplicar. Sin embargo, esta simplificación puede llevar a una pérdida de información y menor precisión en la evaluación de la sarcopenia. La elección de utilizar estos sistemas dependerá de los objetivos clínicos y las limitaciones de recursos en un entorno médico específico. Además, en situaciones clínicas reales, los pacientes a veces están postrados, deformaciones óseas, inflamaciones articulares, o bien el personal no dispone del equipo de medición, limita la posibilidad de extraer mediciones antropométricas precisas.

Como trabajo a futuro se plantea emplear esta metodología de trabajo con bases de datos como los generados en la Encuesta Nacional de Salud y Nutrición para determinar la factibilidad teórica y técnica de la actual propuesta.

DECLARACIÓN ÉTICA

El proyecto fue aprobado por el Comité de Bioética del Instituto Estatal de Cancerología, verificado por la Comisión Nacional de Bioética en México bajo el registro: CEICANCL08012021-MARCOGP-01-RA1.

AGRADECIMIENTOS

Los autores agradecen al Consejo Nacional de Humanidades, Ciencias y Tecnologías, por la beca posdoctoral otorgada al CVU 231269, que permitió el desarrollo de este trabajo.

CONTRIBUCIÓN DE CADA AUTOR

S.A.D. conceptualizó la propuesta, desarrolló los algoritmos, analizó los resultados y participó en la redacción del documento. X.A.R.T.T. gestionó la obtención de los registros antropométricos de los participantes ante la Comisión Nacional de Bioética de México, y además colaboró en la revisión y retroalimentación del manuscrito. J.A.B.B. recolectó y curó los datos y contribuyó en la conceptualización de la propuesta. E.E.B.B. conceptualizó y desarrolló la metodología, y participó en la redacción del manuscrito. M.R.S. y M.H.V. contribuyeron en la redacción de la discusión y revisión bibliográfica. J.R.G.F. realizó pruebas de campo e ingresó las mediciones al software para evaluar la usabilidad y compatibilidad con los sistemas operativos Android e iOS.

REFERENCIAS

[1] B. Ballina Negrete, "La construcción institucional de la educación inclusiva en México," Tesis de pregrado, Univ. Nac. Aut. Mex., Ciudad de México, México, 2023. [En línea]. Disponible en: https://ru.dgb.unam.mx/bitstream/20.500.14330/TES01000842167/3/0842167.pdf

[2] L. F. Faúndez-Pino, "Envejecimiento saludable y enfoque salutogénico," Rev. Mex. Med. Fam., vol. 10, no. 1, pp. 39-45, 2023. [En línea]. Disponible en: https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2696-12962023000100039

[3] M. M. Hechavarría Ávila, M. Ramírez Romaguera, H. García Hechavarria y A. García Hechavarria, "El envejecimiento. Repercusión social e individual," Rev. Inf. Cient., vol. 97, no. 6, pp. 1173-1188, 2018. [En línea]. Disponible en: http://scielo.sld.cu/scielo.php?pid=S1028-99332018000601173&script=sci_arttext

[4] M. Casati, A. S. Costa, D. Capitanio, L. Ponzoni, E. Ferri, S. Agostini y E. Lori, "The Biological Foundations of Sarcopenia: Established and Promising Markers," Front. Med., vol. 6, art. no. 184, 2019, doi: http://dx.doi.org/10.3389/fmed.2019.00184

[5] H. Simsek, R. Meseri, S. Sahin, A. Kilavuz, et al., "Prevalence of sarcopenia and related factors in community-dwelling elderly individuals," Saudi Med. J., vol. 40, no. 6, pp. 568-574, 2019, doi: https://doi.org/10.15537%2Fsmj.2019.6.23917

[6] P. B. V. Fernández, A. E. Domínguez, I. B. Vázquez, C. A. Acosta, S. C. Diaz, D. D. A. Navarro, "Sarcopenia and related factors in women over 40 years," Rev. Cuba. Endoc., vol. 31 no. 3, art. no. e198, 2020. [En línea]. Disponible en: https://www.medigraphic.com/cgi-bin/new/resumenI.cgi?IDARTICULO=100636

[7] A. Casirati, G. Vandoni, S. Della Valle, G. Greco, M. Platania, S. Colatrugio, L. Lalli, C. Gavazzi, "Nutritional status and body composition assessment in patients with a new diagnosis of advanced solid tumour: Exploratory comparison of computed tomography and bioelectrical impedance analysis," Clin. Nutr., vol. 40, no. 3, pp. 1268-1273, 2021, doi: http://dx.doi.org/10.1016/j.clnu.2020.08.009

[8] W. A. Sepulveda Loyola, G. A. Luna Corrales, F. Ganz, H. Gonzalez Caro y V. Suziane Probst, "Sarcopenia, definición y diagnóstico: ¿Necesitamos valores de referencia para adultos mayores de Latinoamérica?," Rev. Chil. Ter. Ocup., vol. 20, no. 2, pp. 259-267, 2020, doi: https://doi.org/10.5354/0719-5346.2020.53583

[9] T. M. Nicolalde Cifuentes y S. I. Heredia Aguirre, "Parámetros antropométricos y su relación con la actividad física en una población adulta," Polo Conoc., vol. 6 no. 9, pp. 723-736, 2021. [En línea]. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=8094563

[10] K. Keller, "Sarcopenia," Wien. Med. Wochenschr., vol. 169, pp. 157-172, 2019, doi: https://doi.org/10.1007/s10354-018-0618-2

[11] M. C. Espinel-Bermúdez, S. Sánchez-García, C. García-Peña, X. Trujillo, M. Huerta-Viera, V. Granados-García, S. Hernández-González, y E. D. Arias-Merino, "Factores asociados a sarcopenia en adultos mayores mexicanos: Encuesta Nacional de Salud y Nutrición 2012," Rev. Med. Inst. Mex. Seguro Soc., vol. 56, no. 1, pp. 46-53, 2018. [En línea]. Disponible en: https://www.redalyc.org/articulo.oa?id=457754907008

[12] G. S. Collins y K. G. Moons, "Reporting of artificial intelligence prediction models," The Lancet, vol. 393, no. 10181, pp. 1577-1579, 2019, doi: https://doi.org/10.1016/s0140-6736(19)30037-6

[13] M. Banerjee, E. Reynolds, H. B. Andersson y B. K. Nallamothu, "Tree-based analysis: a practical approach to create clinical decision-making tools," Cir. Cardiovasc. Qual. Outcomes, vol. 12, no. 5, art. no. e004879, 2019, doi: https://doi.org/10.1161%2FCIRCOUTCOMES.118.004879

[14] S. O. Lugo-Reyes, G. Maldonado-Colín y C. Murata, "Inteligencia artificial para asistir el diagnóstico clínico en medicina," Rev. Alerg. Mex., vol. 61, no. 2, pp. 110-120, 2014, doi: https://doi.org/10.29262/ram.v61i2.33

[15] L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5-32, 2001, doi: https://doi.org/10.1023/A:1010933404324

[16] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava y T. Baker, "Analysis of dimensionality reduction techniques on big data," IEEE Access, vol. 8, pp. 54776-54788, 2020, doi: https://doi.org/10.1109/access.2020.2980942

[17] R. Krzymińska-Siemaszko, E. Deskur-Śmielecka, A. Kaluźniak-Szymanowska, A. Styszyński y K. Wieczorowska-Tobis, "Polish version of SARC-F to assess sarcopenia in older adults: An examination of reliability and validity," PLoS One, vol. 15, no. 12, art. no. e0244001, 2020, doi: http://dx.doi.org/10.1371/journal.pone.0244001

[18] S. Huang, J. Yang, S. Fong y Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," Cancer Lett., vol. 471, pp. 61-71, 2020, doi: http://dx.doi.org/10.1016/j.canlet.2019.12.007

[19] F. Morandín-Ahuerma, "Diez recomendaciones de la Unesco sobre ética de la inteligencia artificial," En: Principios normativos para una ética de la inteligencia artificial. Puebla, México: Consejo de Ciencia y Tecnología del Estado de Puebla (CONCYTEP). pp. 86-94.

[20] Cámara de Diputados del H. Congreso de la Unión. (7 de febrero de 1984). Ley General de Salud. [En línea]. Disponible en: https://www.diputados.gob.mx/LeyesBiblio/pdf/LGS.pdf

[21] J. W. Osborne, Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data, Thousand Oaks, CA: Sage Publications Inc., 2013, doi: https://doi.org/10.4135/9781452269948

[22] J. F. Ávila-Tomás, M. A. Mayer-Pujadas y V. J. Quesada-Varela, "La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas," Aten. Primaria, vol. 53, no. 1, pp. 81-88, 2020, doi: https://doi.org/10.1016/j.aprim.2020.04.014

[23] S. H. Park y K. Han, "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction," Radiology, vol. 286, no. 3, pp. 800-809, 2018, doi: http://dx.doi.org/10.1148/radiol.2017171920

[24] L. Lera, C. Albala, H. Sánchez, B. Angel, M. J. Hormazabal, C. Márquez y P. Arroyo, "Prevalence of Sarcopenia in Community-Dwelling Chilean Elders According to an Adapted Version of the European Working Group on Sarcopenia in Older People (EWGSOP) Criteria," J. Frailty Aging., vol. 6, no. 1, pp. 12-17, 2017, doi: https://doi.org/10.14283/jfa.2016.117

[25] J. Moncada Jiménez, Estadística para ciencias del movimiento humano. Costa Rica: Editorial de la Universidad de Costa Rica, 2005.

[26] S. Huang, J. Yang, S. Fong y Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," Cancer Lett., vol. 471, pp. 61-71, 2020, doi: http://dx.doi.org/10.1016/j.canlet.2019.12.007

[27] A. T. Azar y S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," Neural Comput. Applic., vol. 23, pp. 2387-2403, 2013, doi: http://dx.doi.org/10.1007/s00521-012-1196-7

[28] D. Blázquez García, "Análisis de datos de expresión en retina para la identificación de biomarcadores de DMAE: posible herramienta diagnóstica," Tesis de pregrado, Fac. Med., Univ. Vall., Valladolid, España, 2023.

[29] F. T. T. Lai, P. E. Beeler, B. H. K. Yip, M. Cheetham, et al., "Comparing multimorbidity patterns among discharged middle-aged and older inpatients between Hong Kong and Zurich: A hierarchical agglomerative clustering analysis of routine hospital records," Front. Med., vol. 8, art. no. 651925, 2021. https://doi.org/10.3389/fmed.2021.651925

[30] P. Lemenkova, "R Libraries {dendextend} and {magrittr} and Clustering Package scipy.cluster of Python For Modelling Diagrams of Dendrogram Trees," Carpathian J. Electr. Comp. Eng., vol. 13, no. 1, pp. 5-12, 2020, doi: https://doi.org/10.2478/cjece-2020-0002

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, no. 85, pp. 2825–2830, 2011. [En línea]. Disponible en: http://jmlr.org/papers/v12/pedregosa11a.html

[32] A. K. Castañeda Escobar, "Diseño e implementación de un algoritmo para determinar la mejor agrupación utilizando k-means y los índices de validación," Tesis de maestría, Univ. Aut. Edo. Mor., Morelos, México, 2023. [En línea]. Disponible en: http://riaa.uaem.mx/handle/20.500.12055/3115

[33] E. Goffinet, M. Lebbah, H. Azzag, L. Giraldi y A. Coutant, "Multivariate time series multi-coclustering: Application to advanced driving assistance system validation," in Europ. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn., Oct. 2021, pp. 463-468, doi: https://doi.org/10.14428/esann/2021.es2021-119

[34] S. J. Rigatti, "Random forest," J. Insur. Med., vol. 47, no. 1, pp. 31-39, 2017, doi: http://dx.doi.org/10.17849/insm-47-01-31-39.1

[35] M. Bakator y D. Radosav, "Deep Learning and Medical Diagnosis: A Review of Literature," Multimodal Technol. Interact., vol. 2, no. 3, art. no. 47, 2018, doi: https://doi.org/10.3390/mti2030047

[36] A. Yuki, F. Ando, R. Otsuka, Y Matsui, A. Harada, H. Shimokata, "Epidemiology of sarcopenia in elderly Japanese," J. Phys. Fitness Sports Med., vol. 4, no. 1, pp. 111-115, 2015, doi: http://dx.doi.org/10.7600/jpfsm.4.111

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 70 - 83

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.5

E-LOCATION ID: 1387

Sensor for Real-Time Glucose Measurement in Aqueous Media based on Nanomaterials Incorporating an Artificial Neural Network Algorithm on a System-On-Chip

Sensor para Medición de Glucosa en Tiempo Real para Medios Acuosos basado en Nanomateriales Incorporando un Algoritmo de Red Neuronal Artificial en un Sistema en Chip

Xenia Azareth Ayon-Gómez[1], *Ulises Jesús Tamayo-Pérez*[1], *Oscar Roberto López-Bonilla*[1],
Oscar Adrian Aguirre-Castro[1], *Eunice Vargas-Viveros*[1], *Enrique Efrén García-Guerrero*[1]

[1]Facultad de Ingeniería, Arquitectura y Diseño (FIAD), Universidad Autónoma de Baja California (UABC,) Baja California- México

ABSTRACT

The aim of this paper is to present the development of a real-time measurement system for glucose in aqueous media. The proposed system incorporates two lines of research: i) design, synthesis, and implementation of a non-enzymatic electrochemical sensor of Multi-Walled Carbon Nanotubes with Copper nanoparticles (MWCNT-Cu) and ii) design and implementation of a machine learning algorithm based on an Artificial Neural Network Multilayer Perceptron (ANN-MLP), which is embedded in an ESP32 SoC (System on Chip). From the current data that is extracted in real-time during the oxidation-reduction process to which an aqueous medium is subjected, it feeds the algorithm embedded in the ESP32 SoC to estimate the glucose value. The experimental results show that the nanostructured sensor improves the resolution in the amperometric response by identifying an ideal place for data collection. For its part, the incorporation of the algorithm based on an ANN embedded in a SoC provides a level of 97.8 % accuracy in the measurements. It is concluded that incorporating machine learning algorithms embedded in low-cost SoC in complex experimental processes improves data manipulation, increases the reliability of results, and adds portability.

PALABRAS CLAVE: ANN-MLP, multi-walled CNT, electrochemical sensor, nanostructured glucose sensor, ESP32 SoC

RESUMEN

El objetivo de este artículo es presentar el desarrollo de un sistema de medición en tiempo real de glucosa en medios acuosos. El sistema que se implementa incorpora dos lineas de investigación: i) diseño, síntesis e implementación de un sensor electroquímico no enzimático de Nanotubos de Carbono de Pared Múltiple con nanopartículas de Cobre (NTCPM-Cu) y ii) diseño e implementación de un algoritmo de aprendizaje automático basado en una Red Neuronal Perceptrón Multicapa (RN-PM), embebido en un ESP32 SoC (Sistema en Chip). Un dato de corriente que se extrae en tiempo real durante el proceso de oxidación-reducción a la que se somete un medio acuoso, alimenta el algoritmo embebido en el ESP32 para estimar el valor de glucosa. De los resultados experimentales se demuestra que el sensor nanoestructurado mejora la resolución en la respuesta amperométrica al identificar un lugar ideal para la toma de datos. Por su parte, la incorporación del algoritmo basado en una RN embebido en SoC otorga un nivel de 97.8 % de exactitud en la mediciones. Se concluye que incorporar algoritmos de aprendizaje automático embebidos en SoC de bajo costo en procesos experimentales complejos, mejora la manipulación de datos, incrementa la confiabilidad en resultados y adiciona portabilidad.

KEYWORDS: ANN-MLP, NCT de pared múltiple, sensor electroquímico, sensor nanoestructurado de glucosa, SoC ESP32

Corresponding author

TO: **Enrique Efrén García-Guerrero**
INSTITUTION:: **Universidad Autónoma de Baja California**
ADDRESS: **Carretera Transpeninsular Ensenada - Tijuana 3917, Zona Playitas, 22860 Ensenada, B.C.**
EMAIL: **eegarcia@uabc.edu.mx**

Received:

25 October 2023

Accepted:

24 November 2023

INTRODUCTION

The scientific community is constantly working to develop highly efficient, accessible, and stable glucose sensors to monitor glucose levels. However, those based on enzymes are not very stable since, in some cases, they are susceptible to pH and temperature [1][2][3]. With the advent of nanoscience and nanotechnology, scientists have developed and fabricated advanced nanostructured materials over the past two decades to develop glucose sensors with high sensitivity and selectivity [4][5]. Carbon nanotubes (CNT), metal nanoclusters, and nanoparticles are among the nanostructured materials used. CNT and transition metal nanoclusters have had numerous applications as sensors and biosensors [6]. Nanostructured carbon systems such as CNT, graphene, and mesoporous carbon are the most studied materials for glucose detection thanks to their different properties, such as large surface area, better conductivity, biocompatibility, and chemical stability [7][8]. For their part, smaller metal clusters have shown catalytic activity not exhibited by their massive analog or nanoparticles (NP); this generates excellent promise for their possible application as catalysts [9].

Additionally, several metallic clusters have been revealed to have tremendous and selective catalytic activity when deposited on suitable support [10][11]. Hybrids based on the combination of metal and CNT have been positioned as the new trend to produce a synergistic effect in glucose detection due to the outstanding catalytic activity of the metal and the exceptional properties of carbon materials, such as the large surface area, exceptional electrical conductivity and excellent chemical resistance [12]. For example, copper (Cu) is one of the widely investigated metal catalysts, and its electrodeposition has been studied on various carbon nanostructures [13][14][15]. Cu-based nanomaterials are among the most studied nanostructures in sensor and biosensor applications due to their high efficiency, low toxicity, low cost, long-term stability, and high catalytic activity. In [16], the synthesis of nanostructures with different morphologies of copper oxide (CuO) for the detection of glucose; however, the synthesis is complex to carry out. In [17], a nanocomposite formed from Cu and CNT (Cu-CNT) exhibits high sensitivity and stability, fast response with a wide linear range, a low detection limit, and exemplary performance detecting glucose in blood serum. This process is developed by synthesizing carbon nanotubes grown on a Tantalum (Ta) substrate from chemical vapor deposition (CVD) and electrochemically depositing Cu nanocubes on the CNT using a simple cathodic potentiostatic technique. In [18], a non-enzymatic glucose sensor with catalytic oxidation through the electrodeposition of Cu nanoclusters on an electrode modified with carbon nanotubes solubilized with Nafion (Nf). This sensor is applied for glucose analysis in blood samples and provides high sensitivity and stability, fast response, good reproducibility, and selectivity. These qualities make the Cu-CNT nanocomposite-based electrode a promising candidate for enzyme-free glucose detection.

Implementing artificial intelligence (AI) algorithms in systems based on nanostructured biosensors is of great help because it can estimate expected values quickly, accurately, automatically, and in real time. Additionally, an adequate database can train algorithms to estimate particular variables. Machine learning can efficiently process large amounts of messy or low-resolution data and discriminate overlapping signals from each other [19]. In [20], a method based on using artificial intelligence algorithms to process the interference between the oxidation currents of insulin and glucose using cyclic voltammetry with a $Ni(OH)_2$ electrode. The method allows for separating and providing insulin and glucose concentrations, obtaining high precision (at the mmol level) and correlation, with prediction errors of 6.515 % for insulin and 4.36 % for glucose. In [21], an application of an electronic tongue formed by several electrodes modified with nanoparticles for the detection and electrochemical quantification of carbohydrates in sugarcane waste samples. It uses an artificial neural network to process the data and predict carbohydrate concentrations with high precision and correlation.

This article presents the design and implementation of glucose measurement in real-time of a system in aqueous media based on machine learning algorithms embedded in an ESP32 SoC. The instrument is made up of two main blocks: i) the electrochemical stage, where an oxidation-reduction process of an aqueous medium with glucose is carried out and ii) the information processing stage and the estimation of the glucose level, based on machine learning algorithms implemented in an ESP32 SoC. The first block's main contribution is designing, synthesizing, and implementing a non-enzymatic sensor built from a glassy carbon electrode functionalized with MWCNT and copper nanoparticles. Considering that the synergy between copper nanoparticles and MWCNT increases the amperometric response, the resolution and sensitivity of the sensor are improved. The second block estimates the level of glucose present in an aqueous medium. The main contribution at this stage is the method to estimate glucose levels in the solution. It is from current data obtained in real-time in the oxidation-reduction stage of the aqueous medium with glucose, which is fed to the embedded machine-learning algorithm. The embedded algorithm based on an ANN with MLP architecture determines the corresponding glucose level. Therefore, the implemented measurement system incorporates the advantages of nanotechnology through a nanostructured MWCNT-Cu sensor (Multi-Walled Carbon Nanotubes with Copper nanoparticles), the benefits of AI through the training of machine learning algorithms based ANN-MLP (artificial neural network multilayer perceptron) and the potential of digital electronic by the ANN-MLP algorithms, embedded in an ESP32 system-on-chip (SoC). The integration of these elements makes a robust system that estimates the glucose level in real-time of an aqueous medium in an oxidation-reduction process under cyclic voltammetry.

The article is structured as follows: The Materials and Methods Section presents the design, synthesis, implementation, and validation of a non-enzymatic MWCNT-Cu electrochemical sensor. The type of architecture of the ANN-MLP implemented is also presented, as well as the mathematical model of the embedded algorithm in an ESP32 SoC. Finally, the experimental setup is shown. In the Results and Discussion Section, the behavior of the nanostructured sensor in obtaining experimental data on glucose concentration in aqueous measurements using cyclic voltammetry is presented. Implementing the machine learning algorithm is validated through two training processes, one for the current data set corresponding to the voltage level of 1.3 V and another for the voltage level of -0.455 V, both for different glucose concentrations. Finally, the Conclusions of this work are presented.

MATERIALS AND METHODS

The first stage is the design, synthesis, implementation, and validation of a non-enzymatic glucose sensor manufactured with electrodeposition of Cu nanoclusters on the film of a glassy carbon electrode modified with MWCNT solubilized with Nafion (CNT-Nf). The manufacturing of the sensor begins with the synthesis of MWCNT using the spray pyrolysis method; subsequently, they are functionalized, and their composition is validated by X-ray diffraction (XRD). Based on a glassy carbon electrode, MWCNT are added, and the proposed sensor is structured under Cu electrodeposition.

MWCNT Synthesis

Figure 1 shows the process used to obtain the MWCNT. Initially, a solution composed of 572 mg of ferrocene ($C_{10}H_{10}Fe$) with 26 mL of toluene ($C_6H_5CH_3$) is prepared and subjected to an ultrasonic bath for 10 min. This solution is fed to a nebulizer that uses argon (Ar) as a purge gas. This arrangement is placed in a tubular oven where the nebulization is passed through a small quartz tube (SiO_2), bringing it to 850 °C and cooling it to room temperature.

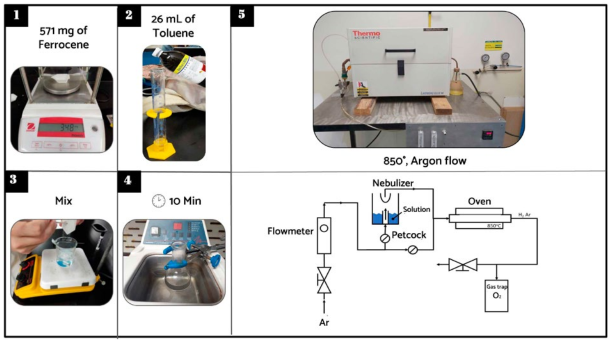


FIGURE 1. **MWCNT Synthesis.**

MWCNT Functionalization

Figure 2 shows the process used to functionalize the MWCNT. Initially, 340 mg of MWCNT are calcined at 400 °C for 30 min. These MWCNT are suspended in 60 mL of 6 M hydrochloric acid (HCl), subjected to an ultrasonic bath for 4 h, and allowed to rest for 12 h. Taking 170 mg of MWCNT, they are suspended in 20 mL of nitric acid (HNO_3) with 40 mL of sulfuric acid (H_2SO_4) and subjected to an ultrasonic bath for 4 h. Finally, filtration is done using a vacuum pump with a cellulose and water (H_2O) filter.

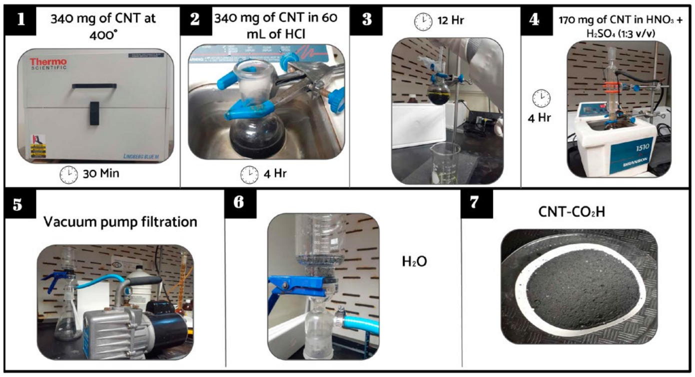


FIGURE 2. **MWCNT Functionalization.**

Characterization of MWCNT

The crystalline nature of MWCNT is validated by X-ray diffraction. Figure 3a shows the diffractograms of the pristine carbon nanotubes, which show the typical peaks at 26.3° and 43.7°, corresponding to the reflections of the graphite in (002) y (100) (Joint Committee of Powder Diffraction Standards (JCPDS) # 96-101-1061) [21]. Figure 3b shows the synthesized and functionalized MWCNT diffractogram obtained with the D2 Phaser X-ray diffractometer. It is observed that the characteristic peaks of pristine MWCNT are also present (Figure 3a). It can be seen that functionalized MWCNT have a lower intensity than pristine nanotubes motivated by the formation of carboxylic groups on the MWCNT. A slight peak at 30° is also observed due to the functionalization process.

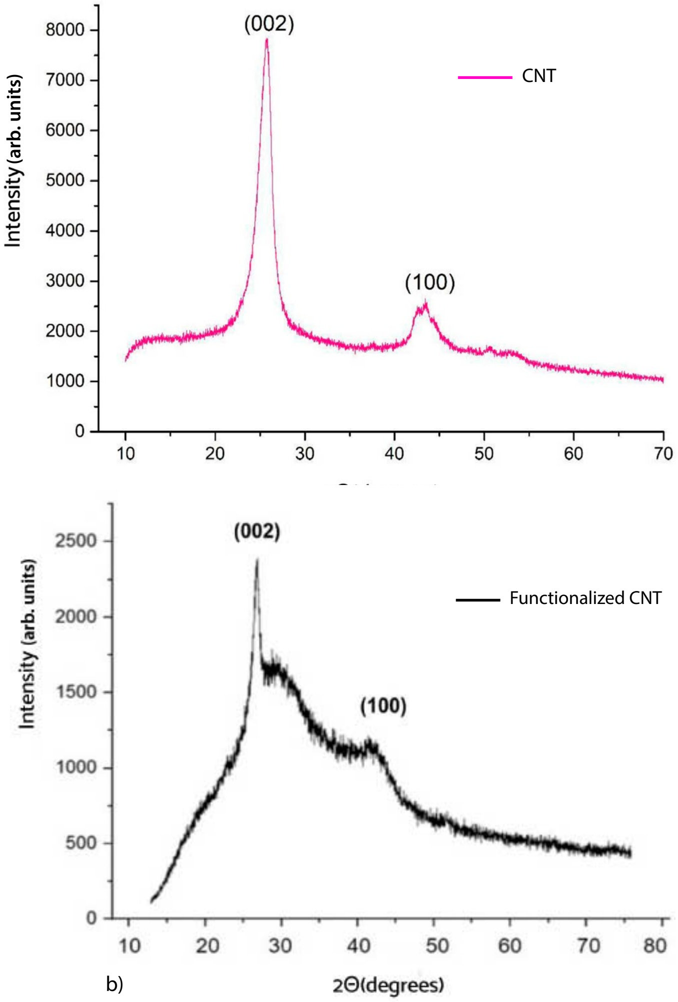


FIGURE 3. **a) Pristine MWCNT, b) Functionalized MWCNT.**

Sensor manufacturing

The proposed sensor is fabricated in two main stages: i) adhesion of the functionalized MWCNT on a glassy carbon electrode and ii) electrodeposition of Cu. Figure

4 shows the first stage of the proposed electrode preparation process. Initially, a glassy carbon electrode (GCE) is taken as a base, and its free end is polished with titanium oxide TiO_2. A 10-minute ultrasonic bath is applied in 15 mL of a solution of nitric acid (HNO_3), ethanol (C_2H_6O), and double-distilled water (1:1 v/v) left to dry at room temperature. In parallel, 1.0 mg of functionalized MWCNT are dispersed in 65 µL of 0.5% Nafion ($C_7HF_{13}O_5S{\cdot}C_2F_4$). 15 µL of the MWCNT solution with Nafion (NTC-Nf) are taken and deposited on the polished surface of the glassy carbon electrode, allowing it to dry at room temperature.

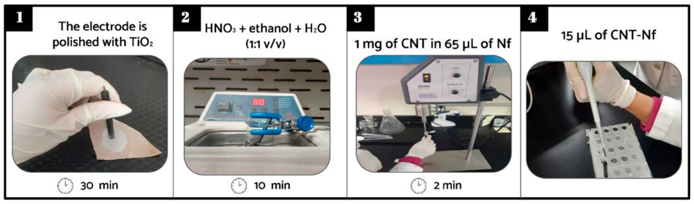


FIGURE 4. **Preparation of the glassy carbon electrode modified with MWCNT-Nf solution.**

FIGURE 5. **Electrodeposition of Cu on MWCNT-Nf.**

Figure 5 shows the second stage in the preparation of the proposed electrode. First, an electrolyte composed of 35.6 mg of sodium sulfate (Na_2SO_4) and 8 mg of copper sulfate ($CuSO_4$) dissolved in 25 mL of deionized water is prepared, applying a sonic bath for 10 min. The electrodeposition of Cu is carried out in an electrochemical cell connected to the CHI920C Scanning Electrochemical Microscope. The electrochemical cell comprises three electrodes immersed in the electrolyte prepared for this purpose. The reference electrode (RE) is a silver/silver chloride (Ag/AgCl) electrode, the counter electrode (CE) is platinum, and the working electrode (WE) is the glassy carbon electrode (GCE) with MWCNT-Nf. Cyclic voltammetry is applied with a potential window of between -0.6 to 0.6 V for 30 cycles at a scanning speed of 100 mVs^{-1}. Under this cyclic electrodeposition process, Cu nanoclusters adhere to the WE (MWCNT-Cu). The prepared work electrode (WE) is rinsed with distilled water, allowed to dry at room temperature, and refrigerated.

Machine learning algorithm and embedded system proposed

The estimation of the glucose value is based on a machine learning or ANN algorithm based on the MLP architecture embedded in an ESP32 SoC that integrates a 32-bit core RISC-V microcontroller with a maximum clock speed of 160 MHz, programmed in Python. The data that feed the ANN for training and validation are obtained from the experimental results, to which data augmentation is applied by estimating the mathematical model of the sensor's behavior. The data obtained correlates the molar concentration and the electrical current obtained from the sensor. Figure 6 shows a block diagram of the main stages of the implementation of the ANN in the ESP32 SoC. The training of the ANN is carried out with the data from the experimental arrangement through a potentiostat. Once the MLP model is trained, it is optimized to obtain the weights required to fit the corresponding mathematical model. The optimized and trained algorithm is embedded in the ESP32 SoC.

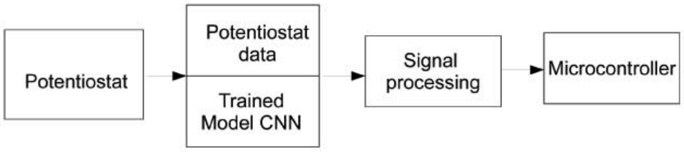


FIGURE 6. **Block Diagram of the main stages for the implementation of the ANN.**

Figure 7 shows the general architecture of the ANN-MLP. For each of the implemented models, various hyperparameter values are used until the appropriate adjustment to the behavior of the sensor is found. Each model is designed with a ReLu-type activation function and presents a sequential architecture with different hidden layers, epochs, and learning rates.

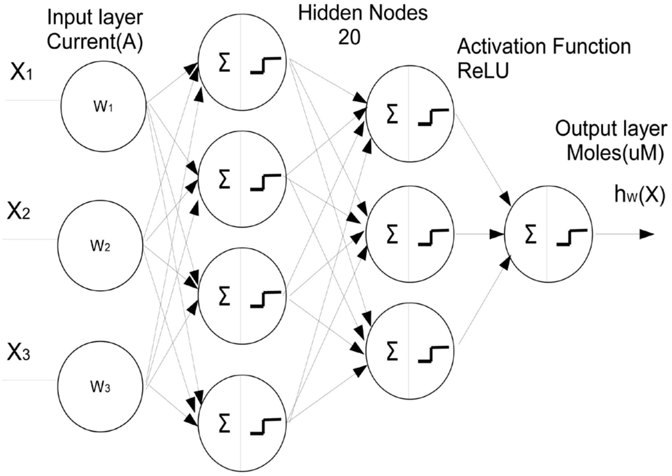


FIGURE 7. Architecture ANN-MLP.

The mathematical model used for each ANN model is given by Equation (1).

$$h_w(x) = f^n(w^n \ldots (f^2(w^2(f(wx+b))+b^2)) \ldots + b^n). \tag{1}$$

Where:

$h_w(x)$	is the expected value of glucose concentration,
f^n	is the limiting function or activation function,
w^n	corresponds to the associated weight matrices in each layer,
x	is the input data vector, and
b^n	is a bias convention from the processing node.

Algorithm 1 shows the pseudocode embedded in the ESP32 SoC, based on the ANN-MLP model trained in Python, about the mathematical model described by Equation 1.

ALGORITHM 1. Pseudocode embedded in the ESP32 SoC.

```
// Weights and bias for model ANN-MLP
double a0[1]
double w1[20][1]
double a1[20]
double w2[20][20]
double a2[20]
double w3[1][20]
double a3[1]
double b1[20]
double b2[20]
double b3[1]
// Preprocessing of ANN-MLP
double mean [1]
double dstd [1]
// Voltage measure and current estimated
I = currentEstimated
a[0] = dataNormalized (I, mean [0], dstd [0])
// ANN-MLP architecture
for (int 1=0; 1<20; i++) {aux=0.0; for (int j=0; j<1; j++) {aux=aux+W1[i][j]*a0[j];}
    a1[i]=relu(aux+b1[i]);}
for (int 1=0; 1<20; i++) {aux=0.0; for (int j=0; j<20; j++) {aux=aux+W2[i][j]*a1[j];}
    a2[i]=relu(aux+b2[i]);}
for (int 1=0; 1<1; i++) {aux=0.0; for (int j=0; j<20; j++) {aux=aux+W3[i][j]*a2[j];}
    a3[i]=linear(aux+b3[i]);}
// Glucose concentration
double GC = dataNormalized (a3[0], mean [0], dstd[0])
// Result display
lcd.print("Concentration:")
lcd.print(GC)
```

Figure 8 shows the proposed experimental setup for real-time glucose measurement for aqueous media based on an ESP32 SoC with an embedded machine-learning algorithm. From the electrochemical cell where the oxidation-reduction process of the glucose solution is carried out, a current value is obtained that feeds the ESP32 SoC. The algorithm based on an ANN-MLP estimates the value of glucose in the solution.

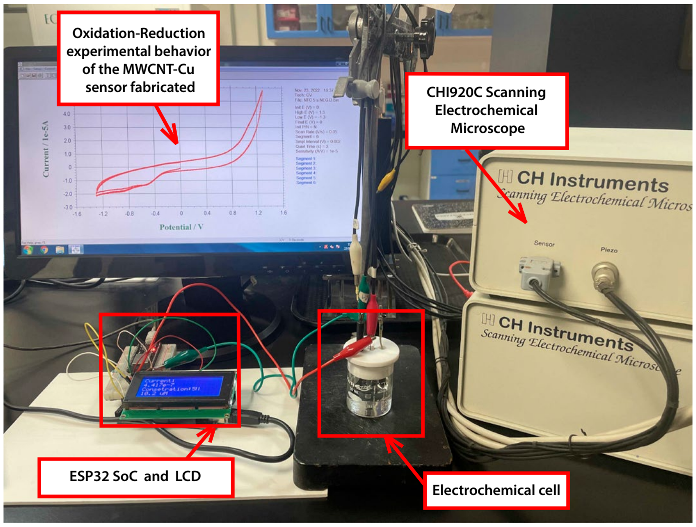


FIGURE 8. Setup for measuring glucose in aqueous media.

Experimental data on glucose concentration are obtained by cyclic voltammetry in an electrochemical cell connected to the CHI920C Scanning Electrochemical Microscope as is illustrated in Figure 8. In the electrochemical cell, three electrodes are immersed in the aqueous solution with a specific concentration of glucose (electrolyte) prepared for this purpose: i) the working electrode (WE) corresponds to the MWCNT-Cu electrode proposed in this work, ii) the reference electrode (RE) is silver/silver chloride (Ag/AgCl), and iii) the counter electrode (CE) of Platinum (Pt). The CHI920C Scanning Electrochemical Microscope supplies voltage to the WE and RE electrodes and collects current data between the WE and CE electrodes. The applied voltage is 50 mVs-1 from -1.3 V to 1.3 V, in a 20 mM alkaline solution of Sodium Hydroxide (NaOH). An electric current flows through the electrolyte in the electrochemical cell (aqueous solution with glucose) in response to the applied potential range. The electric current that flows depends on the concentration of glucose present and the voltage applied at the moment. Therefore, an electric current corresponding to each voltage value is generated when executing a potential sweep. A cyclic voltammogram (CV) is generated by graphing the applied voltage range and the obtained current (Figure 8). Therefore, each solution with a particular glucose concentration has its specific CV. Figure 9 shows how the CV is formed in three segments. In the first segment (from 0 to 1.3 V), the glucose oxidation process occurs from 0.4 to 1.3 V, corresponding to the change of oxidation state from Cu(II) to Cu(III). The second segment (from 1.3 to -1.3 V) is where the copper reduction process occurs from 1.3 to 0.23 V, corresponding to the change from Cu (III) to Cu(II), the peak at -0.455 V corresponds to the change of Cu(II) to Cu(I), where glucose is reduced, allowing its concentration to be measured. Finally, in the third segment (from -1.3V to 0 V), it is reduced from Cu(I) to Cu(0) at the -1.3V peak, Cu(0) to Cu(I) at the peak -0.34V, and Cu(I) to Cu(II) at the -0.1V peak. This oxidation-reduction process of Cu and glucose in an alkaline medium is well reported in [22] and [6]. Figure 9 shows the behavior of the MWCNT-Cu work electrode in a cyclic voltammogram (CV) oxidation-reduction process of Cu and glucose at 100µM in an alkaline medium, in the potential range from -1.3 V to 1.3 V.

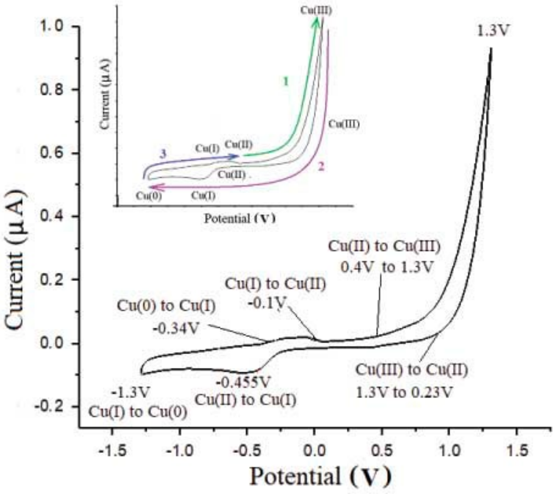


FIGURE 9. MWCNT-Cu sensor in CV.

This methodology allows the same CV to have 2 points of interest to compare and measure the glucose concentration in real-time, at -0.455 V and 1.3 V.

Thus, a database is formed for each aqueous solution with a certain glucose concentration, with the experimental voltage and current data obtained. These database are used to train the machine learning algorithm embedded in the ESP32 Soc.

RESULTS AND DISCUSSION

Figure 10 shows the oxidation-reduction behavior of the MWCNT-Cu sensor fabricated for glucose measurement in aqueous media.

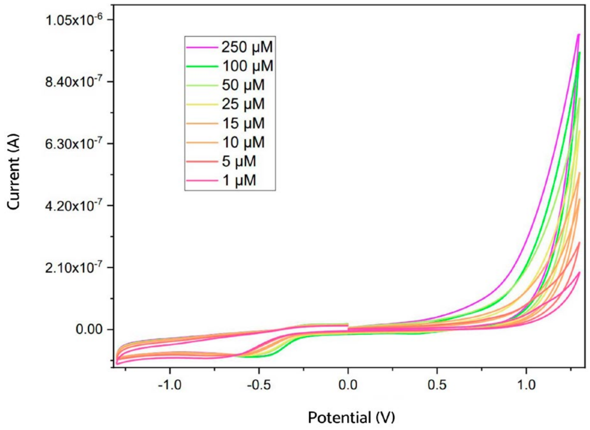


FIGURE 10. **MWCNT-Cu sensor at different glucose concentrations.**

Figure 10 shows results for solutions with different concentrations of D-glucose: 1, 5, 10, 15, 25, 50, 100, and 250 μM in 20 mM an alkaline medium. The curves are obtained in the potential range from -1.3 V to 1.3 V and a scanning speed of 50 mVs-1. The amount of glucose in a solution is directly proportional to its oxidation (amount of oxygen consumed), manifesting in an increase in current, as evidenced. In this case, oxidation for different glucose concentrations can be observed within the potential range of 0.4 V to 1.3 V. For the voltage level of 1.3 V, a clear separation is presented between the current vs. voltage curves, for the different glucose concentrations, which allow us to distinguish up to the concentration of 250 μM; this condition is attributed to the synthesized MWCNT-Cu sensor. For a voltage of -0.455 V, the curves for different glucose concentrations are more concentrated, which reduces the sensitivity of glucose measurement to a maximum of 100 μM.

The second stage in developing and implementing a glucose measurement system in aqueous media is determining the glucose level in real time. The proposed method uses current data in the oxidation-reduction stage of the glucose solution to feed a machine-learning algorithm embedded in the ESP232 SoC. The algorithm based on an ANN-MLP finally estimates the corresponding glucose level. The design of the algorithm and its training are developed in Python 3.8, using various libraries for machine learning. Each solution with a particular glucose concentration has a database formed with the experimental voltage and current data obtained. These data used to train the ANN model are obtained from the experimental values shown in Figure 10 relative to the behavior of the MWCNT-Cu sensor. Two training sessions are carried out to validate the machine learning algorithm, one for the current data set corresponding to the voltage level of 1.3 V and another for the voltage level of -0.455 V for different glucose concentrations. The proposed ANN-MLP architecture has two input layers for the two case training: the electric current and the molar concentration. For the voltage level of 1.3 V it comprises four hidden layers of 16, 20, 20, and 1 neurons and an output layer corresponding to the glucose level estimation, as shown in Table 1. The activation function of the input layer is a Relu function, the learning rate of the classifier is defined as a constant equal to 0.018.

TABLE 1. **Proposed ANN-MLP for training at 1.3 V.**

Input Layer	Hidden Layers	Output Layer
Current Measure, Molar Concentration	16-20-20-1	Molar Concentration Estimated

For the current data set corresponding to the 1.3 V voltage level, Table 2 provides hyperparameter information for this model.

TABLE 2. **Hyperparameters of the proposed ANN-MLP model at 1.3 V.**

Hyperparameters	Proposed ANN-MLP
Kernel	Linear
Optimizer	Adam
Activation Function	Relu
Error	MAE
Epoch	8500
Degree	1
Learning Rate	0.018
Bias	Boolean

The experimental value of glucose concentration is obtained by feeding the ANN-MLP algorithm trained in real-time with a current value obtained from the electrochemical cell for a specific voltage value while performing cyclic voltammetry (Figure 8). For the 1.3 V case, Table 3 shows the molar concentration data used for the first training of the proposed ANN-MLP at 1.3 V and the results obtained by the algorithm after training. It is observed that there is an average error of 0.0138 between the glucose concentration and the estimated value.

TABLE 3. **Glucose concentrations at 1.3 V.**

Experimental currents $\times 10^{-7}$ (A)	Glucose concentrations (μM)	Glucose concentration obtained with ANN-MLP (μM)
1.94	1	0.979088
2.95	5	4.98703
4.41	10	9.99455
5.31	15	14.9987
6.73	25	25.0102
7.83	50	50.0158
9.39	100	100.023
10.0	250	250.021

Figure 11 shows the final correlation between the proposed ANN-MLP training results and the experimental data. It is observed that the red curve corresponding to the results obtained by the RN tends to fit well with the blue curve corresponding to the input values, which indicates adequate algorithm training. These results are obtained with a sequential architecture with four hidden layers of 16, 20, 20, 1, a learning rate of 0.018, and a ReLu-type activation function.

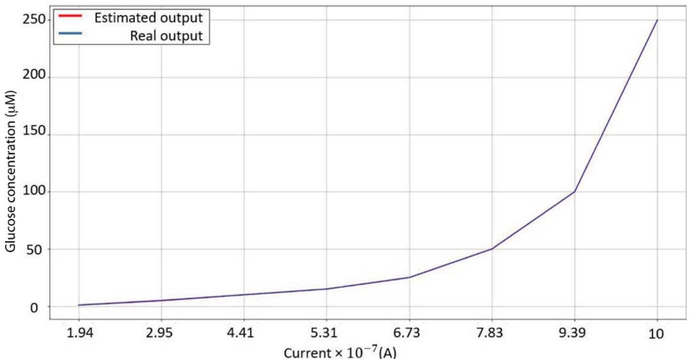


FIGURE 11. **Input data and values estimated by the proposed ANN-MLP for experimental data at 1.3 V.**

The training in the proposed ANN-MLP is adjusted for different epochs, and the error trend is evaluated. Figure 12 shows that the error tends to zero for 8500 epochs.

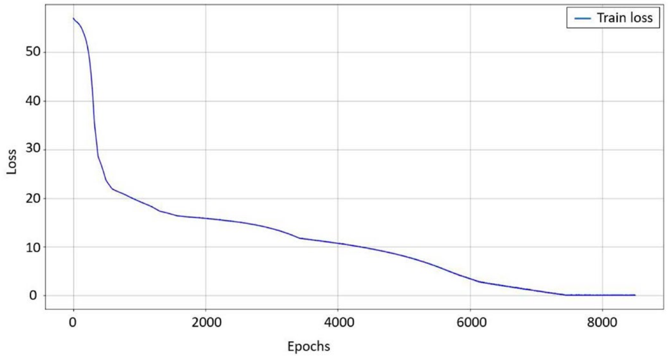


FIGURE 12. **Error trend vs. epochs of the proposed ANN-MLP for experimental data at 1.3 V.**

The procedure is similar to estimating glucose considering the current data set corresponding to the voltage level of -0.455 V. The proposed ANN-MLP architecture has two input layers: the electric current and the molar concentration. For the voltage level of -0.445 V it comprises four hidden layers of 12, 10, 10, and 1 neurons and an output layer corresponding to the glucose level estimation, as shown in Table 4. The activation function of the input layer is a Relu function, the learning rate of the classifier are defined as a constant equal to 0.018.

TABLE 4. Proposed ANN-MLP for training at -0.445 V.

Input Layer	Hidden Layers	Output Layer
Current Measure, Molar Concentration	12-10-10-1	Molar Concentration Estimated

The experimental value of glucose concentration is obtained by feeding the proposed ANN-MLP algorithm trained in real-time with a current value obtained from the electrochemical cell for a specific voltage value while performing cyclic voltammetry (Figure 8). Table 5 provides hyperparameter information for this model.

TABLE 5. Hyperparameters of proposed ANN-MLP model at -0.445 V.

Hyperparameters	Proposed ANN-MLP
Kernel	Linear
Optimizer	Adam
Activation Function	Relu
Error	MAE
Epoch	4500
Degree	1
Learning Rate	0.018
Bias	Boolean

For the -0.445 V case, Table 6 shows the molar concentration data used for the second training of the proposed ANN-MLP at -0.455 V and the results obtained by the algorithm after training. It is observed that there is an average error of 0.0341 between the glucose concentration and the estimated value.

TABLE 6. Glucose concentrations at -0.455 V.

Experimental currents $\times 10^{-8}$ (A)	Glucose concentrations (µM)	Glucose concentration obtained with ANN-MLP (µM)
-4.10	1	1.01725
-4.64	5	5.01976
-5.76	10	10.0197
-6.32	15	15.0241
-7.50	25	25.0483
-7.90	50	50.0492
-8.92	100	100.061

The results for this second training are similar to the previous ones. Figure 13 shows the final correlation between the proposed ANN-MLP training results and the experimental data. It is observed again that the red curve corresponding to the results obtained by the ANN tends to have a good fit with the blue curve corresponding to the input values, which indicates adequate algorithm training.

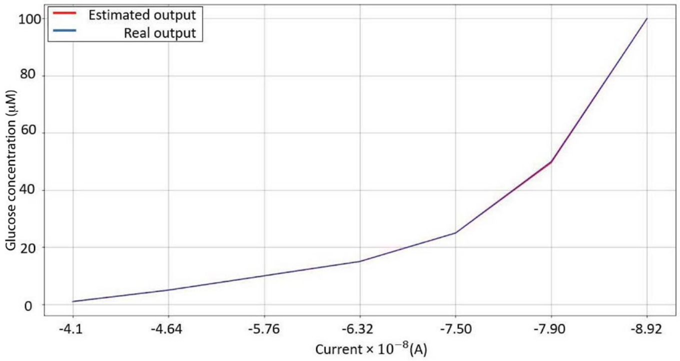


FIGURE 13. Input data and values estimated by the proposed ANN-MLP for experimental data at -0.455 V.

For this second training, the proposed ANN-MLP architecture changes to 10 hidden layers, a learning rate of 0.018, and the limit of the error that tends to zero reached its minimum value of 0.017 at 4500 epochs, as shown in Figure 14.

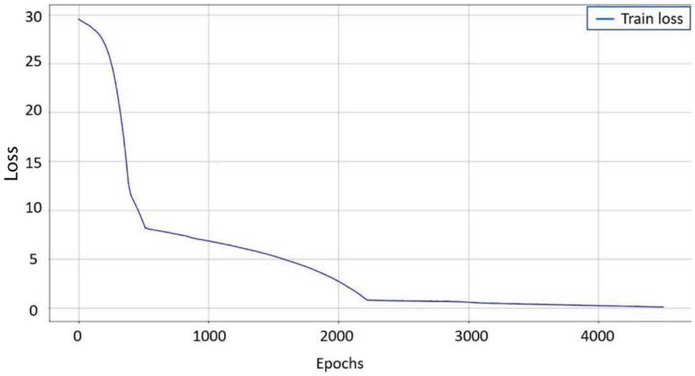


FIGURE 14. **Error trend vs. epochs of the proposed ANN-MLP for experimental data at -0.455 V.**

We can deduce the great advantage of incorporating machine learning algorithms into complex experimental systems from the results obtained. About Figure 10, to validate the machine learning algorithm, the two training sessions of the proposed ANN-MLP are compared, one for the set of current data corresponding to the voltage level of 1.3 V and another for the voltage level of -0.455 V for different glucose concentrations. It is observed that for the case of the voltage at 1.3 V, the current data are relatively easy to identify, unlike those presented for the voltage level of -0.455 V. For the obtained data at the point -0.455 V, the algorithm automatic learning is fed by two current values, one referring to the oxidation and the other to the reduction of glucose, eliminating the one with the lowest magnitude of current, because it is in the middle of the cyclic voltammetry. The increase in glucose presents a more significant negative current. On the other hand, the implementation acquisition and processing of the data at the 1.3 V point appears directly proportional to the positive current and the glucose concentration through a single current value, this being the maximum value, making the implementation of the algorithm more straightforward. The contribution of the system based on the training of the ANN-MLP is independent of the experimental conditions in which the information is found. Therefore, obtaining an ANN architecture according to the experimental system is feasible to facilitate manipulating information and obtaining results. In this experiment, an accuracy of 96.7 % is obtained for the model with 20 hidden layers and 97.8 % for the model with ten hidden layers, which indicates a good level of training of the proposed algorithms.

CONCLUSIONS

It is concluded that it is feasible to synthesize a higher resolution nanostructured sensor for glucose measurement in real-time in aqueous media, aided by machine learning algorithms. The implemented measurement system incorporates nanotechnology through an MWCNT-Cu nanostructured sensor, AI through the training of machine learning algorithms based on ANN-MLP, and digital electronics through ANN-MLP algorithms embedded in an ESP32 SoC . With these elements, a robust system that estimates the glucose level in real time of an aqueous medium in an oxidation-reduction process under cyclic voltammetry is integrated. The proposed method based on the current obtained by cyclic voltammetry saves time in data processing and increases its accuracy, therefore getting glucose concentrations more efficiently. This methodology allows the same CV to have two points of interest to compare and measure the glucose concentration in real-time, at -0.455V and 1.3V. The training of the ANN at the 1.3 V point resulted in an accuracy of 96.7 %, and the model implemented at -0.455 V reached an accuracy of 97.8 %. However, its operating and glucose measurement range was a maximum of 100 μM. For its part, with the data provided from the sensor at 1.3 V, it is possible to distinguish up to 250 μM, which validates the increased resolution of the synthesized sensor. This experimental system based on ANN-MLP training can consider other input variables, such as temperature or voltage scanning speed. Therefore, it is feasible to obtain an ANN architecture according to the experimental system to facilitate the manipulation of information and obtaining results.

From the point of view of technological development, future work considers optimizing the sensor proposed in this work (MWCNT-Cu) without considering the glass carbon electrode as a base. The goal is to develop a more compact and high-accuracy device. From the point of view of basic research, it is considered to improve the machine learning algorithm by including algorithms of the Echo State Network (ESN) type, which is one the most used methods in machine learning to predict complex dynamics, such as chaotic time series [23]. Implementing these algorithms with different topologies can help us predict the experimental behavior of some physical phenomenon in an experimental study under other circumstances.

ETHICAL STATEMENT

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

AUTHOR CONTRIBUTIONS

U.J.T.P. conceptualized and led the research, oversaw the experiments. X.A.A.G. carried out experiments and analyses, O.R.L.B. carried out analyses and validated the results. O.A.A.C. contributed to the design and implementation of A.I. algorithms and validated the results. E.V.V. interpreted and visualized the resulting data and participated in the reviewing and editing of the final manuscript. E.E.G.G. participated in the designed and development of methodology, carried out formal analyses and wrote the original draft. All authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] R. Wilson and A. P. F. Turner, “Glucose oxidase: an ideal enzyme,” Biosens. Bioelectron., vol. 7, no. 3, pp. 165-185, Jan. 1992, doi: https://doi.org/10.1016/0956-5663(92)87013-F

[2] P. N. Bartlett and F. A. Al-Lolage, “There is no evidence to support literature claims of direct electron transfer (DET) for native glucose oxidase (GOx) at carbon nanotubes or graphene,” J. Electroanal. Chem., vol. 819, pp. 26-37, Jun. 2018, doi: https://doi.org/10.1016/j.jelechem.2017.06.021

[3] P. Bollella and L. Gorton, “Enzyme based amperometric biosensors,” Curr. Opin. Electrochem., vol. 10, pp. 157-173, Aug. 2018, doi: https://doi.org/10.1016/j.coelec.2018.06.003

[4] G. Wang, X. He, L. Wang, A. Gu, Y. Huang, B. Fang, B. Geng, X. Zhang, “Non-enzymatic Electrochemical Sensing of Glucose,” Microchim. Acta, vol. 180, pp. 161-186, Feb. 2013, doi: https://doi.org/10.1007/s00604-012-0923-1

[5] D.-W. Hwang, S. Lee, M. Seo, and T. D. Chung. “Recent Advances in Electrochemical Non-enzymatic Glucose Sensors - A Review,” Anal. Chim. Acta, vol. 1033, pp. 1-34, Nov. 2018, doi: https://doi.org/10.1016/j.aca.2018.05.051

[6] X. Kang, Z. Mai, X. Zou, P. Cai, and J. Mo. “A sensitive nonenzymatic glucose sensor in alkaline media with a copper nanocluster/multiwall carbon nanotube-modified glassy carbon electrode,” Anal. Biochem., vol. 363, no. 1, pp. 143-150, Apr. 2007, doi: https://doi.org/10.1016/j.ab.2007.01.003

[7] A. B. Radwan, S. Paramparambath, J.-J. Cabibihan, A. K. Al-Ali, P. Kasak, et al., “Superior Non-Invasive Glucose Sensor using Bimetallic CuNi nanospecies coated mesoporous carbon,” Biosensors, vol. 11, no. 11, art. no. 463, Nov. 2021, doi: https://doi.org/10.3390/bios11110463

[8] C. Tiwari, S. S. Jha, R. Kumar, M. Chhabra, B. D. Malhotra, and A. Dixit, “Exfoliated graphite carbon paper-based flexible nonenzymatic glucose sensor,” Mater Sci. Eng. B, vol. 285, art. no. 115931, Nov. 2022, doi: https://doi.org/10.1016/j.mseb.2022.115931

[9] Q. Fang, H. Wang, X. Wei, Y. Tang, et al., “Cu Aerogels with Sustainable Cu(I)/Cu(II) Redox Cycles for Sensitive Nonenzymatic Glucose Sensing,” Adv. Healthc. Mater., Jun. 2023, doi: https://doi.org/10.1002/adhm.202301073

[10] Y. Zhao, Y. Jiang, Y. Mo, Y. Zhai, J. Liu, A. C. Strzelecki, X. Guo, C. Shan, “Boosting electrochemical catalysis and nonenzymatic sensing toward glucose by Single-Atom PT supported on Cu@CuO Core-Shell nanowires,” Small, vol. 19, no. 18, art. no. 2207240, Jan. 2023, doi: https://doi.org/10.1002/smll.202207240

[11] A. Osorio Villa, “Oxidación electroquímica de glucosa con nanopartículas de oro soportadas en pasta de grafito/carbon (AuNPs/C),” M.S. tesis, CIDETEQ, Tijuana, México, 2017. [Online]. Available: http://cideteq.repositorioinstitucional.mx/jspui/handle/1021/163

[12] L. N. T. Mai, T. H. Tran, Q.-B. Bui, and H.-T. Nhac-Vu, “A novel nanohybrid of gold nanoparticles anchored copper sulfide nanosheets as sensitive sensor for nonenzymatic glucose detection,” Colloids Surf. A: Physicochem. Eng. Asp., vol. 582, art. no. 123936, Dec. 2019, doi: https://doi.org/10.1016/j.colsurfa.2019.123936

[13] T. T. Aun, N. M. Salleh, U. F. M. Ali, and N. S. A. Manan, “Non-Enzymatic glucose sensors involving Copper: An Electrochemical perspective,” Crit. Rev. Anal. Chem., vol. 53, no. 3, pp. 537-593, Sep. 2021, doi: https://doi.org/10.1080/10408347.2021.1967720

[14] G. Martínez-Saucedo, M. Ugalde-Reygadas, J.J. Alcántar Peña, G. Lastra-Medina, J. Márquez-Marín, G. Torres-Delgado, R. Castanedo-Pérez, I.R. Chávez-Urbiola, “Non-enzymatic glucose sensor using nanostructured copper oxide thin films deposited by spray pyrolysis,” Surf. Interfaces, vol. 37, art. no. 102702, Apr. 2023, doi: https://doi.org/10.1016/j.surfin.2023.102702

[15] O. Ghodbane, L.Roué, and D. Bélanger, “Copper electrodeposition on pyrolytic graphite electrodes:Effect of the copper salt on the_ electrodeposition_process,” Electrochim. Acta, vol. 52, no. 19, pp. 5843-5855, May 2007, doi: https://doi.org/10.1016/j.electacta.2007.03.009

[16] X. Zhang, G. Wang, X. Liu, J. Wu, M. Li, J. Gu, H. Liu, and B. Fang, “Different CUO nanostructures: synthesis, characterization, and applications for glucose sensors,” J. Phys. Chem. C, vol. 112, no. 43, pp. 16845-16849, Oct. 2008, doi: https://doi.org/10.1021/jp806985k

[17] J. Yang, W.-D. Zhang, and S. Gunasekaran, “An amperometric non-enzymatic glucose sensor by electrodepositing copper nanocubes onto vertically well-aligned multi-walled carbon_nanotube_ arrays,” Biosens. Bioelectron, vol. 26, no. 1, pp. 279-284, Sep. 2010, doi: https://doi.org/10.1016/j.bios.2010.06.014

[18] A. Singh, A. Sharma, A. Ahmed, A. K. Sundramoorthy, H. Furukawa, S. Arya and A. Khosla, “Recent advances in Electrochemical biosensors: applications, challenges, and future scope,” Biosensors, vol. 11, no. 9, art. no. 336, Sep. 2021, doi: https://doi.org/10.3390/bios11090336

[19] Y. Zhao, H. Zhang, Y. Li, X. Yu, et al., “AI powered electrochemical multi-component detection of insulin and glucose in serum,” Biosens. Bioelectron., vol. 186, art. no. 113291, Aug. 2021, doi: https://doi.org/10.1016/j.bios.2021.113291

[20] A. C. de Sá, A. Cipri, A. González-Calabuig, N. R. Stradiotto, and M. del Valle, “Resolution of galactose, glucose, xylose and mannose in sugarcane bagasse employing a voltammetric electronic tongue formed by metals oxy-hydroxide/MWCNT_modified_electrodes,” Sens. Actuators B Chem., vol. 222, pp. 645-653, Jan._2016, doi: https://doi.org/10.1016/j.snb.2015.08.088

[21] M.-L. Chen and W.-C. Oh, “Synthesis and highly visible-induced photocatalytic activity of CNT-CdSe composite for methylene blue solution,” Nanoscale Res. Lett., vol. 6, art. no. 398, May 2011, doi: https://doi.org/10.1186/1556-276x-6-398

[22] N. Torto, T. Ruzgas, and L. Gorton, “Electrochemical oxidation of mono- and disaccharides at fresh as well as oxidized copper electrodes in alkaline media,” J. Electroanal. Chem., vol 464 no. 2, Mar. 1999, pp. 252-258, doi: https://doi.org/10.1016/S0022-0728(99)00041-8

[23] A. M. Gonzalez-Zapata, L. G. de la Fraga, B. Ovilla-Martinez, E. Tlelo-Cuautle, I. Cruz-Vega, “Enhanced FPGA implementation of Echo State Networks for chaotic time series prediction,” Integration, Vol. 92, pp. 48-57, Sep. 2023, doi: https://doi.org/10.1016/j.vlsi.2023.05.002

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 84 - 104

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.6

E-LOCATION ID: 1386

A Novel Detector of Atypical Beats for Early Diagnosis of Heart Diseases Based on the Stacked Beats Representation of 12-lead ECG

Detector Novedoso de Latidos Atípicos para el Diagnóstico Temprano de Enfermedades Cardíacas Basado en la Representación de Latidos Apilados de un Electrocardiograma de 12 Derivadas

Evgeny Kurmyshev[1], *Deysy Galeana-Pérez*[1]

[1]Centro Universitario de los Lagos, Universidad de Guadalajara, Guadalajara - México

ABSTRACT

We have developed and present in this work a series of algorithms that display a long-duration electrocardiogram (ECG) in a compact form of stacked beats, extracting and visualizing the basic features and facilitating the tedious and time-consuming process of ECG analysis for cardiologists. The expert system based on this representation provides detection of atypical heartbeats, precursors of cardiovascular disease, and their locations in each of the 12 leads. This system was extensively tested with two public databases, MIT-BIH arrhythmia database and China Physiological Signal Challenge (CPSC2018), showing its rapid ECG processing and high efficiency in detecting abnormalities in beat morphology. In particular, tests for atypical beats based on the CPSC2018 database revealed that the set of ECGs marked as normal contains a considerable number of leads with atypical beats. The system is used as a classifier into two classes, normal beats, and atypical beats, the latter being the precursors or indicators of cardiovascular diseases (CVD). It is considered potentially useful for routine studies in groups at high risk of CVD in early stages, as a preventive medicine tool in the public health area. The system allows an intervention of a cardiologist in the intermediate stages of ECG analysis to corroborate the diagnosis in ambiguous cases.

KEYWORDS: atypical beats detection, cardiovascular diseases at an early stage, ECG computer-assisted analysis, preventive medicine, stacked beats representation of ECG

RESUMEN

Desarrollamos y presentamos una serie de algoritmos que muestran un electrocardiograma (ECG) de larga duración en forma compacta de latidos apilados, extrayendo y visualizando características básicas y facilitando el tedioso y lento proceso de análisis de ECG para cardiólogos. El sistema experto basado sobre esta representación provee detección de latidos cardíacos atípicos, precursores de enfermedades cardiovasculares (ECV) y su ubicación en cada uno de las 12 derivadas. Este sistema se probó exhaustivamente con dos bases de datos públicas, base de datos de arritmias del MIT-BIH y China Physiological Signal Challenge (CPSC2018), lo que demostró su rápido procesamiento de ECG y alta eficiencia en la detección de anomalías en la morfología de los latidos. En particular, las pruebas en la base de datos CPSC2018 revelaron que el conjunto de ECG marcados como normales contiene una cantidad considerable de derivadas con latidos atípicos. El sistema se utiliza como clasificador en dos clases, latidos normales y atípicos, siendo estos últimos indicadores de enfermedades cardiovasculares (ECV). Se considera potencialmente útil para estudios de rutina en grupos con alto riesgo de ECV en etapas tempranas, como herramienta de medicina preventiva en el área de salud pública. El sistema permite la intervención del cardiólogo en etapas intermedias del análisis del ECG para corroborar el diagnóstico en casos ambiguos.

PALABRAS CLAVE: detección de latidos atípicos, enfermedades cardiovasculares en etapa temprana, análisis asistido por computadora de ECG, medicina preventiva, representación de latidos apilados de ECG

Corresponding author

TO: **Deysy Galeana-Pérez**
INSTITUTION: **Centro Universitario de los Lagos, Universidad de Guadalajara**
ADDRESS: **Av. Enrique Díaz de León 1144 , Col. Paseos de la Montaña, 47460 Lagos de Moreno, Jalisco**
EMAIL: **deysy.galeana@lagos.udg.mx**

Received:

21 October 2023

Accepted:

26 November 2023

INTRODUCTION

High mortality in the world population (see World Health Organization (WHO) 2023 data) caused by cardiovascular diseases (CVDs), demands the attending early diagnosis of CVD in extensive risk groups of the population. An electrocardiogram (ECG) depicts the electrical activity of the heart and is the most common clinical technique that provides essential information for the diagnosis of CVD. Analysis of ECG and CVD diagnosis is done by a cardiologist mainly. The visual inspection of the ECG by an expert is a delicate and somehow tedious process that takes time. As a consequence, it is difficult to expect a quick, precise, complete, and systematic analysis of 12 leads ECGs of mid and long-term. So, computer-assisted techniques for ECG analysis should be developed and applied to assist a cardiologist in accurate and prompt evaluation of heart conditions [1][2][3].

In the last two decades, several computer-assisted expert systems and classifiers for the characterization of ECGs have been proposed. Methods of ECG analysis and characterization (diagnosis) include three groups of elements: algorithms of preprocessing, characteristic spaces, and classifiers themselves acting on the characteristic space. ECG analysis methods and techniques can differ in any of the three mentioned components. In the first group, thresholding techniques have been used to reduce the amount of information present in the image, and keep only the relevant information [2] and normalization process in time and amplitude used to avoid that feature extraction processes is affected due to physiology, sex and age of the patient, and parameters of the measurement system [4][5][6]. The wavelet transform at this stage is highly appreciated for its properties of multiresolution decomposition of ECG signal and the providing of multi-scale features, to extract the most representative characteristics for being used by a Component Analysis algorithm (PCA), linear discriminant analysis (ICA) or simply as input to a neural network [7][8][9][10][11][12]. Certain frequency ranges have been selected using band-pass, median, average, and finite impulse response (FIR) filters [4][13][14][15][16][17] to detect main energy components of the signal and locate the QRS complexes mainly, and facilitate the segmentation of heartbeats, QRS regions and RR intervals. The accuracy in segmentation is crucial because it determines the performance of the classifier used. In this type of signals, a very important process is the location of R-wave peaks, since it is the basis for measuring intervals, QRS complexes and waves. The Pan-Tompkins algorithm is widely used for peak wave location task [13][18][19].

Second, the algorithms for feature extraction in the ECG analysis used to reduce dimensionality are the algorithm PCA and LDA [6][8][20][21], while the algorithm ICA is used to extract morphological features [8]. Signal transformation methods are widely used, the ECG signals are decomposed into time-frequency representations using Fourier transform and discrete wavelets [5][22], Haar wavelets [2]. Empirical mode decomposition (EMD) is used to detect R-peaks and QRS complex [10][23]. Statistical methods have been applied to obtain statistical characteristics [9][12][17][24][25] to form the feature space that represents the input to the classifier. Third, the most used methods in the classification of ECG signals are machine learning algorithms, such as convolutional neural networks [4][18][19][26][27], also known as deep learning. Those do not need the preprocessing and feature extraction stage because raw data are handled as input to these models. Also, in this approach they have experimented with the data preprocessing using CWT [12] wavelets and IIR filters [15]. They have dabbled applying deep learning with long short-term memory networks [28], recursive neural networks [16][29][30], and random forests [17] to improve mainly the stage of detection and classification of abnormalities in ECG signals. The databases most used in these works is the MIT-BIH [31], followed by the China CPSC2018 database [32].

In deep learning, the feature space is a black box; researchers do not know the selected features of the model. In the approaches without deep learning techniques, combinations of characteristic features have been used. Although they report outstanding results, they do not indicate the location of the cardiac abnormality in the ECG to verify the diagnosis, decreasing the trust of specialists in the computer-aided systems.

To avoid unnecessary repetitions, in this part of the introduction to our article, we suggest reading the introduction in [24] that compiles and describes in a sufficiently complete and concise way the primary knowledge about ECG signals, and explains the necessity of computer method for the cardiovascular diseases (CVD) detection. The recent methods of (approaches to) ECG analysis and classification are also referred to and briefly discussed. For that reason, only a few resumed theses we would like to add to that description to highlight some specific points that will help us to explain our methodology.

First, different CVDs are displayed through the different ECG leads or their combinations; so, for the systematic ECG analysis, it is important to process some specific groups of them, if not all 12 leads, at once. At the early stage of CVD, abnormal heartbeats appear from time to time between normal ones; so, mid to long-term ECGs are needed to be processed to detect them and, due to the amount of information, that has to be computer processed.

Second, ECG signals of different groups of the population depend on many factors such as gender, age, race, stature, and weight among others. In addition, ECGs show a natural time variability. Most of the recent methods are oriented mainly on the detection of specific waveform alterations of heartbeats and deal with the analysis of ECGs of reduced groups of people with notably developed dysfunctions in heart activity. Nevertheless, both in training and validation of CVD classifiers (expert systems, in general) extensive and accurately characterized databases are needed. There are bases accessible for public use but those have some shortness in a systematic database organization and reliable validation by a group of cardiologists.

Third, cardiologists being a conservative group in their methods of disease diagnosis are somehow distrusting to accept technological novelties and new practices in methods of diagnosis if those are not interactive or controlled directly by them. Public health care requires early detection of CVDs at the checking of risk groups of the population to prevent premature death caused by CVD; it means that one has to be able to detect minor changes in beats´ morphology that are not observable easily. So, expert systems for completely automatized and computerized diagnosis are necessary for the preventive checking of public health, but those should be trustful for physicians.

In this work, we develop a computer-assisted expert system for the detection of atypical heartbeats. That is aimed at the early detection of CVD when minor changes in beats´ morphology are difficult to observe. At this stage, we focus on the computer-assisted classification of any of the 12 ECG leads into two classes, ECGs of healthy persons and persons with an alert of heart disease. The final specific diagnosis of risk in CVD is due to a cardiologist or by a higher-level classifier. The system is designed to be open to the intermediate data validation by a cardiologist and his participation in the process, if necessary. To this end, we propose a compact stacked beats representation of 12 lead ECG. Each lead of the raw ECG signal is segmented into the heartbeats using RR-interval statistics and P-wave detection, followed by the correction of the isoelectric and baseline wander. Then, the heartbeats are stacked matching the peaks of R-wave, and the mean of the beats is calculated.

The latter is considered as the reference beat of ECG lead. All main morphological characteristics of ECG such as HRV, intervals, segments, and amplitudes of heartbeat waves are calculated. Statistical features of morphological differences between beats of ECG lead and the respective reference beat are calculated; the latter is used for the atypical beats detection.

The work is organized as follows. In section 2 we describe the basic principles and pseudocodes of algorithms used for the ECG preprocessing and its partition in beats and the compact representation of ECG signals by stacked heartbeats. Based on this representation, we propose the characteristic metric space for the heartbeats analysis and classification into two classes, normal and atypical beats. Section 3 briefly describes the design of experiments for the validation of expert system algorithms and datasets used to this end. Section 4 presents the results of the computer simulation of the expert system on the China Physiological Signal Challenge 2018 database and MIT-BIH arrhythmia database. Conclusions and discussion are given in section 5, followed by references in section 6.

MATERIALS AND METHODS

Basic principles and algorithms

Similarity of heartbeats and natural cardiac variability

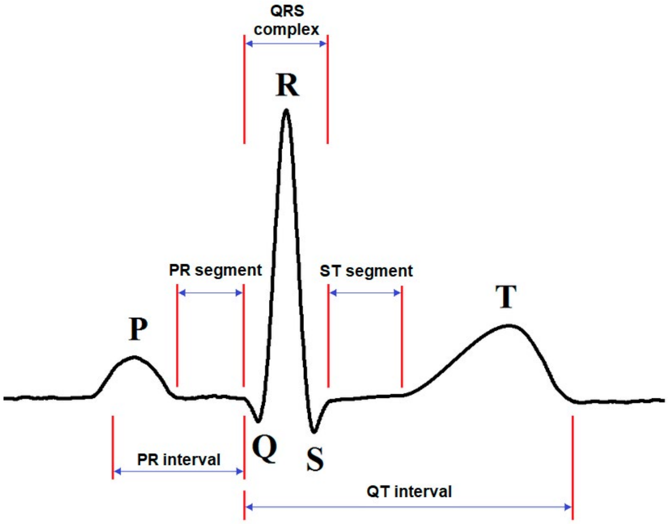


FIGURE 1. Schematic depiction of a standard normal ECG signal [36]

An electrocardiogram (ECG) is the most widely used tool for detecting cardiovascular diseases [33]. ECG maps the heart's electrical activity as a sequence of events associated with the generation and conduction of electrical impulses in the chambers of the heart [34]. The complete cardiac cycle of a standard D1 or D2 lead of ECG demonstrates the sequence of P, QRS, T, and U-waves along with the respective intervals and segments between them [35]. The U-wave located in the ECG after T-wave is generally not presented, due to its dependency on the T-wave amplitude and heart rate [1] [29]. The P-wave is associated with atrial depolarization; the QRS complex, most common in each cardiac cycle, represents the depolarization of the ventricles; and the T-wave shows the repolarization of the ventricles [2]. The QRS complex is considered a crucial waveform in the ECG which consists of three deflection points: Q, R, and S. R-peak is the most prominent peak in the ECG. Changes in the characteristics of these waves and intervals may indicate the presence of arrhythmia in the heart. So, accurate interpretation of ECG is critical in the detection of cardiac diseases [3]. A typical representation of the D1 or D2 standard lead of the ECG signal is given in Figure 1.

The ideal ECG signal is expected to be a periodic function of a certain morphology, but it is affected by many factors such as natural cardiac variability, personal physiology, electric contacts colocation, and apparatus noise among others. In addition, the ECG of a person is so individual that can be considered as a part of a biometric identification system [37]. A group of the 12 leads of ECG signal is used to be recorded simultaneously in each cardiac test. So, many features have to be taken into account to process, analyze, and classify an ECG. The algorithms presented below constitute our computational system for the processing and classification of ECG heartbeats in sets of normal and atypical ones. The method is based on a compact representation of each ECG lead through its processing and segmentation, and the stacking of heartbeats of each lead in a special way.

R and P wave peak detection

For the segmentation of an ECG into heartbeats, we start with the R-peaks detection in the DI or DII lead, the latter have similar morphology. We use a rather simple semi-empirical algorithm for R-wave peak detection.

The steps of the algorithm to detect an R-wave peak are as follows:

1. The array of the DI or DII lead of ECG is loaded into the system, where the notation R is used for the voltage and T for the corresponding instant of time of the signal.
2. Find the absolute maximum R_{max} in the first 10 seconds subarray of DI or DII.
3. Set the threshold $R_{th}=0.5R_{max}$.
4. In the complete array find all the values R_j and corresponding instants T_j such that $R_j > R_{th}$, and $(T_{j+1} - T_j) < SR/5$, forming the set of pairs $\{(R_j,T_j)_i : i = 1,2, \ldots ,N\}$ that belong to the i-th R-wave; N is the number of heartbeats and SR is the sampling rate.
5. Find the absolute maximum of voltage R in each subset $(R_j, T_j)_i$ that is the peak of the i-th R-wave.
6. If necessary for the revision, plot ECG with R-peaks detected (see Figure 2).

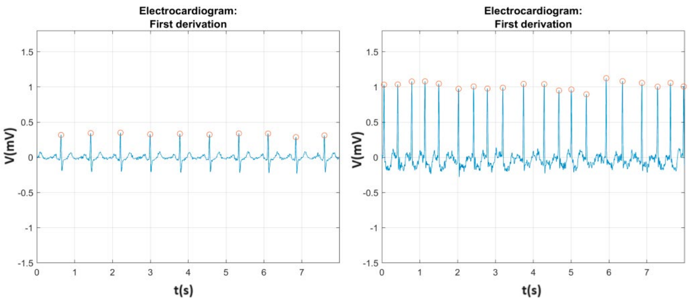


FIGURE 2. The detected R-wave peaks. Normal ECG, left; abnormal ECG, right.

Figure 2 shows an example of R-wave peaks detected in a normal and abnormal ECG. The plot of ECG with the detected R-peaks can be displayed on the monitor for the convenience of analysis by the cardiologist.

Heart rate calculation and depiction

RR intervals are one of the basic features of ECG that provide important information to the physician about the heart rate and its variability. After obtaining the location of all R-peaks d s the time intervals ΔT_i between the current and next R-peaks:

$$\Delta T_i = T_{i+1} - T_i \tag{1}$$

The heart rate (number of beats per minute) is calculated as the inverse of RR intervals:

$$HR_i = \frac{60}{\Delta T_i} \ bpm \tag{2}$$

To characterized the heartbeat variability (HBV) the standard deviation is also calculated:

$$\sigma = \sqrt{E[(HR_i - HR)^2]} \tag{3}$$

where $HR = E[HR_i]$ is the mean of heartbeat rate. In order to facilitate the analysis of ECG for the cardiologist we display HBV in a dot plot. Figure 3 shows the result of the calculation of HBV in two ECGs, one normal and the other abnormal.

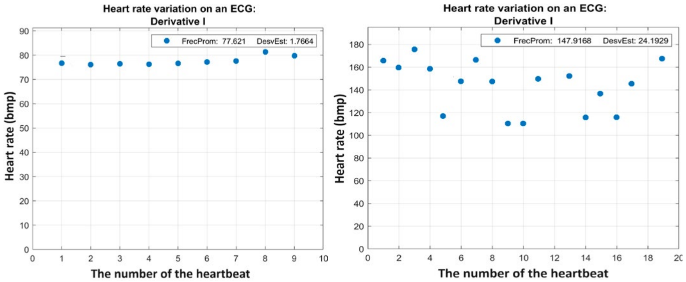


FIGURE 3. Depiction of heartbeat rate variability. Normal ECG, left; abnormal ECG, right.

ECG segmentation and heartbeat stacking

Localization of partition points

One of the essential keys in our compact representation of ECG is the location of partition points to divide the ECG into separate heartbeats; this is an especially subtle step when the heartbeats are atypical. We proceed the following way. The interval $(T_i - 10/SR) < t_P(i) < (T_i - 20/SR)$ is used to search for the location of P wave maximum preceding the detected R peak at point T_i; where *SR* is the sampling rate of ECG signal, and the magnitudes $t_P(i)$ and T_i are instants of corresponding samples. After the location $t_P(i)$ of the P_i peak is calculated, we put the partition point at the distance

$$\Delta_i = \Delta_{st} \cdot \frac{T_i}{T_{st}} \qquad (4)$$

left to the point t_{Pi}. The interval Δ_{st}=0.1 ms was suggested by a cardiologist with the argument for the best positioning of partition point in ECG with the standard (typical) RR interval T_{st}= 60/72 for the HR at 72 bpm (72 beats per 60 seconds); that is done to prevent the excessive cutting of the baseline after the T-wave. In terms of current heartbeat rate HR_i, we get from Equation (4) the relation.

$$\Delta_i = \Delta_{st} \cdot \frac{72}{HR_i} \qquad (5)$$

The latter adjustment, Equation (4) or (5), was made to compensate for the effect of HRV. After being segmented into heartbeats, the array of original ECG is transformed into the matrix, each line of which is a single heartbeat data string. Then, we stack all heartbeats of the ECG for the visual inspection and the following processing. Simple stacking looks somehow unadjusted, and isoelectric-baseline wander is also observed, especially for the abnormal ECGs. To correct the above-mentioned imperfections, we implement isoelectric-baseline correction and then restack heartbeats, matching the R peaks of all plots.

Correction of isoelectric and baseline wander

The isoelectric baseline is defined in the literature in a variety of slightly different ways. In a normal ECG, it denotes resting membrane potentials in the TP interval; those potentials should be ideally equal to zero. So, the beginning of each P wave and the T wave end of the heartbeat has to lie on the straight line of zero potential. Going from this consideration we propose the following algorithm for the isoelectric baseline correction in each heartbeat of segmented ECG.

First, the tangent of the original heartbeat baseline is calculated,

$$\tan\alpha = \frac{V_f - V_{in}}{t_f - t_{in}} \qquad (6)$$

where V_{in} and V_f are the potentials at the initial and final points of a heartbeat, t_f and t_{in} are corresponding instants of time. Then, the corrected value *V* of heartbeat potential for the current instant t is calculated as.

$$V = (V_{or} - (t - t_{in})\tan\alpha) - V_{in} \qquad (7)$$

This way the original, usually deflected heartbeat is settled on a zero-valued isoelectric baseline. Then the stacking of the baseline corrected heartbeats follows. That is the key element of ECG representation and a window for the cardiologist to check the course of automatic ECG processing and analysis.

Stacking of heartbeats with R peaks' matching

The matrix of baseline corrected heartbeats is taken as the input. We find the absolute value maximum of the matrix and then match the peak positions of the R-wave maxima of all heartbeats in a new matrix. To complete the stacked heartbeats representation of ECG we calculate the mean heartbeat that carries the most important morphological information of the ECG lead and simplify the visual analysis of ECG; in addition, it is used as the reference for the later classification of beats. Note that the averaging removes high-frequency noise from the mean heartbeat, so it permits the use of raw ECGs without previous filtering. Figure 4 shows the above-indicated steps graphically: the correction of the isoelectric and baseline to the averaging of beats and identification of the P wave, QRS complex, T wave, and the PP, RR, PR, and QT segments, and their duration in milliseconds.

The set of 12 leads is processed for the ECG to be interpreted systematically. The mean of heartbeats is used as the reference beat of each lead and facilitates visual analysis and interpretation of heartbeats in ECG leads; which is used by cardiologists for computer-assisted diagnosis. Since the stacked heartbeats representation of ECG is based essentially on the P and ORS waves, the former can be efficiently implemented only for the ECGs that contain the DI or DII lead. Each lead of a normal ECG has a standard form of a heartbeat. When quite several heartbeats of ECG lead are distorted significantly or HR is highly variable, the reference beat can differ notably from the standard beat of the lead and, in addition, the ECG partition into beats can fail or produce segments of multiple beats.

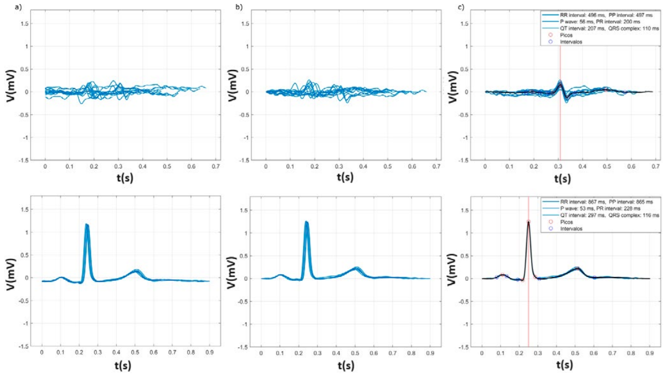


FIGURE 4. The stacked heartbeats representation of DI lead, column: a) isoelectric line correction, b) isoelectric and baseline correction, c) the mean of the stacked heartbeats matching R-wave peaks (black line) and the calculation calculus of main elements of ECG. The first line of graphs is for an abnormal ECG signal, A0005 from CPSC2018 (50 beats); the second one is for a normal ECG signal, A0016 from CPSC2018 (25 beats)

Computer-assisted analysis of ECG for early detection of heart disease

Morphological distortion of heartbeats and abrupt HR variations detected in the ECG signal is the indicator of CVD. Since abnormal heartbeats appear in ECG occasionally, especially in the early stages of CVD, they can be detected by comparing them with the reference heartbeat; we take the mean heartbeat as the reference. To consider heartbeat morphology as the feature space in the mathematical sense for the heartbeats classification one needs to define a measure to quantify the morphological distortions of a heartbeat. We propose to use the Hamming distance between the array of a current heartbeat of interest and the mean heartbeat of the same lead of ECG. The distance between the mean heartbeat V_{mean} and a given heartbeat V_i of one lead is calculated by the formula,

$$d_i = d(V_i, V_{mean}) = \frac{1}{N}\sum_n |V_i(n) - V_{mean}(n)| \quad (8)$$

where V_i *(n)* is the n-th sample of the i-th heartbeat V_i, and N is the dimension of the array. The defined measure permits us to construct the detector of atypical beats and the classifier of heartbeats into 2.5 classes; those differ from each other in the algorithms used.

The detector acts as follows. In a given lead of ECG, the distance $d_i = d\,(V_i, V_{mean})$ of each heartbeat from the mean heartbeat, Equation (8), and the mean of distances d, and the standard deviation σ_d of heartbeats' distances are calculated,

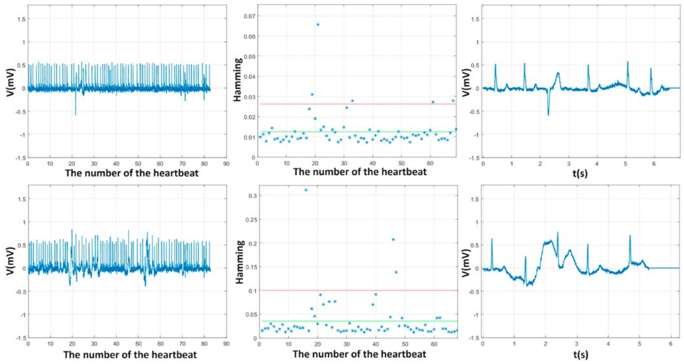
$$d = E[d_i] = \frac{1}{M}\sum_{i=1}^{M} d_i \quad (9)$$

$$\sigma_d = \sqrt{E[(d_i - d)^2]} \quad (10)$$

The standard deviation is the characteristic parameter for the detector and classifier. We distinguish the beats of the two types, normal and atypical beats. The normal ones are close to the reference beat in the characteristic space, and the atypical beats are distant from the prototype at a distance larger than the threshold value (see Equation (11)). The abnormal beats are a subset of the class of atypical beats, usually, they are at distances notably larger than the threshold value. The detector assigns a heartbeat i into the class of atypical beats if

$$d_i \geq d + A\sigma_d \quad (11)$$

where A is the adjusting parameter of the classifier. The less the latter the more sensitive to the morphological distortion of beats is the classifier. As an example, Figure 5 shows the distribution of distances and the separation of atypical beats from the normal (typical) ones; the classifier's threshold $(d + A\sigma_d)$ is the red line at $A = 2.5$.



a) b) c)

FIGURE 5. Example of separation of atypical beats in abnormal ECG, column: a) original DI and DII lead of A0713 register from CPSC2018; b) Distribution of heartbeats' Hamming distance (the mean of distances – green line, separator of atypical heartbeats- red line); c) depiction of one of the atypical beats.

Setting the adjusting parameter A at a desired value (see Equation 11), the detector can be used for the rapid separation of atypical heartbeats in mid- and long-term ECG leads for the posterior visual analysis by a cardiologist or the computer classification of beats.

The borderline between the classes of normal and abnormal beats can't be defined; instead, there is a border band class of warning beats that can be considered as belonging to the two classes in the sense of fuzzy logic. Starting from criterion (11), we propose the following algorithm for the detection of atypical or warning heartbeats. The beats of a lead, most distant from the reference heartbeat, are indicative for seeing if they are abnormal or not. At this stage of development, the expert system is used in two ways. Mode 1 detects a given number of the most distant (atypical) beats from the reference one. Mode 2 detects all heartbeats for the calculated σd and a fixed adjusting parameter A (see Equations (10) and (11)).

In mode 1, we set for the detector the desired number of the most distant beats, for example, M=3. Next, the algorithm runs a cycle for a descending set of values of A from 9 to 1, with step 0.1, counting the number of atypical beats detected at each value of A. When the number of detected atypical beats equals or exceeds 3, then the cycle is stopped and M vicinities of atypical beats (2 previous and 2 posterior beats are included) are displayed for the revision by a cardiologist. The pseudocode of the algorithm is as follows:

```
1.  Choose an ECG lead to be processed, say $L$ (=1,2,...,12)
2.  Set the desirable number ($M$=3 is recommended) of the atypical beats to be analyzed
3.  Run processes for the partition of $L$ into the beats
4.  Calculate the matrix of ECG lead representation through the stacking of beats
5.  Calculate the mean of the beats as the reference one
6.  Calculate the Hamming distance between each beat and the reference beat
7.  Calculate the standard deviation $\sigma d$
8.  For $i$=0,1,2,...,80:
9.  Calculate $A_i$=9 - 0.1$i$
10. Calculate threshold $d+A_i\sigma d$
11. Run classifier at $A_i$ for L and calculate the number $m_i$ of atypical beats, see Equation (11)
12. If $m_i \geq M=3$ then display the vicinities of atypical beats
13. End
```

The pseudocode of Mode 2 is slightly different from the previous one:

```
1.  Choose an ECG lead to be processed, say $L$ (=1,2,...,12)
2.  Run processes for the partition of $L$ into the beats
```

3. Calculate the matrix of ECG lead representation through the stacking of beats
4. Calculate the mean of the beats as the reference one
5. Calculate the Hamming distance between each beat and the reference beat
6. Calculate the standard deviation σd
7. Set a desirable value of adjusting parameter A
8. Calculate threshold d+Aσd
9. Run classifier for the lead L at a given value of A (see Equation (11)) to detect atypical beats
10. Display the vicinities of atypical beats
11. End

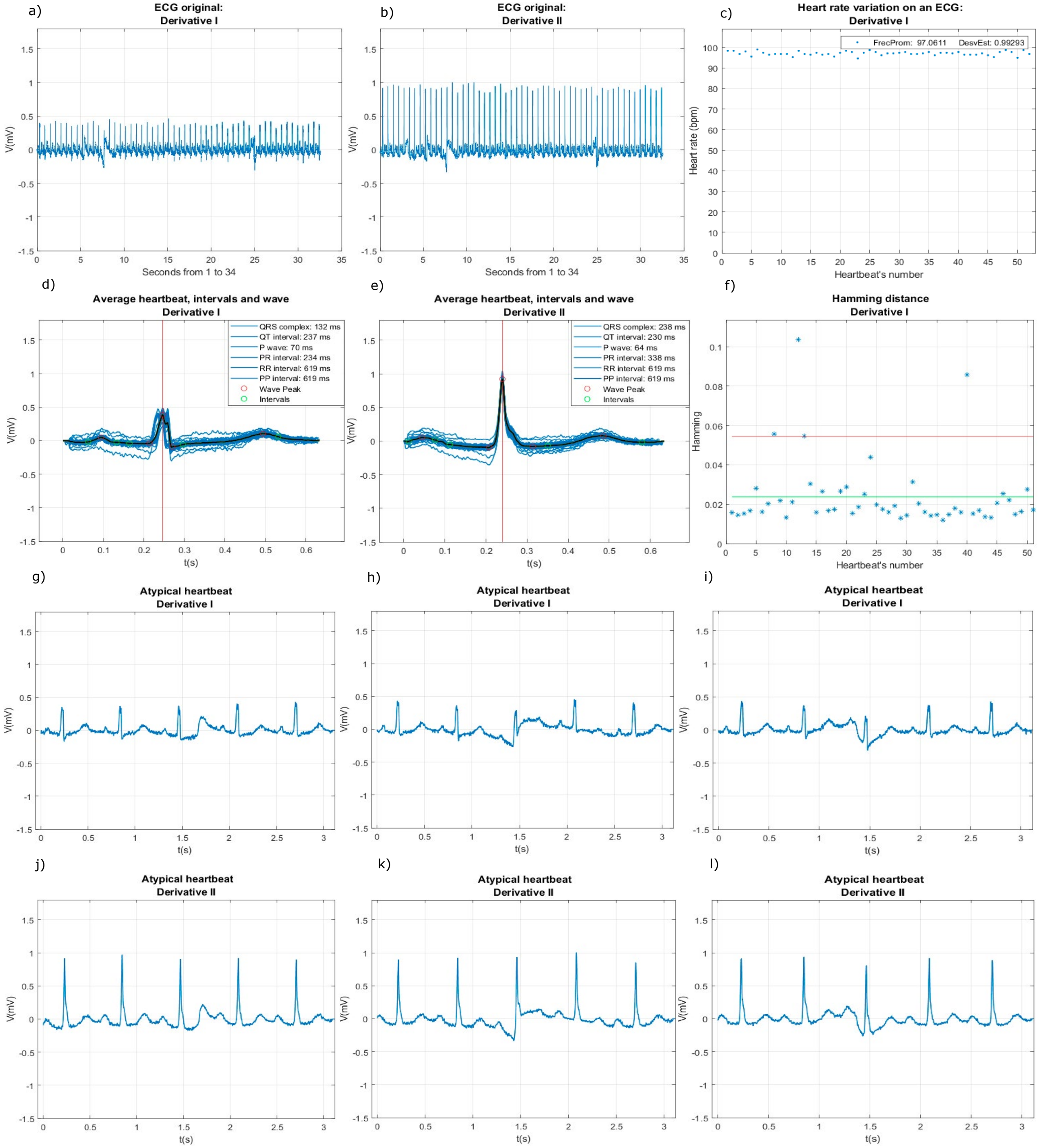


FIGURE 6. **Typical stages of ECG leads processing; example of DI and DII leads from A1437 register of CPSC2028 database.**

The detected atypical heartbeats can be displayed along with the neighboring beats on the computer screen for visual inspection by a cardiologist. As an example, we present in Figure 6 the results of ECG processing by expert system executed in mode 2, for DI and DII leads from A1437 register of CPSC2018 database. Slides a) and b) show original signals, followed by the heartbeat rate in slide c). Slides d) and e) depict stacked heartbeats representation followed by graphical representation of the classifier execution in slide f). The rest of slides demonstrate the vicinities of the detected atypical heartbeats. The stacked heartbeats representation of DI lead makes evident the splitting of the R peaks. If no one of the detected atypical beats can be considered abnormal, then the ECG lead is considered to be normal. The detector of atypical beats is auto-sufficient; it does not need any external training set of ECGs because both the reference heartbeat and the characteristic parameter σ_d are calculated using the heartbeats of the lead of interest.

Datasets and design of experiments

To facilitate the use of the algorithms we develop the interface with the following options:

1. Load a desired digitalized lead of ECG into the expert system
2. Display the original ECG - for visual revision and analysis, if necessary
3. Choose the interval of ECG to be processed
4. Detect and display the R peaks - for visual revision and analysis, if necessary
5. Divide the ECG into heartbeats
6. Calculate and display (the heartbeat rate) HR and its standard deviation - for visual revision and analysis, if necessary
7. Correct the isoelectric and baseline of each heartbeat
8. Stack the processed beats, calculate their mean, and display them for compact representation of ECG - for visual revision and analysis, if necessary
9. Calculate Hamming distances between each heartbeat and the mean of the heartbeats (reference heartbeat), and the standard deviation of this distribution
10. Apply the detector of atypical heartbeats and display the results, with the vicinities of atypical heartbeats being included - for visual revision and analysis by a specialist for the final diagnosis, if necessary

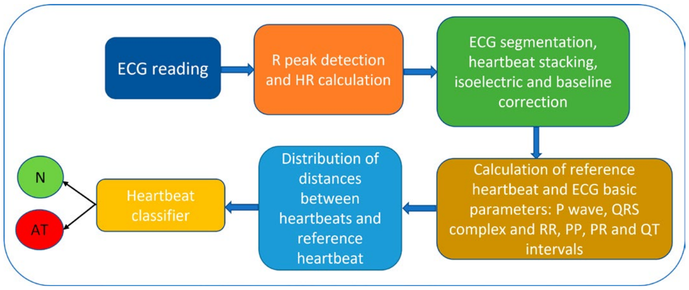


FIGURE 7. **Flow chart of expert system**

The functionality of the ECG stacked beats representation and the performance of the atypical heartbeat detector were evaluated on the two ECG databases: China Physiological Signal Challenge in 2018 (CPSC2018), and the MIT-BIH public database. The CPSC2018 database consists of two subsets, 3,699 males and 3,178 females (6,877, in total) 12 lead records lasting from 6 seconds to 60 seconds; each recording was sampled at 500 Hz. The database contains 918 ECGs marked as normal and 5959 ones labeled as abnormal; neither the location of abnormal beats in the ECG nor their number is indicated. The public dataset from the MIT-BIH Arrhythmia Database contains 48 half-hour excerpts of two-channel ambulatory ECG recordings obtained from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979. Those are sampled at the rate of 360 Hz. The MIT-BIH database has announced the number of abnormal beats for each ECG both for the five first minutes and the rest of the ECG record. From the description of the database, it is not clear if the number of abnormal beats is attributed to the two leads of the record or one of the two leads; in the last case, the user does not know to which of the two leads the number of abnormal beats is attributed. Note that the numbers of abnormal heartbeats of different leads should not be equal.

Every ECG of those databases was processed by the expert system to see first its ability for the compact stacked heartbeats representation of ECG. Then the detection of atypical beats was implemented and analyzed. All ECGs went through the same processing: ECG segmentation in heartbeats, isoelectric and baseline correction, and stacking of the heartbeats with matching on R peaks.

The processing was used for feature extraction such as: 1. HR and its standard deviation; 2. reference heartbeat calculation of each ECG lead; 3. Calculation of Hamming distances and their distribution; and finally, 4. Detection and displaying of atypical heartbeats. The results of each stage of ECG processing and analysis can be displayed on the computer screen if necessary for visual revision and validation. The flow chart in Figure 7 resumes the overall execution of the expert system.

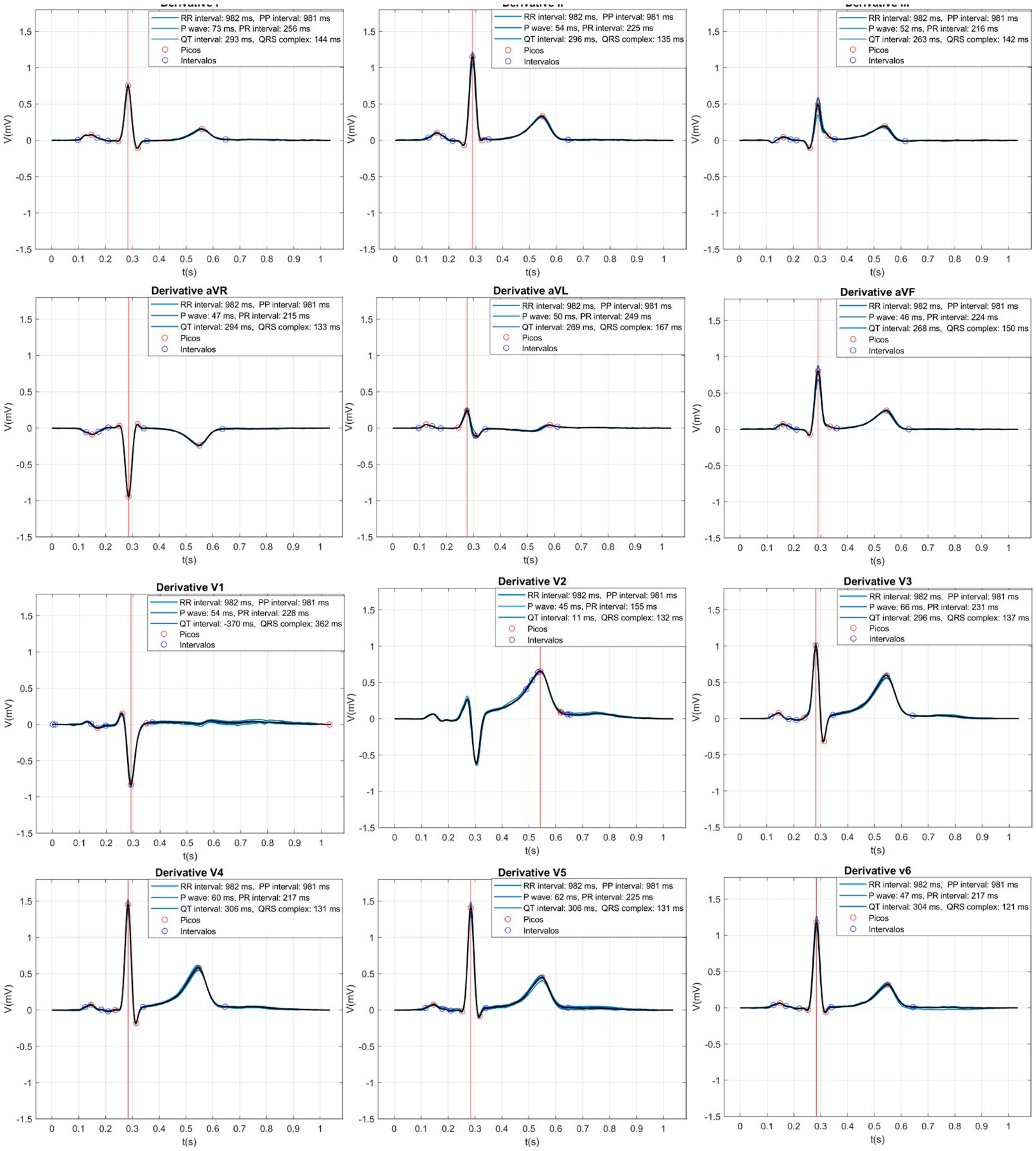


FIGURE 8. **The stacked beats representation of 12-Lead ECG signals registered in CPSC2018 as A0030 normal signal. The black line represents the average of the beats of each derivative, the vertical red line shows the point of the R peaks matching.**

RESULTS AND DISCUSSION

Result of computer simulation

China Physiological Signal Challenge 2018 database

The expert system was evaluated on the CPSC2018 database to be sure it can process 12-lead ECG records and detect atypical heartbeats. As a drawback, the database has relatively short records of the heart's electrical activity, 6 to 60 seconds. To show that the proposed stacked beats representation works for both normal and abnormal ECGs, in the following graphs we give examples of its application. Figure 8 shows a typical stacked heartbeats representation of 12 lead normal ECG signals. In this representation it is easy to see that the morphological dispersion of the beats in each derivative is minimal, which confirms the classification of ECG normality in CPSC2018. Compared to the normal ECGs Figure 9 shows a larger scatter in the form of heartbeats, which is consistent with the ECG labeling as abnormal in the Chinese CPSC2018 database. Therefore, this representation allows a cardiologist to execute a visual diagnosis quickly and efficiently.

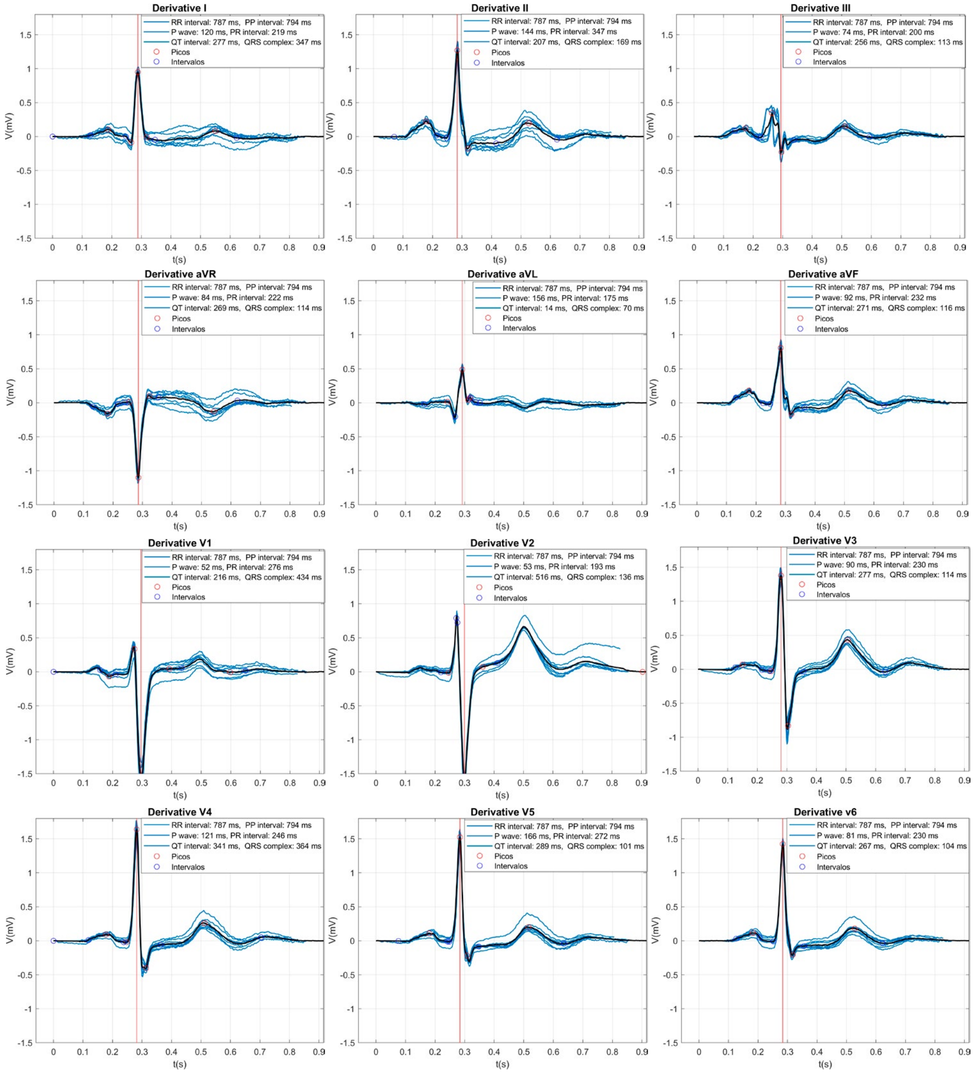


FIGURE 9. The stacked beats representation of 12-Lead ECG signals registered in CPSC2018 as A0013 abnormal signal. The black line represents the average of the beats of each derivative, the vertical red line shows the point of the R peaks matching.

Our system processed 6,837 of the 6,877 ECG records, running all processes. Of the 40 ECG records with the interrupted process, 37 correspond to abnormal ECGs, and 3 (A0945, A1035, A1 582) to the normal ECGs. Figure 10 shows electrocardiograms with the interrupted process labeled in the database as normal ECG. The process was interrupted because a) the signal has peaks much larger than the rest of the R peaks; this irregularity of signal is present in Figure 10a; b) the voltage alteration at the beginning of the signal is extremely large, see Figure 10b; c) irregular R wave peaks, less or equal in size to T wave peaks. So, the ECG leads with interrupted preprocessing have parts of the signal so much distorted that cannot be considered normal and have to be analyzed by a specialist to verify if they were misclassified or badly registered. As for the 37 abnormal leads with the interrupted process, they all have strongly distorted beat morphology. So, the interruption of the ECG preprocessing serves as the coarse filter for the abnormalities in heartbeats.

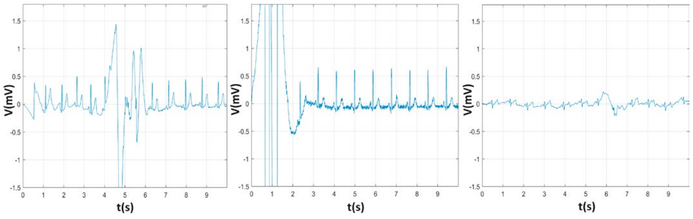


FIGURE 10. Three ECG signals of the DI lead labeled as normal in the CPSC2018 database that failed to be segmented in heartbeats.

One of the main goals of our computer algorithms is the early detection of CVD, in other words, the detection of relatively small morphological changes in ECG beats. For that reason, the expert system has to be evaluated first on a confident set of normal ECGs. The public CPSC2018 database contains 918 ECGs of 12 lead signals that are labeled as normal. Before the systematic testing of the expert system, we processed and analyzed a number of them in detail. To our surprise we found that a noticeable number of ECG leads contain beats that are suspicious or cannot be considered normal; this finding forced us to change the traditional way of classifier validation. To evaluate the functionality of the expert system and to find the atypical beats in the ECG leads labeled as normal in the CPSC2018 database, a variety of computer experiments were implemented.

The measure of morphological variability of heartbeats in a lead is the standard deviation σ_d of heartbeats' distances, eq. (10). So, first we have analyzed the changes of σ_d in the group of leads of the same type of different ECG and each set of 12 leads of a group of ECGs. Fifty DI leads, chosen at random from the set of 915 ECGs labeled as normal, were processed. The standard deviation σ_d of Hamming distances d_i were calculated for each DI lead; we found that the dispersion σ_d is varied in the wide range from 0.0011 up to 0.0129. Then, the four 12 lead ECGs (A0166, A4790, A5094, A4512), were processed also. We found that the standard deviation σ_d varies significantly from one lead to another. For example, the first six leads of A0166 get σ_d in the range from 0.0025 to 0.0044, while σ_d for the rest of the leads is significantly higher and ranges from 0.0136 to 0.2243, the latter for the DV6 of A0166. This result accords with the cardiologists' experience that different combinations of leads have to be analyzed for the assertive CVD diagnosis.

Our expert system is aimed to assist a cardiologist in the analysis and diagnosis of mid-long time ECG signals, at the requirement that the system has to be interactive. To simulate and evaluate the functioning of the system, two sets of twenty 12 lead ECGs labeled as normal were randomly drawn from the CPSC2018 database (480 leads in total). Set #1 consists of registers A0166, A0173, A0177, A0221, A0365, A1971, A2014, A2473, A2507, A2591, A2773, A3975, A4286, A4512, A4550, A4790, A5094, A5940, A6306 and A6832. Registers A0141, A0281, A0588, A0690, A0774, A0985, A1208, A1437, A1463, A1534, A2043, A2484, A2770, A3424, A3751, A4036, A4366, A4867, A5004 and A6560 constitute the set #2. Each of the 480 leads was processed to obtain the stacked beats representation. Then, the reference heartbeat was calculated for each lead and the distribution of Hamming distances of beats from the reference one. The M=3 most distant beats of each lead were detected using criterion (11) and algorithm of Mode 1; the vicinities of the detected beats were depicted on a screen and visually corroborated as being normal or atypical ones. The first set of twenty 12 lead ECG signals labeled as normal has shown that 87 leads of the 240 ones are normal, while in the rest of the 240 leads were detected 77 alert and 76 abnormal heartbeats. The second set of twenty 12 lead ECG signals labeled as normal has shown that 133 leads of the 240 ones are normal, while in the rest of the 240 leads were detected 68 alert and 39 abnormal heartbeats.

The detector of atypical beats at the adjusting parameter A=3 was applied in Mode #2 to a wider set of ECGs labeled as normal in the CPSC2018 database, showing several irregularities in heartbeat morphology. As an example, we present in Figure 8 the result of the processing of DI and DII leads in the A1437 register. Panels a) and b) give the general view of the ECG leads, which can be seen also by sections for more details. Panel c) represents the HRV and its distribution in time. Panels d) and e) show the stacked beats representation of the leads with the reference (mean) heartbeat that helps to see the defects of atypical beats.

The reference heartbeat in panel d) shows the flection of R-wave emerged from the double peaks seen in the spanning of individual R-waves. The deflection in the R-wave is also observed in the reference beat of DII. Panel f) shows the morphological closeness of beats or their difference from the reference beat in the characteristic space with the Hamming distance metric for the DI lead. In the case of A1437, panel f) is followed by panels g), h), and i) for DI, and j), k), and l) for DII, which show sections of the ECG leads with notably deformed heartbeats. The analysis of ECG leads from the CPSC2018 database has shown the efficiency of the proposed detector (classifier) of atypical heartbeats, even when the waveform distortions are minors and not visible easily; that is important for the detection of precursors of a CVD at an early stage.

An ECG with elevated T-waves presents a challenge for the segmentation in heartbeats because T- and R-wave picks can be confused by the algorithms. We develop an efficient algorithm to overcome this difficulty. As an example of its application, we present the DI lead of the A0281 ECG register in Figure 11; the ECG was labeled as abnormal and diagnosed as an elevated ST segment. Having the mean beat of the stacked representation as the reference, one easily observes the noise and distortion in the T wave of beats in panel c) of Figure 11.

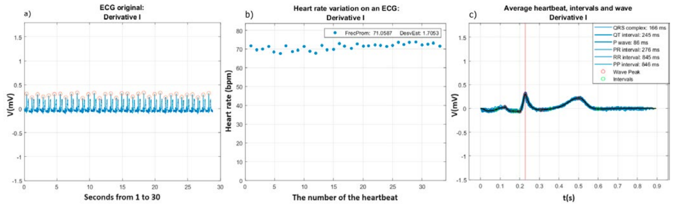


FIGURE 11. DI lead of A0281 ECG register with the elevated ST-segment: a) R-wave peaks detection, b) HRV, and c) stacked beat representation with the reference heartbeat (black).

When ECG is examined by the classifier at a desired value of adjusting parameter A (we suggest A=3), the atypical beats ($d_i \geq d + A\sigma_d$) are detected and displayed for the posterior detailed analysis by a cardiologist, to decide if there are abnormal ones. Finally, our detector has identified the abnormal beats in all processed ECGs, marked as abnormal in the CPSC2018 database. These trials show that the developed algorithms are versatile tools for the ECG analysis and detection of atypical and abnormal beats.

MIT-BIH arrhythmia database

The database has different pairs of leads and each beat of a lead is labeled as Normal (N), Left bundle branch block (L), Right bundle branch block (R), Atrial premature (A), Aberrated atrial premature (a), Nodal (junctional) premature(J), Supraventricular premature (S), Premature ventricular contraction (V), Fusion of ventricular and normal (F), and so on [31]. Since the ECGs of the database were pre-classified as abnormal, the database is suitable for testing our algorithms as the detector of atypical heartbeats. For that purpose, we chose three half-hour excerpts of two-channel ECG recordings of the MIT-BIH database, # 100, # 106, # 112, and set the adjusting parameter at the value A=3.5 (see Equation 11). The six leads were processed following the Mode #2 algorithm.

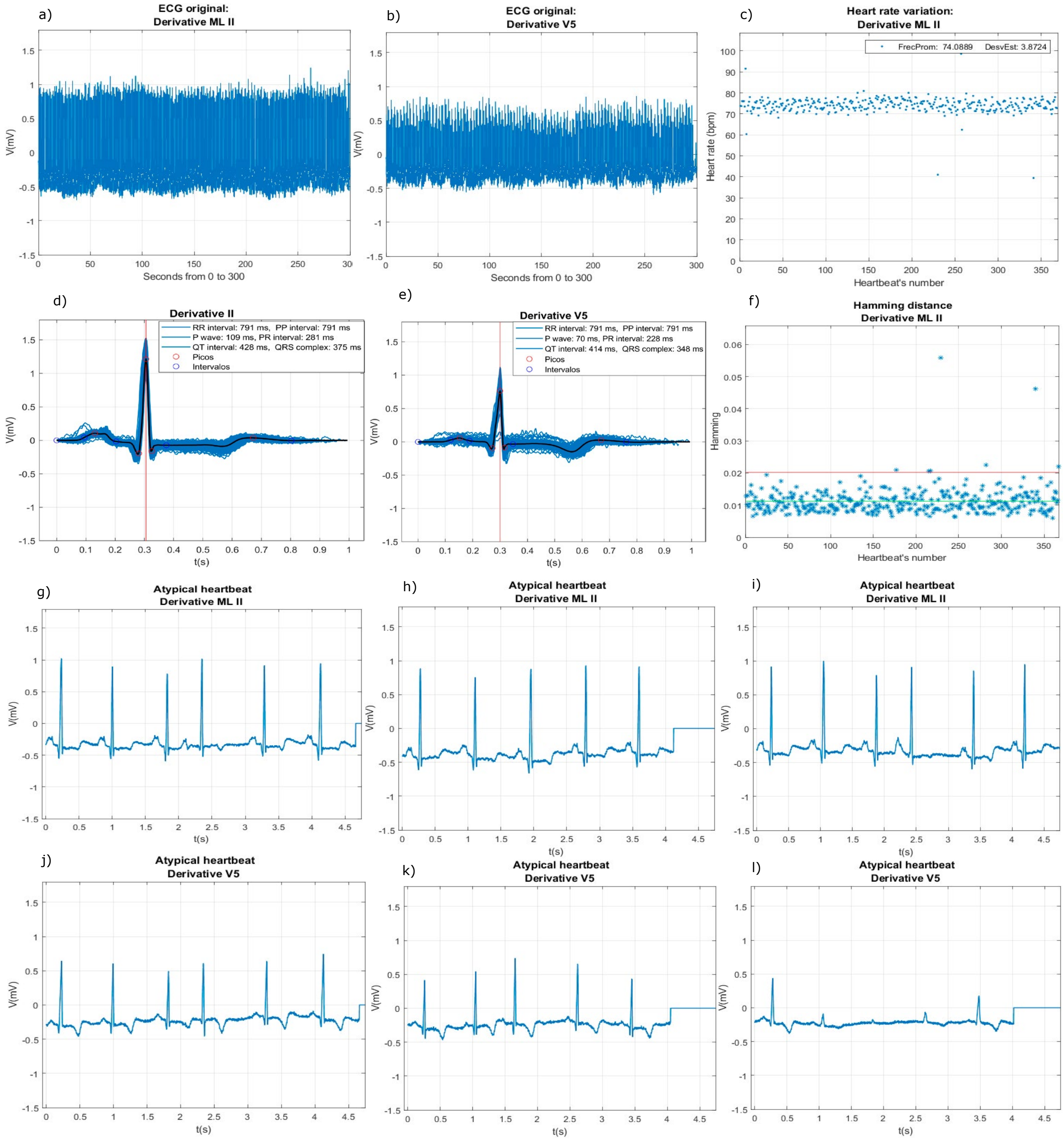


FIGURE 12. Processing of ECGs # 100 from MIT-BIH dataset: The micro-panels of the last two lines are examples of excerpts of abnormal heartbeats.

In register # 100, for the DII lead were detected 41 atypical beats and 46 atypical beats for the V5 lead (34 reported in the MIT-BIH database). In register # 106 were detected 54 atypical beats for the DII lead and 50 atypical beats for the V1 lead (520 reported in MIT-BIH database). In register # 112, 81 atypical beats were detected for the DII lead and 106 atypical beats for the V1 lead, in contrast to the 2 reported in the MIT-BIH database. We found that the number of atypical beats of one lead differs from that for the other of the same ECG. Also, the number of abnormal beats reported in the MIT-BIH database does not coincide with that detected by our classifier. It should be noted that the number of abnormal beats reported in the MIT-BIH database does not indicate to what of the leads this number is attributed. In addition, our detector identifies atypical beats, a number of them can be abnormal and the rest can be artifact errors. Figure 12 shows the processing of the first five minutes register # 100: the original ECG leads (slides a) and b)), their stacked beats representation matching peaks of R waves (slides d) and e)), the HR distribution (slide c)), the distribution of Hamming distance of beats from the reference one (slide f)), followed by the three exampled vicinities of detected atypical beats for the ML II and V5 lead. For the first five minutes in the ML II lead of register # 100 were detected 7 atypical beats and 5 atypical beats for the V5 lead (4 reported in MIT-BIH database). Figure 12 is similar to Figure 6, but it depicts the capability of expert system to process the long-term ECG.

The ECGs resemble chaotic signals, the register # 106 especially. The stacked beats representation of register # 106 looks confusing, but the morphology of the reference (mean) beat is seen clearly. That is due to the presence of double (multiple) beat partitions because of the abrupt HR changes and significant or critical morphological distortions of heartbeats' waveform. In case of long-time ECG registers with abrupt HR changes and significant morphological distortions of heartbeats we suggest processing ECG leads in required time intervals to see more details. Typical beats of an ECG lead in the stacked beat representation are accommodated (lodged, setting) in a compact group around the reference (mean) beat, so that the atypical (abnormal) beats are noticeable on that background. Figure 13 shows the first five minutes of register # 106, for the DII lead were detected 22 atypical beats and 13 atypical beats for the V5 lead (60 reported in MIT-BIH database).

CONCLUSIONS

The proposed algorithms of ECG analysis are in between human analysis and highly computerized completely autonomous expert systems. The reason and vantage of the approach are that cardiologists, being quite a conservative group in their methods of diagnosis due to the great responsibility for public health and the principles of education, are not willing to accept the technological novelties easily if those are not comprehensible or controlled directly by them. It is a kind of psychological resistance to using an expert system of completely automatized and computerized diagnosis which is a black box for them. Nevertheless, this kind of expert system is necessary for public health monitoring

To overcome the indicated contradiction, we proposed an intermediate state computer-assisted expert system for the processing and analysis of middle and longtime ECG and identification of atypical beats for early detection of heart diseases; the system permits a cardiologist to intervene, observe, and control to some extent the process of ECG analysis and take a final decision. Based on the stacked beats representation of ECG, were proposed a characteristic space and the detector of atypical beats for the posterior analysis by a cardiologist.

This is an interactive expert system for its use as an assistant of a cardiologist, with the potential of the next extension to the unsupervised classifier of heartbeats and CVD detection. Currently, the expert system is used for the detection of atypical heartbeats at an early stage of CVD. The classifier does not need an external training dataset; it is trained on the heartbeats of ECG under the analysis. This feature of the system is a noticeable advantage compared to those requiring training because the training usually needs an extensive database with the well previously classified entries. In addition, constructing a database one is subjected to be wrong unwillingly. The set of computer algorithms was developed in MatLab^{TM} language and is accompanied by the interface.

The expert system was simulated and evaluated on a set of ECGs labeled as normal in the CPSC2018 database and demonstrated its capability to detect up to minor morphological heartbeat distortions, precursors of CVD. Our expert system has detected a noticeable number of leads, labeled as normal in the CPSC2018 database, but contaminated by atypical and abnormal heartbeats; from 480 analyzed leads only 220 leads are considered free of atypical or abnormal heartbeats, artifacts being included in atypical beats.

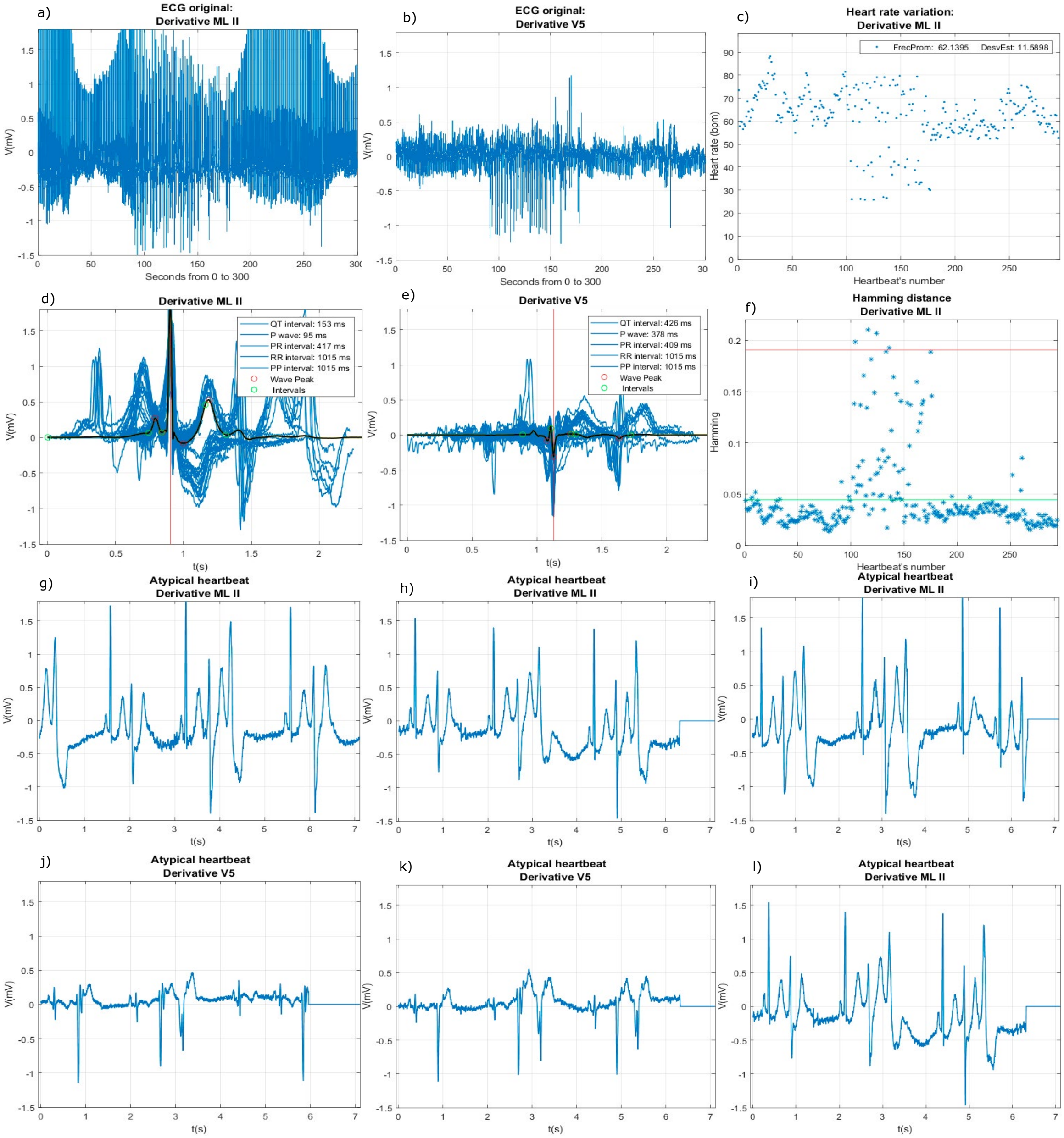


FIGURE 13. **Processing of ECGs # 106 from MIT-BIH dataset: The micro-panels of the last two lines are examples of excerpts of abnormal heartbeats.**

This inconsistency can be caused by two reasons: 1.- some of the contaminated ECGs are a result of artifacts produced by external or internal interference, or patient motion; 2.- In general, people go to the hospital to check their cardiovascular health status when they have some symptoms; so it is very likely that their ECG will present atypical or abnormal heartbeats. In addition, the system was evaluated on ECGs of the MIT-BIH database with clearly expressed abnormalities. In this case, the stacked beats representation often shows the cuts of double or even more beats, mainly because of strong morphological deformations in beats due to the artifacts and proper arrhythmias; for example, see register (record) # 112. Due to that, our system detects more atypical heartbeats than those abnormal reported in the database. Along with the stacked beats representation, all intervals, segments, and complexes are calculated and displayed on the screen for revision by a cardiologist.

For future work, we consider the improvement of the algorithm for the R-wave peaks detection in case of strong isoelectric and baseline wander. More clinical trials are certainly needed to introduce our system in practice.

Based on the results of the evaluation, the proposed expert system can be used in a real-time monitoring system and/or as a complementary system in hospitals to help physicians and cardiologists with ECG analysis at higher speed and accuracy to detect persons with ECD. The expert system can be potentially used in the following specific areas:

1. Support tool for an expert cardiologist in visualization of long ECGs, for precision calculation of intervals, segments, QRS complex and waves, visualization and exploration of details of atypical beats.
2. General practitioner with the possibility of obtaining a kind of electrocardiogram can use the expert system for prior diagnosis and redirect to a cardiologist if necessary.
3. Use in public health campaigns to monitor cardiac health in the population.
4. Incorporate the expert system into a personal cardiovascular health monitoring bracelet.

FUNDING

The authors sincerely acknowledge the *Consejo Nacional de Humanidades, Ciencia y Tecnología* (CONAHCYT), México for the support to this research trough the grant No. I1200/320/2022 to Dra. Deysy Galeana-Perez.

CONSENT TO PARTICIPATE

Informed consent was obtained from all individual participants included in the study.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

AUTHOR CONTRIBUTIONS

E.K. conceptualized and oversaw the project, participated in the designed and development of methodology, carried out formal analyses, validated results and wrote the manuscript. D.G.P. carried out formal analyses, contributed to the design and implementation of algorithms, validated and visualized the results and wrote the manuscript. Both authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] C. K. Jha & M. H. Kolekar, "Empirical Mode Decomposition and Wavelet Transform Based ECG Data Compression Scheme," IRBM, vol. 42, no. 1, pp. 65-72, Feb. 2021, doi: https://doi.org/10.1016/j.irbm.2020.05.008

[2] B. Mohamed, A. Issam, A. Mohamed, & B. Abdellatif, "ECG Image Classification in Real time based on the Haar-like Features and Artificial Neural Networks," Procedia Comput. Sci., vol. 73, pp. 32-39, 2015, doi: https://doi.org/10.1016/j.procs.2015.12.045

[3] I. Güler & E. D. Übeyli, "ECG beat classifier designed by combined neural network model," Pattern Recognit., vol. 38, no. 2, pp. 199-208, Feb. 2005, doi: https://doi.org/10.1016/j.patcog.2004.06.009

[4] L.-H. Wang, Y.-T. Yu, W. Liu, L. Xu, et al., "Three-Heartbeat Multilead ECG Recognition Method for Arrhythmia Classification," IEEE Access, vol. 10, pp. 44046-44061, Apr. 2022, doi: https://doi.org/10.1109/ACCESS.2022.3169893

[5] Z. Dokur & T. Ölmez, "ECG beat classification by a novel hybrid neural network," Comput. Methods Programs Biomed., vol. 66, no. 2-3, pp. 167-181, Sep. 2001, doi: https://doi.org/10.1016/S0169-2607(00)00133-4

[6] J.-S. Wang, W.-C. Chiang, Y.-L. Hsu, Y.-T. C. Yang, "ECG arrhythmia classification using a probabilistic neural network with a feature reduction method," Neurocomputing, vol. 116, pp. 38-45, Sep. 2013, doi: https://doi.org/10.1016/j.neucom.2011.10.045

[7] S. M. Anwar, M. Gul, M. Majid, M. Alnowami, "Arrhythmia Classification of ECG Signals Using Hybrid Features," Comput. Math. Methods Med., vol. 2018, art. no. 1380348, 2018, doi: https://doi.org/10.1155/2018/1380348

[8] F. A. Elhaj, N. Salim, A. R. Harris, T. T. Swee, T. Ahmed, "Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals," Comput. Methods Programs Biomed., vol. 127, pp. 52-63, Apr. 2016, doi: https://doi.org/10.1016/j.cmpb.2015.12.024

[9] M. K. Moridani, M. Abdi Zadeh, Z. Shahiazar Mazraeh, "An Efficient Automated Algorithm for Distinguishing Normal and Abnormal ECG Signal," IRBM, vol. 40, no. 6, pp. 332-340, Dec. 2019, doi: https://doi.org/10.1016/j.irbm.2019.09.002

[10] S. Sahoo, M. Mohanty, S. Behera, & S. Kumar Sabut, "ECG beat classification using empirical mode decomposition and mixture of features", J. Med. Eng. Technol., vol. 41, no. 8, pp. 652-661, 2017, doi: https://doi.org/10.1080/03091902.2017.1394386

[11] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification," Comput. Biol. Med., vol. 96, pp. 189-202, May 2018, doi: https://doi.org/10.1016/j.compbiomed.2018.03.016

[12] R. A. Alharbey, S. Alsubhi, K. Daqrouq, A. Alkhateeb, "The continuous wavelet transform using for natural ECG signal arrhythmias detection by statistical parameters," Alex. Eng. J., vol. 61, no. 12, pp. 9243-9248, Dec. 2022, doi: https://doi.org/10.1016/j.aej.2022.03.016

[13] J. Park & K. Kang, "PcHD: Personalized classification of heartbeat types using a decision tree," Comput. Biol. Med., vol. 54, pp. 79-88, Nov. 2014, doi: https://doi.org/10.1016/j.compbiomed.2014.08.013

[14] Y. Kutlu & D. Kuntalp, "A multi-stage automatic arrhythmia recognition and classification system," Comput. Biol. Med., vol. 41, no. 1, pp. 37-45, Jan. 2011, doi: https://doi.org/10.1016/j.compbiomed.2010.11.003

[15] J. Kim & H. Shin, "Simple and Robust Realtime QRS Detection Algorithm Based on Spatiotemporal Characteristic of the QRS Complex," PLoS ONE, vol. 11, no. 3, art. no. e0150144, 2016, doi: https://doi.org/10.1371/journal.pone.0150144

[16] N. Katsaouni, F. Aul, L. Krischker, S. Schmalhofer, L. Hedrich, M. H. Schulz, "Energy efficient convolutional neural networks for arrhythmia detection," Array, vol. 13, art. no. 100127, 2022, doi: https://doi.org/10.1016/j.array.2022.100127

[17] X. Yang, X. Zhang, M. Yang, L. Zhang, "12-Lead ECG arrhythmia classification using cascaded convolutional neural network and expert feature," J. Electrocardiol., vol. 67, pp. 56-62, 2021, doi: https://doi.org/10.1016/j.jelectrocard.2021.04.016

[18] S. G. Begum, E. Priyadarshi, S. Pratap, S. Kulshrestha, V. Singh, "Automated Detection of Abnormalities in ECG signals using Deep Neural Network," Biomed. Eng. Adv., vol. 5, art. no. 100066, Jun. 2023, doi: https://doi.org/10.1016/j.bea.2022.100066

[19] S. C. Mohonta, M. A. Motin, & D. K. Kumar, "Electrocardiogram Based Arrhythmia Classification Using Wavelet Transform with Deep Learning Model," Sens. Bio-Sens. Res., vol. 37, 2022, doi: http://dx.doi.org/10.2139/ssrn.4088025

[20] J. H. Abawajy, A.V. Kelarev, M. Chowdhury, "Multistage approach for clustering and classification of ECG data," Comput. Methods Programs Biomed., vol. 112, no. 3, pp. 720-730, Dec. 2013, doi: https://doi.org/10.1016/j.cmpb.2013.08.002

[21] W.-H. Jung & S.-G. Lee, "An Arrhythmia Classification Method in Utilizing the Weighted KNN and the Fitness Rule," IRBM, vol. 38, no. 3, pp. 138-148, Jun. 2017, doi: https://doi.org/10.1016/j.irbm.2017.04.002

[22] L. Lu, J. Yan, C. W. de Silva, "Feature selection for ECG signal processing using improved genetic algorithm and empirical mode decomposition," Measurement, vol. 94, pp. 372-381, Dec. 2016, doi: https://doi.org/10.1016/j.measurement.2016.07.043

[23] U. R. Acharya, P. Subbanna Bhat, S.S. Iyengar, A. Rao, Sumeet Dua, "Classification of heart rate data using artificial neural network and fuzzy equivalence relation," Pattern Recognit., vol. 36, no. 1, pp. 61-68, Jan. 2003, doi: https://doi.org/10.1016/S0031-3203(02)00063-8

[24] J. Rahul, M. Sora, L. D. Sharma, V. K. Bohat, "An improved cardiac arrhythmia classification using an RR interval-based approach," Biocybern. Biomed. Eng., vol. 41, no. 2, pp. 656-666, 2021, doi: https://doi.org/10.1016/j.bbe.2021.04.004

[25] R. Rohmantri & N. Surantha, "Arrhythmia Classification using 2D Convolutional Neural Network," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, 2020, doi: http://dx.doi.org/10.14569/IJACSA.2020.0110427

[26] S. L. Oh, E. Y. K. Ng, R. S. Tan, U. R. Acharya, Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats, Comput. Biol. Med., vol. 102, pp. 278-287, Nov. 2018, doi: https://doi.org/10.1016/j.compbiomed.2018.06.002

[27] B. Chen, D. M. Maslove, J. D. Curran, A. Hamilton, P. R. Laird, P. Mousavi, S. Sibley, "A deep learning model for the classification of atrial fibrillation in critically ill patients," Intensive Care Med. Exp., vol. 11, art. no. 2, Jan. 2023, doi: https://doi.org/10.1186/

s40635-022-00490-3

[28] O. Yildirim, U. B. Baloglu, R.-S. Tan, E. J. Ciaccio, U. R. Acharya, “A new approach for arrhythmia classification using deep coded features and LSTM networks,” Comput. Methods Programs Biomed., vol. 176, pp. 121-133, Jul. 2019, doi: https://doi.org/10.1016/j.cmpb.2019.05.004

[29] N. Du, Q. Cao, L. Yu, N. Liu, E. Zhong, Z. Liu, Y. Shen, K. Chen, Fm-ecg: A fine-grained multi-label framework for ecg image classification, Information Sciences 549 (2021) pp. 164-177, doi: https://doi.org/10.1016/j.ins.2020.10.014

[30] A. Aranda Hernandez, P. Bonizzi, R. Peeters, J. Karel, “Continuous monitoring of acute myocardial infarction with a 3-Lead ECG system,” Biomed. Signal Process. Control, vol. 79, art. no. 104041, Jan. 2023, doi: https://doi.org/10.1016/j.bspc.2022.104041

[31] G. B. Moody & R. G. Mark, “The impact of the MIT-BIH Arrhythmia Database,” IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45-50, 2001, doi: https://doi.org/10.1109/51.932724

[32] F. Liu, C. Liu, L. Zhao, X. Zhang, et al., “An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection,” J. Med. Imaging Health Inf., vol. 8, no. 7, pp. 1368-1373, 2018, doi: http://dx.doi.org/10.1166/jmihi.2018.2442

[33] M. Bellamoli, F. Marin, L. Maritan, D. Prati, et al., “New-onset extreme right axis deviation in acute myocardial infarction: clinical characteristics and outcomes,” J. Electrocardiol., vol 60, pp. 60–66, 2020, doi: https://doi.org/10.1016/j.jelectrocard.2020.03.003

[34] X. Gu, J. Hu, L. Zhang, J. Ding, F. Yan, “An Improved Method with High Anti-interference Ability for R Peak Detection in Wearable Devices,” IRBM, vol. 41, no. 3, pp. 172–183, Jun. 2020, doi: https://doi.org/10.1016/j.irbm.2020.01.002

[35] C. Cakir, Y. Ceylan, O. Y. Akbal, R. Sarikaya, S. Barutcu, “Clinical characteristics and angiographic findings of non-st-elevation acute coronary syndrome patients admitted with normal electrocardiogram,” J. Electrocardiol., vol. 60, pp. 77–81, 2020, doi: https://doi.org/10.1016/j.jelectrocard.2020.03.002

[36] N. Dugarte, E. Dugarte, N. Dugarte, Electrocardiografía de Alta Resolución Técnicas Aplicadas de Adquisición y Procesamiento, Argentina: Universidad Tecnológica Nacional (UTN-FRM), 2018.

[37] R. Srivastva, A. Singh, Y. N. Singh, “PlexNet: A fast and robust ECG biometric system for human recognition,”

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 105 - 116

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.7

E-LOCATION ID: 1394

Performance Evaluation of Biomedical Time Series Transformation Methods for Classification Tasks

Evaluación del Rendimiento de Métodos de Transformación de Series Temporales Biomédicas para Tareas de Clasificación

Carlos Alejandro Ku-Maldonado[1], *Erik Molino-Minero-Re*[2]

[1]Universidad Nacional Autónoma de México, Yucatán - México

[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, Unidad Académica, Yucatán - México

ABSTRACT

The extraction of time series features is essential across various fields, yet it remains a challenging endeavor. Therefore, it's crucial to identify appropriate methods capable of extracting pertinent information that can significantly enhance classification performance. Among these methods are those that translate time series into different domains. This study investigates three distinct time series transformation approaches for addressing time series classification challenges within biomedical data. The first method involves a response vector transformation, while the other two employ image transformation techniques: RandOm Convolutional KErnel Transform (ROCKET), Gramian Angular Fields, and Markov Transition Fields. These transformation methods were applied to five biomedical datasets, exploring various format configurations to ascertain the optimal representation technique and configuration for input, which in turn improves classification performance. Evaluations were conducted on the effectiveness of these methods in conjunction with two classification algorithms. The outcomes underscore the significance of these time series transformation techniques as facilitators for enhanced classification algorithms documented in current literature.

KEYWORDS: biomedical data, classification, convolutional neural networks, time series, transformations

RESUMEN

La extracción de características de series temporales es esencial en diversos campos, pero sigue siendo un desafío. Por lo tanto, es crucial identificar métodos apropiados capaces de extraer información pertinente que pueda mejorar significativamente el rendimiento de clasificación. Entre estos métodos se encuentran aquellos que traducen las series temporales a diferentes dominios. Este estudio investiga tres enfoques distintos de transformación de series temporales para abordar los desafíos de clasificación de series temporales en datos biomédicos. El primer método implica una transformación de vector de respuesta, mientras que los otros dos emplean técnicas de transformación de imagen: RandOm Convolutional KErnel Transform (ROCKET), Gramian Angular Fields y Markov Transition Fields. Estos métodos de transformación se aplicaron a cinco conjuntos de datos biomédicos, explorando diversas configuraciones de formato para determinar la técnica y configuración de representación óptima para la entrada, lo que a su vez mejora el rendimiento de clasificación. Se realizaron evaluaciones sobre la efectividad de estos métodos en conjunción con dos algoritmos de clasificación. Los resultados subrayan la importancia de estas técnicas de transformación de series temporales como facilitadoras para mejorar los algoritmos de clasificación documentados en la literatura actual.

PALABRAS CLAVE: clasificación, datos biomédicos, redes neuronales convolucionales, series temporales, transformaciones

Corresponding author

TO: **Carlos Alejandro Ku-Maldonado**
INSTITUTION: **Universidad Nacional Autónoma de México**
ADDRESS: **Carretera Mérida-Tetiz Km. 4.5, Ucú, Yucatán, México, C. P. 97357**
EMAIL: **ku.carlos@aries.iimas.unam.mx**

Received:

31 October 2023

Accepted:

07 Dicember 2023

INTRODUCTION

The study of time series classification has gained a great deal of interest due to the large amounts of information generated from the study of various phenomena that evolve in time. This dynamic behavior is observed in the biomedical areas such as disease studies, drug efficacy assessments, treatment analyses, signal processing, and image analysis, etc. where the importance of finding patterns that can help us to segment behaviors arises.

The complexity in time series classification stems from the inadequacy of traditional similarity metrics, which fail to account for the temporal aspect inherent in such data. These datasets frequently exhibit temporal relationships, shifts over time, and variations in magnitude, posing challenges for classification methods reliant on linear comparisons. Certain conventional approaches opt for an initial feature extraction phase to derive pertinent information from the data. However, given these limitations, there's a pressing need to explore alternative representations that remain invariant to shifts, preserve magnitudes, and retain temporal dependencies.

In [1], various techniques for time series classification across different domain representations are explored. These include methods based on dictionary representation, which involves counting the frequency of specific patterns within the data. Another approach involves extracting distinctive shapes, termed as shapelets, present within each class, enabling effective differentiation between groups. Additionally, one analysis strategy involves converting signals into images, leveraging our current computational capabilities to develop more sophisticated models, often relying on neural networks [2][3][4][5][6][7].

This article's contribution lies in its exploration of three distinct methods for representing time series data, specifically applied to biomedical signals. These methods unveil various temporal relationships within the information, potentially enhancing classification performance. The study also delves into analyzing which representation proves most effective for each dataset and classifier, thereby shedding light on the optimal approach for different data contexts and classification algorithms.

In this paper, we aim to review three cutting-edge time series transformation methods, each tested with diverse hyperparameter configurations. These methods will be applied across five distinct biomedical datasets, each possessing unique properties. The objective is to thoroughly evaluate their efficacy in handling classification tasks within these varied biomedical contexts.

MATERIALS AND METHODS

Datasets

The article focuses on five datasets obtained from the UCR Time Series Classification Archive [8], which are widely acknowledged and frequently employed in state-of-the-art methodologies:

A. 'SemgHandGenderCh2' (SHGC2): Captures surface electromyography signal power spectrum data depicting muscle electric activity during six hand grasp movements by five healthy subjects. These movements are categorized into two classes: Class 1 representing Female and Class 2 representing Male.

B. 'PhalangesOutlinesCorrect' (POC): Centers on hand and bone outline detection, categorizing the output label as either a correct or incorrect image outline.

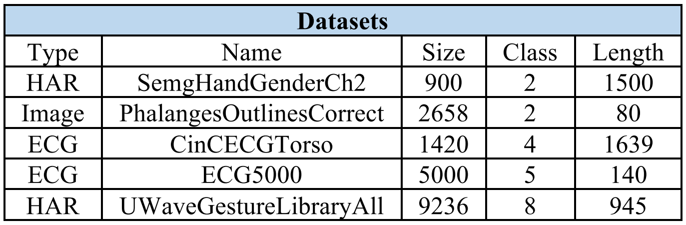
C. 'CinCECGTorso' (CCECGT): Features data derived from ECG chlorine concentration readings across multiple torso-surface sites from four distinct individuals, each individual representing a single class.

D. 'ECG5000': Revolves around ECG heartbeat recordings from a patient with severe congestive heart failure, with class values determined via automated annotation.

E. 'UWaveGestureLibraryAll' (UWGLA): Comprises a collection of eight simple gestures generated from accelerometers.

Further details and characteristics of these datasets are available in Table 1. Notably, these datasets share the common characteristic of being univariate time series data, characterized by ndim = 1.

TABLE 1. **UCR Datasets [8].**



Datasets				
Type	Name	Size	Class	Length
HAR	SemgHandGenderCh2	900	2	1500
Image	PhalangesOutlinesCorrect	2658	2	80
ECG	CinCECGTorso	1420	4	1639
ECG	ECG5000	5000	5	140
HAR	UWaveGestureLibraryAll	9236	8	945

It's interesting to note that these datasets encompass three distinct types of data sources:

Human Activity Recognition (HAR): This category involves datasets like 'SemgHandGenderCh2' (SHGC2) and 'UWaveGestureLibraryAll' (UWGLA), which capture human movements and gestures, often derived from sensors or accelerometers.

Image Contours (Image): 'PhalangesOutlinesCorrect' (POC) falls under this category, focusing on the detection and assessment of image outlines, specifically related to hand and bone outlines.

Electrocardiogram Signals (ECG): Datasets such as 'CinCECGTorso' (CCECGT) and 'ECG5000' involve electrocardiogram data, primarily concerning the recording and analysis of heart-related signals, including chlorine concentration readings and heartbeat patterns.

Transformation methods

RandOm Convolutional KErnel Transform (ROCKET)

Convolutional kernels can be thought of as a matrix of values used to modify the input data by a dot product. These kernels contain certain basic parameters such as length, weights, bias, dilation, and padding. The kernels have the same 1-dimensional vector structure as the input data but are smaller in size. In the case of time series, the kernels are weighted vectors with a bias added to the result of the convolution between the input data and the kernel weights.

These kernels operate as filters, enabling the extraction of diverse shapes and patterns embedded within time series data. Through kernel dilation, these filters can extract patterns across various scales, reflecting different frequency characteristics. By leveraging a combination of multiple kernels, the system can extract complex patterns. While the weights of convolutional neural network kernels are typically learned, random convolutional kernels have demonstrated striking effectiveness.

ROCKET capitalizes on the concept of employing random convolutional kernels as a feature transformation for input in other classifiers. This transformation method offers a low computational cost owing to the random initialization of kernels instead of learned weights. As it involves a single layer, numerous kernels can be generated without significantly escalating the computational demand.

ROCKET employs a mechanism akin to max pooling by extracting maximum values from each feature map. However, it introduces a novel feature: the proportion of positive values. This addition enables ROCKET to discern the prevalence of a distinct pattern within the time series [9]. The process of this method is visually illustrated in Figure 1.

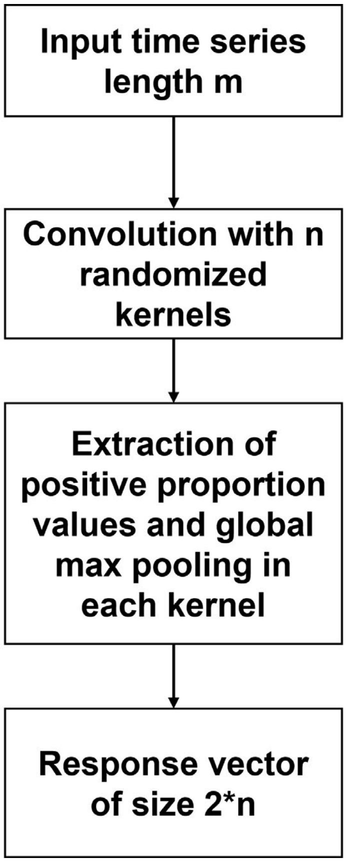


FIGURE 1. **Random convolutional kernel transforms flow diagram.**

Gramian Angular Field

The Gramian Angular Field (GAF) is based on the dot product operation between two vectors, which shows the similarity that exists between them. If we have two vectors u and v with a norm of 1, we get Equation 1.

$$\langle u, v\rangle = ||u|| \cdot ||v|| \cdot cos(\theta) \tag{1}$$

When working with unit vectors, the dot product is characterized only by the angle θ between *u* and *v*, since the magnitude of a unit vector is 1, which is why the result falls in the range [-1,1].

$$G = \begin{pmatrix} \langle v_1, v_1\rangle & \langle v_1, v_2\rangle & \cdots & \langle v_1, v_n\rangle \\ \langle v_2, v_1\rangle & \langle v_2, v_2\rangle & \cdots & \langle v_2, v_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_n, v_1\rangle & \langle v_n, v_2\rangle & \cdots & \langle v_n, v_n\rangle \end{pmatrix} \tag{2}$$

Considering the data as unit vectors, the Gramian matrix depicted in Equation 2 is formulated to capture linear dependencies within a vector set. This matrix serves to retain temporal dependencies, embedding the temporal dimension within the geometry of the matrix [10].

Given a time series $X=x_1, x_2, \ldots, x_n$ of n real observations, X is rescaled so that all values remain in the interval [-1,1] to preserve the unit vector property.

Once the information is scaled, the polar transformation is obtained. In this case, two values are considered:

1. The actual value of the time series (observations).

2. Their respective time labels.

From these values, the angle will be obtained, and this operation is described in Equation 3.

$$\begin{cases} \theta_i = \arccos(x_i) \\ r_i = \frac{i}{N} \end{cases} \tag{3}$$

Some advantages of this coding are that it is a composition of bijective functions and that the temporal dependence is preserved based on the variable r_i. With these new angles, the dot product operation is adapted using Equation 4 so that it can be treated as a Gram matrix.

$$x \oplus y = \cos(\theta_1 + \theta_2) \tag{4}$$

By implementing Equation 4 in G, the following Gramian matrix is obtained.

$$G = \begin{pmatrix} \cos(\theta_1 + \theta_1) & \cos(\theta_1 + \theta_2) & \cdots & \cos(\theta_1 + \theta_n) \\ \cos(\theta_2 + \theta_1) & \cos(\theta_2 + \theta_2) & \cdots & \cos(\theta_2 + \theta_n) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\theta_n + \theta_1) & \cos(\theta_n + \theta_2) & \cdots & \cos(\theta_n + \theta_n) \end{pmatrix} \tag{5}$$

where $\cos(\theta_1+\theta_2)$ is given by Equation 6 in cartesian system.

$$\cos(\theta_1+\theta_2) = x \cdot y - \sqrt{1-x^2} \cdot \sqrt{1-y^2} \tag{6}$$

This transformation generates a density map, portraying values through a color spectrum with intensities spanning from -1 to 1. Figure 2 illustrates a visual depiction of the encoded magnitudes.

Figure 3 showcases the distinct stages of this method, delineating the process involved in achieving the intended image representation.

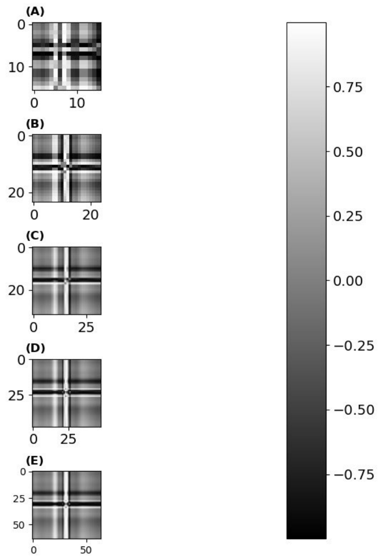


FIGURE 2. Grammian angular field encoding sizes. a) 16X16 image. b) 24X24 image. c) 32X32 image. d) 48X48 image. e) 64x64 image.

Markov transition fields

Markov Transition Fields (MTF) offer an alternative method of transforming a time series into an image, employing a probabilistic framework rooted in Markov chains. This approach provides a distinctive perspective on representing time series data as visual elements.

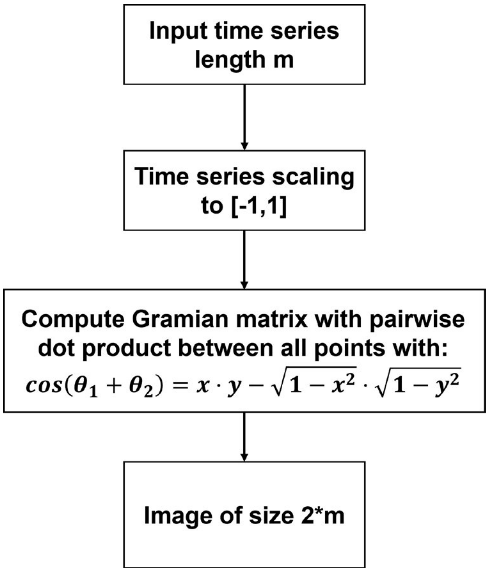


FIGURE 3. Gramian angular field flow diagram.

Given a time series $X = x_1, x_2, \ldots, x_n,$ its Q quantile bins are determined and each value x_i is assigned to its corresponding bin q_j ($j \in [1,Q]$). From these bins, a weighted adjacency matrix W of size $Q \times Q$ is constructed, counting the transitions between bins in the form of a Markov chain along the time axis [10].

$$M = \begin{pmatrix} w_{ij}|x_1 \in q_i, x_1 \in q_j & \cdots & w_{ij}|x_1 \in q_i, x_n \in q_j \\ w_{ij}|x_2 \in q_i, x_1 \in q_j & \cdots & w_{ij}|x_2 \in q_i, x_n \in q_j \\ \vdots & \ddots & \vdots \\ w_{ij}|x_n \in q_i, x_1 \in q_j & \cdots & w_{ij}|x_n \in q_i, x_n \in q_j \end{pmatrix} \tag{7}$$

The calculation of w_{ij} is based on the frequency of occurrence where a point in quantile q_j is succeeded by a point in quantile q_i. After deriving this transition matrix, it undergoes normalization process to yield the matrix M. In this transition matrix, M_{ij} represents the probability of transitioning from q_i to q_j. Leveraging these probabilities, we encode the time series into an $N \times N$ matrix, where N corresponds to the length of the original signal.

The color representation of the original signal is depicted through intensities ranging from [0, 1], delineated by the probability distribution. This image illustrates the likelihood of a point transitioning to other points (states) within the time series. Figure 4 showcases the visual encoding derived from this probability distribution.

For a comprehensive view of the method's process, please refer to Figure 5, which details the sequential steps involved in this encoding procedure.

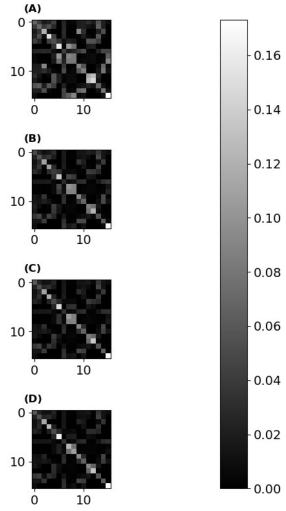


FIGURE 4. **Markov transition fields resolutions. a) 8 bins. b) 16 bins. c) 32 bins. d) 64 bins.**

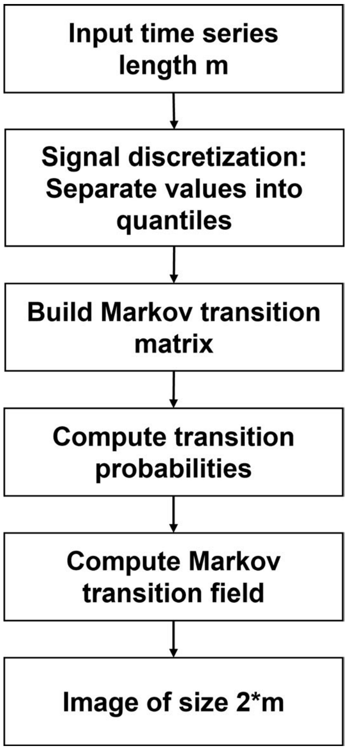


FIGURE 5. **Markov transition field flow diagram.**

All three of these transform representations are shown in Figure 6 for the same set of time series.

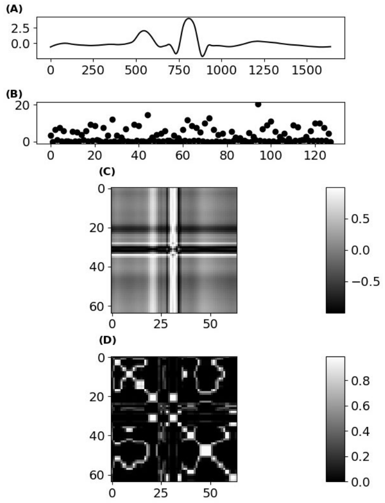


FIGURE 6. **Transformation methods in time series. a) Original time series. b) ROCKET response vector. c) Gramian angular field transformation (GAF). d) Markov transition field transformation (MTF).**

System and software specifications

This methodology was performed in a 11th Gen Intel® Core ™ i7-11800H processor with a NVidia GeForce RTX 3060 mobile GPU and 16 gb of DDR4 Sodimm.

The transformation methods were implemented using Pyts [11], and the classifiers were selected using the Scikit learn and tensorflow-keras packages [12][13].

Pre-processing

Given the common occurrence of artifacts like noise and baseline offsets in biomedical signal data, a preliminary filtering stage was applied to enhance signal quality. To achieve this, a Savitzky-Golay filter was employed across all five datasets. This filter, widely used in biomedical signal processing [14], utilized a window size equivalent to 10 % of the signal length with a polyorder of 2. Notably, this filter retains essential characteristics of the initial distribution, including relative maximums, minimums, and peak widths, resulting in smoother signal behavior.

Additionally, Z normalization was performed to rectify baseline offsets, ensuring a consistent reference across all signal sets. This standardization establishes a zero mean and unit variance across the dataset, a technique known to often enhance model performance.

Once the whole dataset was ready, the training and testing subsets were created. The training set was defined as 70 % of the dataset size, the remaining 30 % was used for testing the classification models. 20 % of the training set was used as validation.

For comparison purpose, all three methods allow to define the output dimensions. On the image-based transformations it is defined by an $N \times N$ image where $N \in$ {16,24,32,40,48\}. In ROCKET the response vector obtained is given by the number of kernels M generated times 2, the number of features extracted in every kernel, this equals to a $2M$ vector where $M \in$ {625,1250,2500,5000,10000}.

Rocket

Other set of parameters that were random selected along all the kernels were, weights, bias, dilation, padding, the implementation can be seen in [9].

Gramian Angular Fields (GAF)

In GAF other variation parameters is the mode, in this case we have summation and difference mode.

Markov Transition fields (MTF)

For MTF the remaining variable parameter was the number of bins, this number gives us the number of states or resolution the data values can transition to. The number of bins was given by *nBins* where *nBins* ∈{8,16,32,64}.

All three transformations—ROCKET, GAF, and MTF—were applied across the five datasets, encompassing various combinations of parameters. These transformed datasets were utilized as inputs for the subsequent classification methods.

Classification methods

The classification stage aimed solely at evaluating transformation performance. The five datasets, transformed using distinct configurations, served as inputs for fixed classifiers to determine the input format that optimizes classification performance. No hyperparameter tuning was conducted on the neural network in this phase.

For the ROCKET-transformed data, a Ridge regressor classifier was utilized. This classifier incorporates a regularization parameter α within the loss function to prevent overfitting. Various values of α, selected from $\alpha \in$ {0.001, 0.01, 1, 10, 100, 1000}, were tested to identify the best-fitting parameter that avoids overfitting the model. Classifier selection was guided by references [9][10].

In the case of image-based transformations, a convolutional neural network (CNN) with fixed parameters was employed. Categorical encoding was applied to the labels to optimize network functionality. The CNN architecture included the following modules:

1. input layer (image size)
2. convolutional layer (6 neurons, 8x8 size kernels)
3. max pool layer (3x3 pool size, padding=same)
4. convolutional layer (6 neurons, 3x3 size kernels)
5. max pool layer (3x3 pool size, padding=same)
6. flattening layer
7. dropout(0.5)
8. dense layer (categorical enconding = # labels, softmax)

The CNN utilized mean square error as the loss function, employed the adam optimizer, updated batch size weights of 50, and ran for 25 epochs. Figure 7 illustrates the various stages involved in evaluating the enhancement of time series transformations for classification.

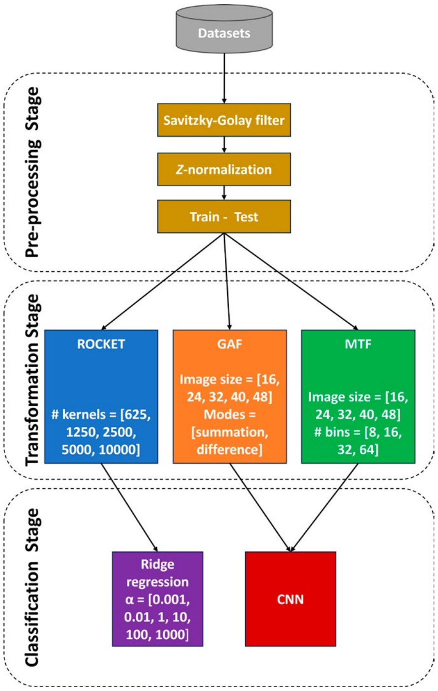


FIGURE 7. **Flow diagram of methodology.**

Evaluation metrics

In the evaluation process, both mean square error (MSE) and accuracy were utilized to assess the performance of various input configurations and different parameter settings within the classifier. These metrics were computed on both the training and validation sets. To identify the most effective input configuration, the validation set's lowest MSE was considered.

For assessing test performance, metrics including accuracy, precision, recall, and f1-score were obtained. These measurements offer a comprehensive understanding of the model's predictive capabilities and effectiveness in correctly classifying instances.

RESULTS AND DISCUSSION

The preprocessing stage involved a stratification split to ensure an equal representation of class examples within the training and test sets. However, only the CinCECGTorso and UWaveGestureLibraryAll datasets exhibited an equal number of examples in each class.

Figure 8 demonstrates that ROCKET coupled with ridge regression exhibited notable performance, particularly with datasets where all classification test accuracies surpassed 0.8. One significant advantage observed in employing this method was its low memory resource consumption, as it did not require batch execution. Additionally, it was noted that the regularization parameter tended to perform better when $\alpha \geq 1$. However, the ECG5000 dataset showed poor precision, potentially due to class imbalance issues. It's important to note that the computing time for this method increased with a higher number of kernels, although the fitting process to the classifier was relatively swift once the data was transformed.

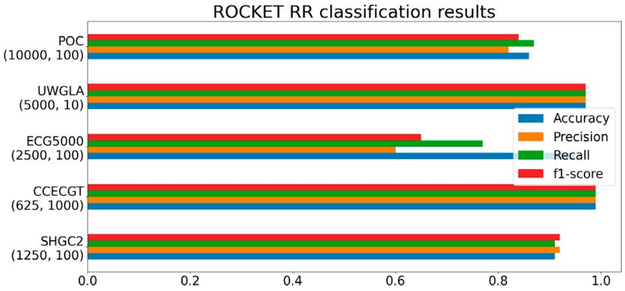


FIGURE 8. **ROCKET transformation performance on the five datasets, number below each dataset correspond to (#kernels, α).**

GAF, similar to ROCKET, showed promising performance on this type of data. However, certain limitations were observed, particularly with respect to memory consumption. Unlike ROCKET, the GAF transformation required batch processing for datasets larger than 1000 samples. Despite this limitation, the transformation process itself was rapid, as were the training times for the classifier.

Regarding the mode selection in GAF, it was noted that the difference mode yielded better results. Additionally, the optimal image size seemed to be 48.

A detailed overview of performance results can be seen in Figure 9.

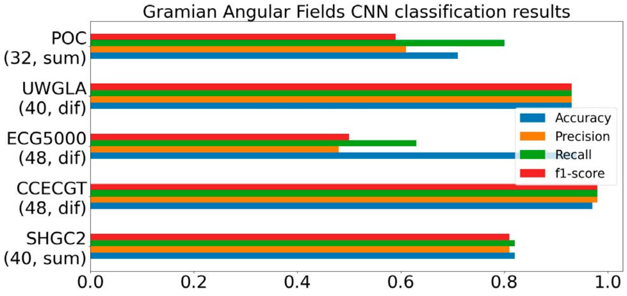


FIGURE 9. Gramian angular field transformation performance on the five datasets, number below each dataset correspond to (image size, transformation mode).

Finally, we conducted the MTF classification, yet this transformation didn't yield superior results compared to the preceding two methods. As illustrated in Figure 10, specifically in the precision column of ECG5000, MTF exhibited heightened sensitivity to quantile selection. Signals with low variability necessitated a smaller quantile size; larger quantile sizes led to certain bins containing only a few data points. Similar to the former method, MTF also required transformation into batches for signals exceeding a length of 1000 samples.

The optimal outcomes were achieved with an image size of 48 and # bins set between 8 to 16.

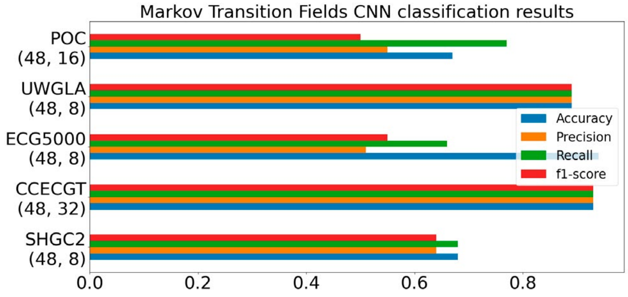


FIGURE 10. Markov transition field transformation performance on the five datasets, number below each dataset correspond to (image size, # bins).

In Figure 11 we summarize the F1 score on all datasets given by the optimal parameters for each transformation methods.

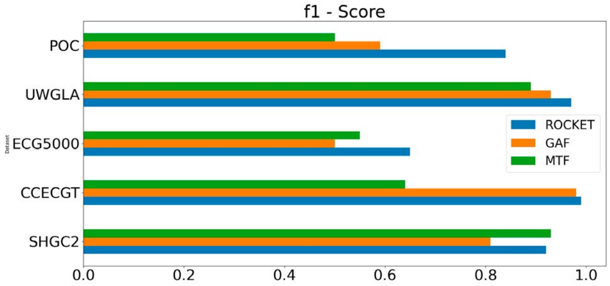


FIGURE 11. f-1 score of all 3 transformation methods on the five datasets.

The computation time was systematically measured to analyze how signal size and the number of examples influence the time complexity of these methods. Observations revealed that while ROCKET consistently demonstrated better accuracy across all datasets, its computation time increased with a higher number of filters.

Conversely, GAF and MTF exhibited shorter transformation times, but they tended to consume more memory, scaling the information by N^2 where N represents the signal size. Additionally, MTF encountered challenges when discretizing time series with large bins and limited variability, preferring a smaller number of bins in such cases.

The time metrics were obtained by averaging the time taken for each transformation configuration across each dataset. These computation times are detailed in Table 2.

TABLE 2. Computation time measurements.

Dataset mean computation time (seconds)			
Name	ROCKET	GAF	MTF
SemgHandGenderCh2	30.7	1.8	5.7
PhalangesOutlinesCorrect	5.9	2.3	2.4
CinCECGTorso	51.8	2.0	9.7
ECG5000	18.2	2.9	3.4
UWaveGestureLibraryAll	97.2	3.1	18.7

It's evident that these transformation methods serve as pivotal preprocessing steps, significantly enhancing the classification performance of these methods. The mean accuracy across all five datasets demonstrates strong performance: 0.936 for ROCKET, 0.874 for GAF, and 0.822 for MTF, all showcasing accuracies exceeding 0.8. This underscores their effectiveness in improving classification outcomes across diverse datasets.

For an overview of the key features and comparative analysis of each transformation method, please refer to Table 3.

TABLE 3. **Methods observed characteristics.**

Comparison Table of methods		
ROCKET	GAF	MTF
-Presents the highest accuracy of the three representations, with f1 scores over 0.6, even with unbalanced datasets. -The computation time increases linearly with the increase of number of kernels and the time series length. -It's not image based, can be used with more traditional methods. -Due to the random initialization of the kernels, no hyperparameter tunning is required.	-Presents the lowest computation time. -The accuracy obtained is lower than ROCKET, but comparable in some datasets. -Performance may increase with hyperparameter tunning on the NN or implementing other architectures. -Involves more hyperparameter for the transformation stage. -Memory consuming method. -Piecewise aggregate approximation is needed for high length data to avoid memory overflow.	-Presents the lowest accuracy of all three methods. Struggles with unbalanced datasets. -Presents stability issues with high number of bins. -Has a fast computing time when transformation hyperparameters are well suited. - Involves more hyperparameter for the transformation stage. -Memory consuming method. -Piecewise aggregate approximation is needed for high length data to avoid memory overflow.

Depending on the specific application and the tradeoffs you're willing to consider, you might opt for one method over another. Among all three, GAF demonstrated the most balanced performance concerning accuracy and utilization of time resources.

Additionally, these scrutinized time series transformation methods could prove beneficial in various machine learning tasks beyond classification, such as forecasting [15][16][17][18] or clustering [19]. This can be explored in future work.

CONCLUSION

In this study, our primary focus was on transforming time series data to enhance classification performance. We introduced three distinct methods—Random Convolutional Kernel Transform, Gramian Angular Fields, and Markov Transition Fields—and rigorously tested their efficacy using five diverse biomedical datasets. The results illustrate that biomedical signals, when treated as time series data, can achieve classification performances akin to state-of-the-art methodologies [20][21].

While our choice of classifiers was based on existing literature, it's crucial to note their adaptability to various other classifier types and architectures. For instance, in [10], a tiled Convolutional Neural Network was employed, with its last layer serving as input for a Support Vector Machine classifier. Different neural network architectures might yield comparable outcomes.

Notably, these imaging methods exhibit sensitivity to noise and generally perform optimally with signals less than 1000 samples in length due to memory constraints. However, they enable the extraction of diverse features from the data, significantly improving classification performance for temporal datasets.

For large-scale applications, the integration of deep learning computer vision approaches with attention mechanisms could leverage these representations, potentially offering promising outcomes in big data contexts.

REFERENCES

[1] M. Middlehurst, P. Schäfer, and A. Bagnall, "Bake off redux: a review and experimental evaluation of recent time series classification algorithms," 2023, arXiv:2304.13029, doi: https://doi.org/10.48550/arXiv.2304.13029

[2] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," Data Min. Knowl. Discov., vol. 33, no. 4, pp. 917-963, 2019, doi: https://doi.org/10.1007/s10618-019-00619-1

[3] C. Li, J. Xiong, X. Zhu, Q. Zhang, and S. Wang, "Fault Diagnosis Method Based on Encoding Time Series and Convolutional Neural Network," IEEE Access, vol. 8, pp. 165232-165246, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3021007

[4] G. R. Garcia, G. Michau, M. Ducoffe, J. Sen Gupta, and O. Fink, "Temporal signals to images: Monitoring the condition of industrial assets with deep learning image processing algorithms," Proc. Inst. Mech. Eng. O J. Risk Reliab., vol. 236, no. 4, pp. 617-627, 2022, doi: https://doi.org/10.1177/1748006X21994446

[5] J. Lines, S. Taylor, and A. Bagnall, "Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification," in 2016 IEEE 16th international conference on data mining (ICDM), Barcelona, Spain, 2016, pp. 1041-1046, doi: https://doi.org/10.1109/ICDM.2016.0133

[6] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 1578-1585, doi: https://doi.org/10.1109/IJCNN.2017.7966039

[7] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, et al., "Inceptiontime: Finding alexnet for time series classification," Data Min. Knowl. Discov., vol. 34, no. 6, pp. 1936-1962, 2020, doi: https://doi.org/10.1007/s10618-020-00710-y

[8] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, E. Keogh, "The UCR time series archive," IEEE/CAA J. Autom. Sin., vol. 6, no. 6, pp. 1293-1305, 2019, doi: https://doi.org/10.1109/JAS.2019.1911747

[9] A. Dempster, F. Petitjean, and G. I. Webb, "ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels," Data Min. Knowl. Discov., vol. 34, no. 5, pp. 1454-1495, 2020, doi: https://doi.org/10.1007/s10618-020-00701-z

[10] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," 2015, arXiv:1506.00327, 2015, doi: https://doi.org/10.48550/arXiv.1506.00327

[11] J. Faouzi and H. Janati, "pyts: A Python Package for Time Series Classification," J. Mach. Learn. Res., vol. 21, no. 46, pp. 1-6, 2020. [Online]. Available: http://jmlr.org/papers/v21/19-763.html

[12] F. Chollet, "Keras." 2015, GitHub. [Online]. Available: https://github.com/fchollet/keras

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, no. 85, 2825-2830. [Online]. Available: http://www.jmlr.org/papers/v12/pedregosa11a.html

[14] Tanu and D. Kakkar, "Accounting for order-frame length tradeoff of Savitzky-Golay smoothing filters," in 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2018, pp. 805-810, doi: https://doi.org/10.1109/SPIN.2018.8474194

[15] X. Li, Y. Kang, and F. Li, "Forecasting with time series imaging," Expert. Syst. Appl., vol. 160, art. no. 113680, 2020, doi: https://doi.org/10.1016/j.eswa.2020.113680

[16] A. M. Gonzalez-Zapata, L. G. de la Fraga, B. Ovilla-Martinez, E. Tlelo-Cuautle, and I. Cruz-Vega, "Enhanced FPGA implementation of Echo State Networks for chaotic time series prediction," Integration, vol. 92, pp. 48-57, 2023, doi: https://doi.org/10.1016/j.vlsi.2023.05.002

[17] A. Ben Said and A. Erradi, "Deep-Gap: A deep learning framework for forecasting crowdsourcing supply-demand gap based on imaging time series and residual learning," 2019, arXiv:1911.07625, doi: https://doi.org/10.48550/arXiv.1911.07625

[18] W. Ni, C. Zhang, T. Liu, Q. Zeng, L. Xu, and H. Wang, "An efficient astronomical seeing forecasting method by random convolutional Kernel transformation," Eng. Appl. Artif. Intell., vol. 127, art. no. 107259, 2024, doi: https://doi.org/10.1016/j.engappai.2023.107259

[19] J. Marco-Blanco and R. Cuevas, "Time Series Clustering With Random Convolutional Kernels," 2023, arXiv:2305.10457, doi: https://doi.org/10.48550/arXiv.2305.10457

[20] B. Dhariyal, T. Le Nguyen, and G. Ifrim, "Back to Basics: A Sanity Check on Modern Time Series Classification Algorithms," 2023, arXiv:2308.07886, doi: https://doi.org/10.48550/arXiv.2308.07886

[21] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall, "HIVE-COTE 2.0: a new meta ensemble for time series classification," Mach. Learn., vol. 110, no. 11-12, pp. 3211-3243, 2021, doi: https://doi.org/10.1007/s10994-021-06057-9

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 117 - 127

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.8

E-LOCATION ID: 1396

Development of an Adaptive Acquisition and Transmission System for Digital Processing of ECG Signals under Variable n-QAM Schemes

Desarrollo de un Sistema Adaptativo de Adquisición y Transmisión para el Procesamiento Digital de Señales de ECG Bajo Esquemas Variables n-QAM

José Ricardo Cárdenas-Valdez[1], *Fortunato Ramírez-Arzate*[1], *Angel Humberto Corral-Domínguez*[1], *Carlos Hurtado-Sánchez*[1], *Andrés Calvillo-Téllez*[2], *Enrique Efrén García-Guerrero*[3]

[1] Tecnológico Nacional de México - Instituto Tecnológico de Tijuana, Baja California - México
[2] Instituto Politécnico Nacional - CITEDI, Baja California - México
[3] Universidad Autónoma de Baja California, Baja California - México

ABSTRACT

In the post-pandemic era, it is critical to monitor and transmit biomedical signals, specifically ECG. This study aims to develop a platform that enables signal acquisition, adaptation, and transmission using different n-QAM modulation schemes. The system comprises an acquisition stage implemented in the 2.5 GHz band employing the Olimex module and electrodes equipped with an Ag/AgCl type sensor. To effectively manage appropriate bandwidths during implementation of the various n-QAM modulation schemes, an adaptive algorithm was developed and applied to the system. The power amplifier was operated in the linear region to enhance the crest factor and achieve an ACPR close to 30 dBc, demonstrating an appropriate demodulation of the electrocardiogram (ECG) signal, it is feasible to shift to modulation schemes above 64-QAM in order to detect high frequencies and perform a subsequent Fourier analysis. As a telemedicine proposal, the developed system offers flexibility in signal acquisition, data storage, and digitalization, in addition to a multivariable n-QAM scheme; the hardware implementation ensures n-QAM scheme compatibility. For the purpose of contributing to telemedicine via RF transmission, the system was executed on an AD9361 transceiver, which removes the requirement for a traditional signal vector generator and enables optimal control of the tones to be transmitted.

KEYWORDS: ECG, n-QAM, RF, telemedicine, transceiver

RESUMEN

El monitoreo y transmisión de señales biomédicas, particularmente ECG, es fundamental en la era pospandemia, este trabajo de investigación se centra en el desarrollo de una plataforma para la adquisición, adaptación y transmisión de señales bajo diversos esquemas de modulación n-QAM. El sistema incluye una etapa de adquisición mediante el módulo Olimex y electrodos con un sensor tipo Ag/AgCl. Se desarrolló un algoritmo adaptativo a los diversos esquemas de modulación n-QAM para la gestión de anchos de banda apropiados durante una implementación en la banda de 2,5 GHz, al amplificador de potencia se operó en la región lineal para mejorar el factor de cresta y obtener un ACPR cercano a 30 dBc, se realizó una demodulación adecuada de la señal ECG y es posible migrar a esquemas de modulación superiores a 64-QAM si se requiere detectar altas frecuencias y un posterior análisis de Fourier. El sistema desarrollado como propuesta de Telemedicina brinda versatilidad para la adquisición de señales, digitalización, almacenamiento de datos y un esquema multivariable n-QAM, la implementación en hardware proporcionó una adecuada adaptabilidad para esquemas n-QAM. El sistema se implementó sobre un transceptor AD9361 que elimina el tradicional generador vectorial de señal y permite un control óptimo de los tonos a enviar, para aporte en el área de Telemedicina a través de transmisión RF.

PALABRAS CLAVE: ECG, n-QAM, RF, telemedicina, transceptor

Corresponding author

TO: **José Ricardo Cárdenas-Valdez**

INSTITUTION: **Tecnológico Nacional de México/Instituto Tecnológico de Tijuana**

ADDRESS: **Blvd. Industrial s/n, Cd Industrial, C.P. 22430, Tijuana, Baja California.**

EMAIL: **jose.cardenas@tectijuana.edu.mx**

Received:

31 oct 2023

Accepted:

07 dic 2023

INTRODUCTION

Cardiovascular disease is a significant contributor to mortality on a global scale, it is estimated that a third of the causes of death in the world are related to this cardiovascular diseases [1]; in addition to the critical nature of detecting cardiac abnormalities for preventive treatment, the rapid growth of telemedicine development since 2021 as a result of the pandemic further emphasizes this fact, despite the fact that the most significant contributions to this field are older than ten years. Telemedicine faces obstacles, especially in the digital field, connectivity systems, and access to devices. In addition, the learning curve of the diverse range of accessories for sensing and transmission, the most crucial challenge is adopting digital technologies, especially for specialized systems such as radio frequency (RF), and Internet of things (IoT) as the main employed methods in high data rate transmissions [2][3]. The electrocardiogram (ECG) signal analysis is used in various study cases, particularly in people who require constant monitoring of heart rhythms, postoperative monitoring, or high-performance athletes, all of them to detect any pathology or measurement in the time domain of PQ, QRS or QT intervals [4][5]. A biomedical application is devised in [6] that utilizes flexible electrodes to enable field communication that is compatible with digital stages. Previous work has examined the feasibility of wirelessly transmitting biomedical signals using the long-term evolution (LTE) standard. Additionally, for the 3rd Generation Partnership Project (3GPP), a proposition was placed forth to reduce latency on machine-type communication technology [7]. Additionally, detection based on ECG signals is a current challenge, especially for the development of classifiers and detection of cardiac arrhythmia; the process of transmitting signals wirelessly and maintaining the originality of the acquired signal is crucial in diagnosis [8][9]. A fundamental aspect of biomedical applications is spectral analysis, which is utilized to identify and invade adjacent bands or to implement in free bands during radio frequency (RF) transmissions [10]. To reduce the power consumption of synthetic data obtained from the PhysioNet database, an embedded application and a study based on fast Fourier transform (FFT) were created [11].

Digital multiplexes such as orthogonal frequency-division multiplexing (OFDM) have been adopted as a basis in state-of-the-art work in 4G-5G applications. In addition to transmission standards such as LTE, the benefits of OFDM multiplexes are sub-channeling efficiency for handling multipath fading channels and, and consequently efficient per-tone multiple-input multiple-output (MIMO) processing support [12].

A reconfigurable system designed to reduce energy consumption is presented in [13]. This system finds utility in biomedical contexts, specifically in the transmission of quadrature amplitude modulation (QAM), frequency shift keying (FSK), and phase shift keying (PSK), unfortunately additional biomedical information is required to validate the developed system. The development of Telemedicine applications in 5th generation (5G) networks in the post-pandemic era is a key factor for advancements in RF performance [14][15]; in [16] is presented an ECG monitor scheme with two-stage end-to-end improving the power consumption. Additionally, the implementation of high-efficiency amplifiers for large signals under 64-QAM enables the processing of large amounts of data in 5G applications [17].

ECG monitoring offers remote control of acquired data and provides continuous and remote surveillance of patients [18]. This research endeavor provides a proposal for hardware implementation and analysis in the frequency domain of ECG signal monitoring employing n-QAM schemes in the 2.5 GHz band, In the post-pandemic era, examine the spectral purity of an RF transmission for telemedicine applications by evaluating figures of merit and ensuring an appropriate adjacent channel power ratio (ACPR).

This work is organized as follows: Section 2 describes the proposed testbed and digitization stage and the adaptive algorithm. Section 3 shows the results of the hardware implementation in the transceiver. Section 4 summarizes the conclusions of the paper.

MATERIALS AND METHODS

n-QAM adaptive system

Digital multiplexes such as OFDM have been adopted as a basis in state-of-the-art work in 4G-5G applications. In addition to transmission standards like LTE, OFDM multiplexes exhibit inherent benefits that have been particularly developed in [12], extend sub-channeling efficiency, a critical characteristic when handling multipath fading channels. OFDM multiplexes enable effective support to per-tone MIMO processing [12]. With the development of 5G exponentially worldwide, the need to use high transfer rate schemes and multimedia applications, as well as lower resource consumption, the modulation scheme to be used plays a crucial role in the management of wide band bandwidth and power efficiency [19]. N-QAM schemes are used in modern communication standards due to flexibility in bandwidth management and power efficiency [20]. QAM is adaptable and versatile enough to be utilized in the sub-channeling and utilization of information within multiplexation such as OFDM. In the Figure 1 is depicted and example of 64-QAM scheme.

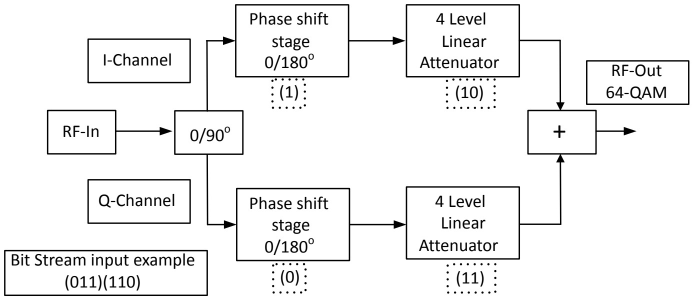


FIGURE 1. Overview of a typical 64-QAM modulator.

In this research work, an adaptable n-QAM scheme is formulated for the purposes of acquiring and preprocessing ECG signals. Within this framework, two channels I and Q, are assumed, whose symbol packaging is of the type Equation (1) and Equation (2).

$$a_k = a_k I + j a_k I, \tag{1}$$

$$b_k = b_k I + j b_k I. \tag{2}$$

The parameters a_k and b_k represent the amplitudes corresponding to the *k - th* symbols within the in-phase (I) and quadrature (Q) channels, respectively. In the context of n-QAM modulation systems, the values of a_k and b_k are selected from the set $\{2i - 1 - f\sqrt{n}\}$ where f takes on values within the range $f = \{1\ldots,\sqrt{n}\}$. Here, f serves as a parameter defining the n-QAM constellation, as expressed by Equation (3):

$$d = \sqrt{\frac{[3log_2(M)E_b]}{2(n-1)}}. \tag{3}$$

In this context, E_b represents the energy per bit, and the per-bit sequence derived from n-QAM is characterized by ck, denoted as the sequence, with w_k representing the pilot sequence. The factor α is intricately linked to the power ratio between the proposed pilot sequence w_k and the power associated with the information sequence bk. It is posited that w_k assumes the form of a periodic sequence of N_s, where the period T_w is equal to $N_s T_s$. The transmitted n-QAM signal is represented in accordance with conventions established by [21][22], denoted by Equation (4).

$$s(t) = \sum c_n p\,(t - nT_s), \tag{4}$$

where *p(t)* is the pulse waveform associated with the transmission process and in the presence of multipath propagation and noise the signal *z(t)* depicted by Equation (5) is received:

$$z(t) = \sum c_n q\,(t - nT_s) + u(t), \tag{5}$$

In this stage *q(t)* the signal filtering stages are taken into account, both the one used in the transmission and reception.

An additional metric of considerable import to assess during the implementation phase is the ACPR. It delineates the power ratio within a telecommunications system by discerning the power in the adjacent channels relative to that within the main bandwidth. ACPR assumes a pivotal role as a primary analytical metric in telecommunications systems, since it evaluates the level of interference that the transmission has in its main bandwidth concerning other bands. In this context, augmenting the ACPR level is imperative to ensure the efficacy of the transmission process. ACPR is represented mainly by Equation (6) involving the Third Order Intercept Point (IP3) and Equation (7) by the average power of the main channels against the adjacent channels.

$$ACPR = -20.75 + 160log\left(\frac{P_{ref}}{P_{adj}}\right) + 2(P_{in} - IP3), \tag{6}$$

$$ACPR(dBc) = 10log_{10}\left(\frac{P_{ref}}{P_{adj}}\right), \tag{7}$$

where P_{ref} is the average power of the main bandwidth, P_{adj} is the average power of the adjacent bands, the crest factor (CF) can also be defined as the power ratio between P_{ref} and P_{adj} .

Figure 2 delineates the spectral distribution of frequency bands, and the average power of the main bandwidth, as well as the adjacent bands. The ACPR serves as a crucial merit figure for assessing the potential for interference with adjacent bands. It is employed to ascertain compliance with regulatory standards, such as those stipulated by the standard for Telecommunications IEEE 802.n [23], to validate that the spectral growth is below the limits established by the FCC. The ACPR, measured in decibels relative to the carrier (dBc), assumes particular significance in the context of mitigating undesirable effects arising from memory and linearity issues inherent in the transmitter stage. This is particularly pertinent to distortions induced by the RF-power amplifier. A robust ACPR value, approaching or exceeding 30 dBc, ensures that unwanted memory and linearity effects of the transmitter stage, particularly those induced by the RF-Power amplifier, have been eliminated.

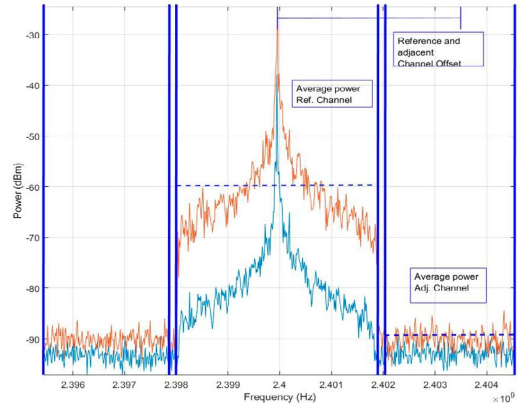


FIGURE 2. Band distribution for ACPR calculation.

ECG acquisition data and RF transmission

The block diagram of the experimental setup is depicted in Figure 3, outlines in the beginning the procedure for obtaining a precordial signal, establishing an AVR: Existence of net potential on the right (aVr), net potential existing on the left side (aVl). and voltage reference (AVF) type lead placement to form the Einthoven triangle for acquiring P, QRS, and T waves [24]. In this case, the used electrodes include polyethylene, they are biocompatible, coated with acrylic copolymer, with an AgCl type sensor; conductive gel was placed on the terminal to improve the acquisition of the ECG signal. In the acquisition stage, an Olimex module is used coupled to the digital I/O port with the Arduino Uno embedded system; the design was built in C++ language in a Java environment. The storage between boards allows heart peak detection, calculation of sample based heart rate, and adequate accuracy of the extracted biomedical signals.

Variations in the magnitude of voltage compensation, amplitude variation, and data array length obtained by the embedded system ECK/EKG board are accounted for in an adaptive stage of Algorithm 1. The algorithm commences by establishing a m-ary number associated with the defined QAM modulation scheme. It then proceeds to identify the length of the array and normalizes it to a unitary amplitude, which is subsequently processed by the AD9361 Toolkit. Furthermore, it supports an amplitude range of 2 to 12 for one of the 2 × 2 channels transceivers that incorporate a 12-bit DAC and ADC. In the M programming language, the digitized and transmitted signal is stored as an environment variable. The data to be transmitted and before demodulation is packaged in n-QAM order packets. The demodulated signal is subsequently encoded as a variable within a Matlab environment. This enables mathematical classifiers for biomedical signals and analysis in the frequency domain.

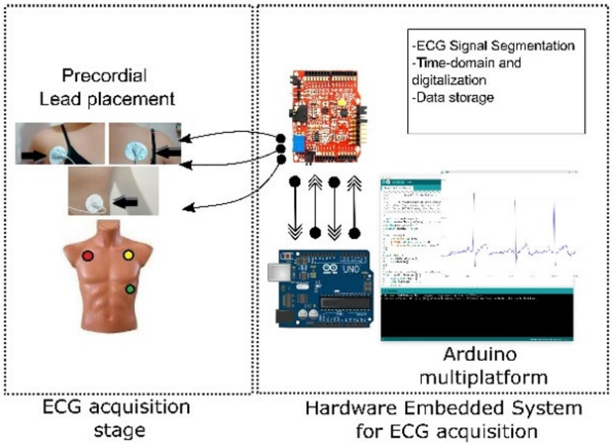


FIGURE 3. **ECG acquisition within an embedded multiplatform.**

As an intermediate stage, the ECG signal segmentation process is carried out based on Algorithm 1, and storage of the digitized variables in the processing of the data to operate the FMCOMMS3 RF transceiver, a variable n-QAM stage is designed; the characteristics of bandwidth (BW), gain mode sampling frequency are adjusted in the AD9361 transceiver toolkit, the transmitter stage comprises an n-QAM modulator, coupled with a raise cosine filter coupled prior to transmission in Matlab environment.

ALGORITHM **1. ECG signal segmentation and n-QAM digital packets.**

```
Require: $n \geq 1$
Ensure: $ECG_{QAM} = ECGstream/n$
  while N= 0 and N ≤length(ECGadq) do
      if N is integer then
          $ECGoffset = ECGadq - min(ECGadq)$
          $ECGnorm = ECGoffset/max(ECGadq)$
          $ECGstream = Dec2Bin(ECGnorm)$
  end if
  StoreDigitalizedSignal ← ECGstream
end while
```

In this work, an RF power amplifier ZX60-5916MA-S+ is used as a device under test (DUT); it is a proper device for ultra-wide band applications when is required high linearity, including the 2.5 GHz band, it improves lower system noise figure performance. The transceiver used has a carrier frequency of 2.45 GHz, the ARRadio card includes the AD9361 transceiver, the modulator system is variable at n-QAM levels. In this case, quadrature phase shift keying (QPSK) was used as the fundamental modulation into the Cyclone V FPGA SoC-Kit development board. The development board provides full control of the transmitted tones, which replaces the traditional signal vector generator (SVG), and the RF transmission is validated in the Spectrum Analizer SSA-3032X. The testbed is protected due to the coupled port attenuation of -20.73 and the attenuator BW-S10W5+ of 10 dB. The directional coupler as a passive device provides high isolation between the medium power levels and the digital schemes, improving the reflected signal transmitted signal.

Figure 4 delineates the schematic representation of the digitization stage, baseband adaptation, and digital modulation processes, data preprocessing procedures that precede the transmission through the AD9361 transceiver. Furthermore, the signal is directed to a ZFBP-2400-S+ low-pass filter designed for operation within the 2.5 GHz band. A directional coupler, specifically the ZHDC-16-63-S+, is employed to mitigate line feedback and suppress spurious noise within the system. Notably, the RF power amplifier as a device under test is the RF-PA ZX60-5916MA-S+. Concurrently, a spectral evaluation of the transmission is carried out using the SSA 3032X spectrum analyzer.

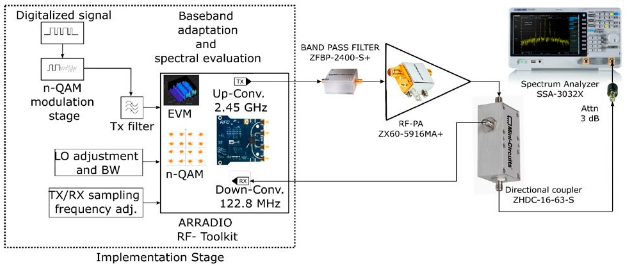


FIGURE 4. Schematic representation of digitization, baseband adaptation and transmission of the AD9361 transceiver.

Figure 5 depicts the devised test bench designed for the comprehensive handling of data storage, preprocessing, and transmission processes, inclusive of the AD9361 RF Agile Transceiver + FPGA-SocKit transceiver. The ARRADIO board, serving as a pivotal component, is purposed to furnish a platform for RF full-tone control. In this specific instantiation, a balun operating within the 2.4-2.5 GHz ranges is employed. The utilization of advanced transceivers enables precise tone regulation, obviating the need for expensive SVR. Furthermore, this setup facilities the discretization, storage, and processing of data.

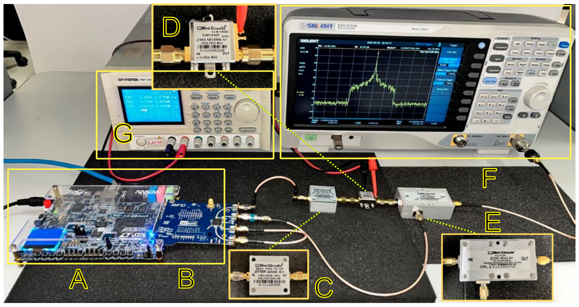


FIGURE 5. Overview of the developed experimental testbed. (A-B) AD9361 RF Agile Transceiver + FPGA-SocKit. (C) ZFBP-2400-S+ 50Ω Narrow bandpass Filter. (D) ZX60-5916MA-S+ RF-Power Amplifier. (E) ZHDC-16-63-S+ Directional Coupler. (F) Spectrum Analyzer SIGLENT SSA3032X.

In the hardware transmission stage, a ZFBP-2400-S+ 50Ω narrow-bandpass filter is integrated, with a pronounced rejection capability of 50 dB, affording protective measures against undesired transmissions in the sidebands. Subsequently, in the intermediate stage, a ZHDC-16-63-S+ high directivity directional coupler is employed, characterized by superior coupling flatness and providing requisite attenuation to mitigate line return effects during the transmission process.

RESULTS AND DISCUSSION

One of the main challenges of Telemedicine applications through RF transmission is the low nonlinearity involved in the transmitter stage; a critical consideration is to ensure that biomedical signals operate within power regimes below the P1dB threshold, thereby mitigating intrinsic nonlinearities. The RF-PA model ZX60-5916MA-S+ operates under the P1dB threshold of 10.63 dBm, ensuring heightened linearity without compromising power during RF transmission. The active directivity of the DUT particularly suitable for minimizing noise in the receiver stage and facilitating the recovery of the demodulated signal.

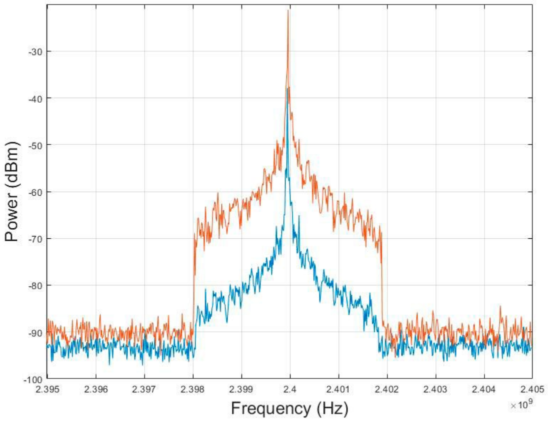


FIGURE 6. Spectral analysis of the acquired ECG signal compared to that amplified by the DUT.

Figure 6 illustrates the relation between input power and output power, emphasizing the utilization of the RF-PA within its linear operational range for optimal power efficiency in telemedicine transmission applications. At the central point of the bandwidth, a power level of -27.9 dBm is noted, while the output power is reached at -11.2 dBm, resulting in the DUT operating with a gain of 16.7 dB.

Figure 7 depicts the acquired signal derived from the embedded ECG/EKG board, wherein the system comprises an acquisition stage utilizing the Olimex module in conjunction with electrodes featuring an AgCl-type sensor. An algorithm has been devised for the discernment of peaks within cardiac signals, facilitating the determination of heart rate, and the subsequent computation of sample-based heart rate.

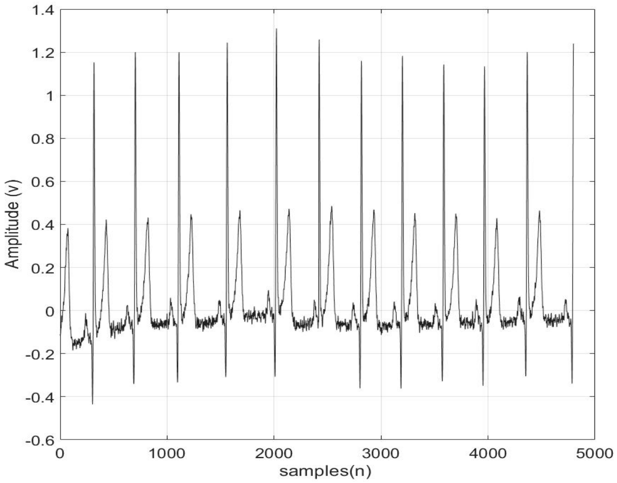


FIGURE 7. **signal recovered in the time domain under that 32-QAM modulation scheme.**

Figure 8 presents the signal recuperated under the 16-QAM scheme, wherein each cycle comprises 16 constituent parts constituting the signal. It is noteworthy that despite the 4-bit packaging characteristic of 16-QAM, the scheme leverages speed advantages during the transmission process. Subsequently, Figure 9 illustrates the signal recovery under the 32-QAM modulation scheme, involving a discretization of 32 segments and a packaging configuration of 5 bits per event.

In addition, the 64-QAM modulation scheme is illustrated, (Figure 10), featuring a significant discretization of the signal and an enhancement of -20 dB in the NMSE. It has been determined that the integration of higher-order n-QAM schemes into subsequent procedures may have the potential to improve NMSE, particularly when high induced frequencies in the signal are required for Fourier domain analysis.

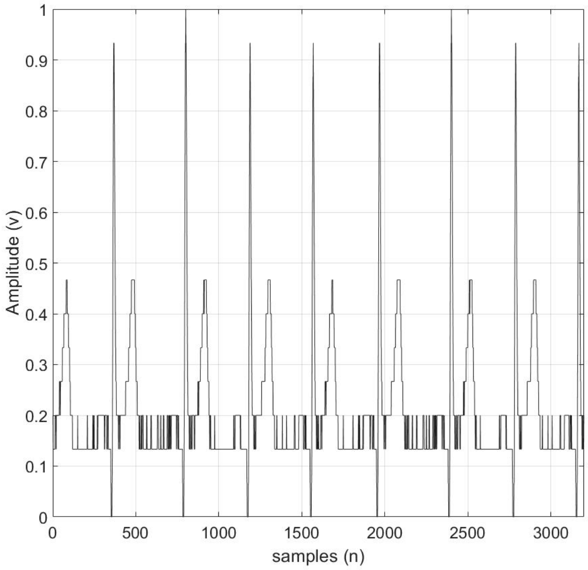


FIGURE 8. **The time domain signal recovered by the 16-QAM modulation scheme.**

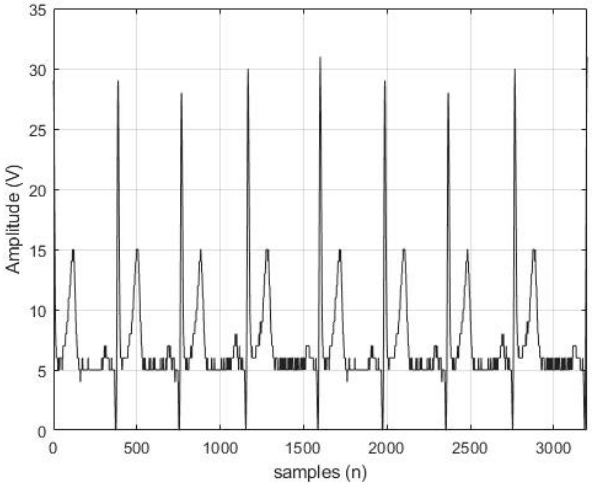


FIGURE 9. **The time domain signal recovered by the 32-QAM modulation scheme.**

Spectral validation is critical to ensure that there is no invasion of adjacent bands during the transmission process; ACPR is a figure of primary merit for this analysis since it defines the relationship of the average power between the main bandwidth and the side channels, the signal is shown in Figure 11; acquired from the transmitter showing -27.89 dBm due to the 10 dB attenuator located at the input port of the Spectrum Analyzer SIGLENT SSA3032X, and Figure 12 shows the signal amplified by the DUT ZX60-5916MA-S+ with a power of -11.21 dBm given that it passes through the 10 dB attenuator of the input port.

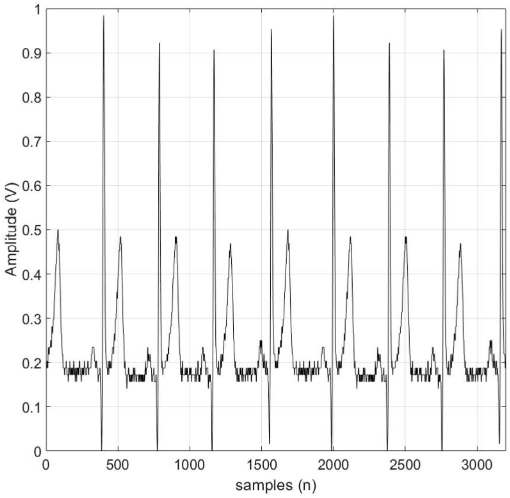
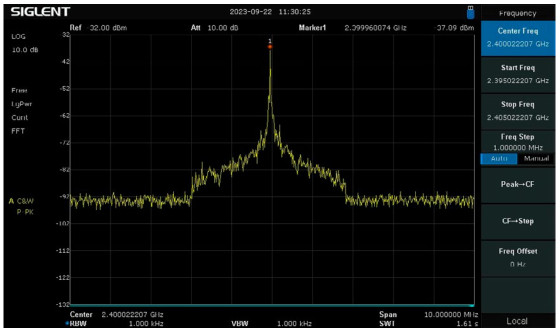
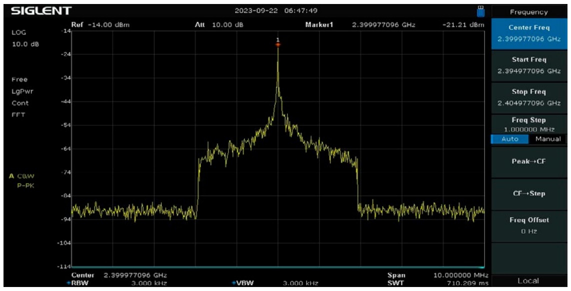


FIGURE 10. **The time domain signal recovered by the 64-QAM modulation scheme.**



Ref -32.00 dBm
Att 10.00 dB
Marker1 2.399960074 GHz -37.09 dBm
Center Freq 2.400022207 GHz
Start Freq 2.395022207 GHz
Stop Freq 2.405022207 GHz
Freq Step 1.000000 MHz
Freq Offset 0 Hz
Center 2.400022207 GHz
Span 10.000000 MHz
RBW 1.000 kHz
VBW 1.000 kHz
SWT 1.61 s

FIGURE 11. **Transmitted stimulus under a 10 dB attenuation.**



Ref -14.00 dBm
Att 10.00 dB
Marker1 2.399977096 GHz -21.21 dBm
Center Freq 2.399977096 GHz
Start Freq 2.394977096 GHz
Stop Freq 2.404977096 GHz
Freq Step 1.000000 MHz
Freq Offset 0 Hz
Center 2.399977096 GHz
Span 10.000000 MHz
RBW 3.000 kHz
VBW 3.000 kHz
SWT 710.289 ms

FIGURE 12. **Amplified signal under a 10 dB attenuation.**

Table 1 describes the metrics achieved during the implementation. A bandwidth of 10 MHz is used because the transmission of ECG signals involves a relatively low level of BW for digital modulation, which provides flexibility to migrate to higher bandwidths if the monitoring of several biomedical signals and transmission is required, a crest factor of 718.85 is obtained showing an appropriate relationship of the main BW concerning the power of the adjacent bands, the achievement of an ACPR of 28.56 dBc enables the implementation of modulation schemes that ensure the highest order n-QAM.

TABLE1. **Power spectral validation**

RF-PA spectral validation	
Average power of main BW	-61.546 dBm
Average power of the adjacent channel	-90.112 dBm
Crest Factor	718.85
BW power validation	7.005×10^{-7} mW
Adjacent channels power validation	9.745×10^{-10} mW
ACPR	28.56 dBc
BW	10 MHz

CONCLUSIONS

This research work introduces a substantive contribution to the field of Telemedicine, characterized by key stages encompassing signal acquisition facilitated by an Olimex module along with electrodes featuring an AgCl-type sensor. Additionally, a M-language algorithm is formulated for the segmentation and storage of biomedical signals. The digitized ECG signal undergoes modulation utilizing a variable n-QAM scheme, with a designated bandwidth of 10 MHz chosen to accommodate the periodic nature of the signals during transmission. By selecting this option, the crest factor is decreased, which encourages the consideration of implementing higher QAM schemes. The application of n-QAM modulation transcends ECG signals and facilitates the transmission of real-time data while managing a wide variety of biological signals.

The transceiver system's intrinsic regulation of the transmission process eliminates the need for expensive signal vector generators in the laboratory, thereby guaranteeing accuracy during signal acquisition, preprocessing, and data transmission. In addition, a flexible algorithm is developed for handling various n-QAM orders in order to discretize, store, and package data within the OFDM framework.

The transition of systems to 5G requires the implementation of adaptive platforms that are capable of meeting digital modulations, consume just the necessary bandwidth, and follow the crest factor and ACPR limits. These limits are imposed by the RF power amplifier being evaluated, which is susceptible to spectral growth resulting from nonlinearities and memory effects.

The platform and algorithm developed are an alternative in Telemedicine applications because they allow the acquisition of biomedical signals and wireless transmission, especially in rural areas where the doctor is located remotely or his location is centralized for monitoring several towns.

AUTHOR CONTRIBUTIONS

J.R.C.V. developed the methodology, participated in the design and development of the algorithms used, oversaw the experiments and participated in the analyses of results. F.R.A. participated in the data gathering and processing, use of specialized software. A.H.C.D. developed hardware and software for data gathering. C.H.S. design and developed programing language and code for data gathering and analysis. A.C.T. design and developed algorithms for data gathering and validated data. E.E.G.G. validated data and results for the discussion. All authors participated in all the writing stages of the manuscript, reviewed and approved the final version of the manuscript.

REFERENCES

[1] World Health Organization, “Cardiovascular diseases (CVDs),” WHO. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] A. Ibaida, A. Abuadbba, N. Chilamkurti, “Privacy-preserving compression model for efficient IoMT ECG sharing,” Comput. Commun., vol. 166, no. 15, pp. 1-8, Jan. 2021, doi: https://doi.org/10.1016/j.comcom.2020.11.010

[3] C. De Capua, A. Meduri and R. Morello, “A Smart ECG Measurement System Based on Web-Service-Oriented Architecture for Telemedicine Applications,” IEEE Trans. Instrum. Meas., vol. 59, no. 10, pp. 2530-2538, Oct. 2010, doi: https://doi.org/10.1109/TIM.2010.2057652

[4] J. Francis, “ECG monitoring leads and special leads,” Indian Pacing Electrophysiol. J., vol. 16, no. 3, pp. 92-95, 2016, doi: https://doi.org/10.1016/j.ipej.2016.07.003

[5] I. -A. Ivanciu, L. Ivanciu, D. Zinca, and V. Dobrota, “Securing Health-Related Data Transmission Using ECG and Named Data Networks,” in 2019 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), Paris, France, 2019, pp. 1-6, doi: https://doi.org/10.1109/LANMAN.2019.8846993

[6] M. Zulqarnain, S. Stanzione, G. Rathinavel, S. Smout, M. Willegems, K. Myny, and E. Cantatore, “A flexible ECG patch compatible with NFC RF communication,” npj Flex. Electron., vol. 4, art. no. 13, Jul. 2020, doi: https://doi.org/10.1038/s41528-020-0077-x

[7] Y. Cho, H. Shin, and K. Kang, “Scalable Coding and Prioritized Transmission of ECG for Low-Latency Cardiac Monitoring Over Cellular M2M Networks,” IEEE Access, vol. 6, pp. 8189-8200, Jan. 2018, doi: https://doi.org/10.1109/ACCESS.2018.2795028

[8] N. Clark, E. Sandor, C. Walden, I. S. Ahn and Y. Lu, “A wearable ECG monitoring system for real-time arrhythmia detection,” in 2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), Windsor, ON, Canada, 2018, pp. 787-790, doi: https://doi.org/10.1109/MWSCAS.2018.8624097

[9] G. Cosoli, S. Spinsante, F. Scardulla, L. D'Acquisto and L. Scalise, “Wireless ECG and cardiac monitoring systems: State of the art, available commercial devices and useful electronic components,” Measurement, vol. 177, art. no. 109243, Jun. 2021, doi: https://doi.org/10.1016/j.measurement.2021.109243

[10] S. M. Noor, E. John, and M. Panday, “Design and Implementation of an Ultralow-Energy FFT ASIC for Processing ECG in Cardiac Pacemakers,” IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 27, no. 4, pp. 983-987, Apr. 2019, doi: https://doi.org/10.1109/TVLSI.2018.2883642

[11] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, pp. e215-e220, 2000, doi: https://doi.org/10.1161/01.cir.101.23.e215

[12] C. Kim, Y. H. Yun, K. Kim, and J. -Y. Seol, “Introduction to QAM-FBMC: From Waveform Optimization to System Design,” IEEE Commun. Mag., vol. 54, no. 11, pp. 66-73, Nov. 2016, doi: https://doi.org/10.1109/MCOM.2016.1600384CM

[13] C. -H. Heng and K. -H. Teng, “Reconfigurable, energy efficient transmitter with band-shaping and multi-channel support for biomedical applications,” in 2016 URSI Asia-Pacific Radio Science Conference (URSI AP-RASC), Seoul, Korea (South), 2016, pp. 986-989, doi: https://doi.org/10.1109/URSIAP-RASC.2016.7601308

[14] B. O. hAnnaidh, P. Fitzgerald, H. Berney, R. Lakshmanan, N. Coburn, S. Geary, B. Mulvey, “Devices and Sensors Applicable to 5G System Implementations,” in 2018 IEEE MTT-S International Microwave Workshop Series on 5G Hardware and System Technologies (IMWS-5G), Dublin, Ireland, 2018, pp. 1-3, doi: https://doi.org/10.1109/IMWS-5G.2018.8484316

[15] T.-H. Tsai and W.-T. Kuo, “An Efficient ECG Lossless Compression System for Embedded Platforms With Telemedicine Applications,” IEEE Access, vol. 6, pp. 42207-42215, Jul. 2018, doi: https://doi.org/10.1109/ACCESS.2018.2858857

[16] N. Wang, J. Zhou, G. Dai, J. Huang, and Y. Xie, “Energy-Efficient Intelligent ECG Monitoring for Wearable Devices,” IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 5, pp. 1112-1121, Oct. 2019, doi: https://doi.org/10.1109/TBCAS.2019.2930215

[17] K. Nakatani, Y. Yamaguchi, Y. Komatsuzaki, S. Sakata, S. Shinjo, and K. Yamanaka, “A Ka-Band High Efficiency Doherty Power Amplifier MMIC using GaN-HEMT for 5G Application,” in 2018 IEEE MTT-S International Microwave Workshop Series on 5G Hardware and System Technologies (IMWS-5G), Dublin, Ireland, 2018, pp. 1-3, doi: https://doi.org/10.1109/IMWS-5G.2018.8484612

[18] D. Lucani, G. Cataldo, J. Cruz, G. Villegas, and S. Wong, “A portable ECG monitoring device with Bluetooth and Holter capabilities for telemedicine applications,” in 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 2006, pp. 5244-5247, doi: https://doi.org/10.1109/IEMBS.2006.260798

[19] P. K. Singya, P. Shaik, N. Kumar, V. Bhatia, and M.-S. Alouini, “A Survey on Higher-Order QAM Constellations: Technical Challenges, Recent Advances, and Future Trends,” in IEEE Open J. Commun. Soc., vol. 2, pp. 617-655, 2021, doi: https://doi.org/10.1109/OJCOMS.2021.3067384

[20] T. S. Rappaport, Wireless Communications: Principles and Practice, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall PTR, 1996.

[21] F. Mazzenga, “Channel estimation and equalization for M-QAM transmission with a hidden pilot sequence,” IEEE Trans. Broadcast., vol. 46, no. 2, pp. 170-176, 2000, doi: https://doi.org/10.1109/11.868934

[22] J. G. Proakis and M. Salehi, Digital Communications, 2nd ed. NJ, USA: McGraw-Hill, Higher education, 2008.

[23] “IEEE Standard for Telecommunications and Information Exchange Between Systems - LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: High Speed Physical Layer in the 5 GHz band,” in IEEE Std 802.11a-1999, pp.1-102, 1999, doi: https://doi.org/10.1109/IEEESTD.1999.90606

[24] B. Abi-Saleh, B. Omar, “Einthoven’s Triangle Transparency: A Practical Method to Explain Limb Lead Configuration Following Single Lead Misplacements,” Rev. Cardiovasc. Med.,” vol. 11, no. 1, pp. 33-38, 2010, doi: https://doi.org/10.3909/ricm0506

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 128 - 139

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.9

E-LOCATION ID: 1392

Aplicación de Redes Neuronales Artificiales para la Clasificación de Actividades de la Vida Diaria en Sujetos con Enfermedad de Párkinson

Classification of Daily Living Activities in subjects with Parkinson's Disease using Artificial Neural Networks

Lizeth Rodriguez Montero[1], *Jose Ambrosio Bastián*[1], *Alberto Isaac Pérez Sanpablo*[1,2]

[1] Universidad La Salle, Ciudad de México - México

[2] Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra, Ciudad de México - México

RESUMEN

El presente trabajo es un seguimiento a la propuesta para la contribución con especialistas en la salud para enriquecer los sistemas de seguimiento y apoyo en pacientes con Enfermedad de Párkinson a través de la clasificación de actividades de la vida diaria (AVDs) utilizando Redes Neuronales Artificiales programadas en lenguaje Python. El método propuesto de aprendizaje supervisado permitió la clasificación de 6 AVDs mediante 22 señales procedentes de haber aplicado Análisis de Componentes Principales; conformando la base de datos utilizada para entrenar un Perceptrón Multicapa, logrando un acercamiento a la clasificación con el 93% de medida F1-score. El presente estudio demuestra la versatilidad de las RNA basadas en MLP combinadas con la técnica de PCA, pues incluso en una base de datos desbalanceada como la utilizada permite alcanzar excelentes valores en la medida F1-score. El uso de Inteligencia Artificial y otras herramientas aplicadas en este trabajo pueden eventualmente ayudar a especialistas a desempeñar una evaluación más certera en el monitoreo de la rehabilitación en pacientes con enfermedad de Párkinson mejorando los registros y así evitar subjetividad en la interpretación de los resultados del tratamiento.

PALABRAS CLAVE: enfermedad de Párkinson, PCA, redes neuronales artificiales

ABSTRACT

This paper is a proposal to contribute with health specialists to enrich the follow-up and support systems in patients with Parkinson's by identifying and classifying Daily Living Activities (DLAs) using Artificial Neural Networks programmed in Python language. The proposed method of supervised learning allowed the classification of 6 DLAs through 22 signals obtained from the application of Principal Component Analysis, creating a database used to train a Multilayer Perceptron. This model achieved an approximation of classification with 93% of the F1-score. The present study demonstrates the versatility of ANNs based on MLP combined with the PCA technique since, even in an unbalanced database such as the one used, it allows excellent values to be achieved in the F1-score measure. The use of Artificial Intelligence and other tools applied in this work may eventually help specialists to perform a more accurate assessment in the monitoring of rehabilitation for patients with Parkinson's disease by improving records and thus avoiding subjectivity in the interpretation of treatment results.

KEYWORDS: artificial neural network, Parkinson’s disease, PCA

Autor de correspondencia

DESTINATARIO: **Alberto Isaac Pérez Sanpablo**
INSTITUCIÓN: **Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra / Universidad La Salle**
DOMICILIO: **Calz México-Xochimilco 289, Coapa, Col. Arenal de Guadalupe, Tlalpan, 14389 Ciudad de México, CDMX**
CORREO ELECTRÓNICO: **albperez@inr.gob.mx**

Recibido:

31 oct 2023

Aceptado:

08 dic 2023

INTRODUCCIÓN

La Enfermedad de Parkinson (EP) se caracteriza por ser una afección crónica y progresiva provocada por el deterioro neuronal en la sustancia nigra y la disminución de los niveles de dopamina en el cuerpo [1]. La enfermedad de Párkinson no es curable, por lo que el objetivo de los tratamientos es exclusivamente sintomático [2]. El manejo farmacológico es el principal tratamiento del paciente con EP. Se requiere de un tratamiento que incluya diversas disciplinas en su manejo para promover la participación del paciente en la sociedad y fomentar la actividad física desde el momento del diagnóstico con la finalidad de adquirir estrategias de movimiento antes de presentar limitaciones físicas y cognitivas más severas [3]. A pesar del óptimo tratamiento 70 % de los pacientes con enfermedad de Parkinson experimentan momentos donde los síntomas motores como tremor, bradicinesia e inestabilidad postural empeoran, estos momentos son conocidos como fluctuaciones motoras [4].

El seguimiento de las fluctuaciones de los síntomas motores en pacientes con enfermedad de Párkinson se realiza actualmente a través de la autoevaluación subjetiva de los pacientes. Sin embargo, los médicos y especialistas requieren información objetiva y más confiable sobre la ocurrencia de las fluctuaciones para permitir una planeación precisa del tratamiento [5].

La evaluación mediante sensores inerciales de los signos motores y fluctuaciones, pueden proporcionar retroalimentación objetiva y confiable sobre los efectos del manejo terapéutico [6]. A la fecha este tipo de sensores han sido estudiados para detectar actividades de la vida diaria, principalmente caminar y en pocos casos acostarse, sentarse y mantenerse de pie [7][8][9]. En estas investigaciones se han utilizado diversos tipos, números, ubicaciones de sensores y métodos de clasificación variables sin alcanzar un consenso a la fecha. Principalmente se ha utilizado una configuración de un sensor predominantemente en la cadera y de tipo acelerómetro [7][8][9]. Aunque los primeros experimentos se concentraban en la validación de laboratorio de los registros obtenidos con los sensores inerciales, actualmente el interés parece concentrarse en probar nuevos métodos de clasificación centrados en un propósito clínico, con precisiones reportadas entre un 55.9% y 90.8% [8][9][10].

La clasificación de actividades de la vida diaria utilizando Redes Neuronales Artificiales (RNA) ha demostrado el alto potencial de aplicar *Deep Learning* a la resolución de los problemas en la clasificación de señales realizadas por pacientes con Enfermedad de Párkinson. Estas herramientas han ayudado a desarrollar nuevas posibilidades de apoyo para que los especialistas puedan llevar a cabo evaluaciones más certeras de la funcionalidad motora y de los tratamientos proporcionados mediante una mejor calidad de monitoreo en la rehabilitación de los pacientes. El uso de inteligencia artificial para ayudar a la detección y tratamiento de la EP es de creciente interés para médicos e investigadores alrededor del mundo [11]. Actualmente se está trabajando en el desarrollo de diversos métodos de inteligencia artificial basados en algoritmos de aprendizaje profundo que permitan la detección y poder hacer un seguimiento de los principales síntomas motores de la Enfermedad de Párkinson [12].

De acuerdo con el estudio donde las RNA son utilizadas para clasificar las AVDs desempeñadas por 5 pacientes con Enfermedad de Párkinson, se obtuvo un porcentaje de exactitud del 99 % en el aprendizaje de clasificación [13]. Por lo que en este trabajo se busca obtener un método con resultados semejantes, pero realizando reducción de dimensionalidad de la base de datos a través de la extracción de la información más relevante para la clasificación de actividades de la vida diaria utilizando sensores inerciales.

MATERIALES Y MÉTODOS

Origen y adecuación de datos

Este trabajo fue desarrollado utilizando una base de datos (BD) de 179,755 observaciones por 64 columnas, empleada y descrita en investigaciones previas [9][13]. Brevemente, para crear la base de datos se reclutó a tres individuos con EP, con una edad promedio de 63.5 años y un estadio Hoehn y Yahr de dos puntos. La base de datos contiene información de los individuos mientras hacían en orden aleatorio un circuito de actividades de la vida diaria como mantenerse de pie, caminar, sentarse, subir escaleras, acostarse y caminar con una charola en las manos. Durante el registro se capturó información de cuatro sensores inerciales colocados en pie, muslo, pelvis y muñeca derecha a una frecuencia de 250 Hz. Esta información fue preprocesada con un filtro Butterworth pasa-bajas de segundo orden y una frecuencia de corte de 5 Hz y se utilizó una ventana móvil de 6.4s para calcular diferentes métricas como promedio, mínimo, máximo, desviación estándar, valor RMS de la señal en cada eje y resultante de los sensores que componen cada sensor inercial.

En esta investigación, se desarrolló un entorno de trabajo en *Anaconda*, donde se importó la base de datos que dispone de 179,755 observaciones y 64 variables definidas como las entradas [13]. A diferencia de la metodología reportada con anterioridad, una vez importada y reagrupada la base, se le aplicó aleatorización y posteriormente se guardaron las entradas en la variable *dataset_01x*; y el vector 65 se guardó en la variable *dataset_01y*; caracterizado por contar con las etiquetas de los tipos de AVDs desempeñadas por los sujetos con EP.

Posteriormente se hizo la división de la información donde el 50% de las entradas se emplearon para entrenamiento y el otro 50% restante para prueba [13].

Análisis de Componentes Principales

Buscando reducir la dimensionalidad de la base de datos, se aplicó Análisis de Componentes Principales *(Principal Component Analysis, PCA)* mediante instrucciones de programación en Python. Esta herramienta permite extraer las características más relevantes para mantener un porcentaje elevado de significancia de la base para así eliminar aquellas variables que no aporten información tan relevante [14].

El PCA es una técnica ampliamente utilizada para reducir la dimensionalidad de conjuntos de datos complejos sin pérdida de información a través de la representación del conjunto de datos original en otro nuevo conjunto de nuevas variables no correlacionadas llamadas componentes principales *(Principal Components, PC)* [15]. Los PC son útiles para representar de forma más compacta los datos. Los PC se obtienen a partir del cálculo de los valores propios del conjunto de datos original por lo que representan direcciones en el espacio de datos originales, lo que la hace del PCA una técnica altamente adaptable a diferentes tipos y estructuras de datos [15]. La técnica de PCA ha sido utilizada previamente para análisis de la marcha en pacientes con enfermedad de Parkinson [16][17].

Como primeros pasos para emplear este método, se normalizaron los datos para eliminar diferencias en las escalas de las características o variables de la base de datos original. En la Figura 1 se tiene un diagrama de flujo de la metodología empleada para el diseño de la red neuronal artificial. Los puntos 1 a 3 pertenecen al origen y la adecuación de los datos, a partir del punto 4 se emplean instrucciones específicas para las entradas y salidas.

Partiendo del paso 4-A y 4-B de la Figura 1, se pueden observar los procesos que permiten adaptar las entradas de entrenamiento X a un escalamiento de tal manera que su media sea 0 y la varianza sea 1. Posteriormente el PCA calcula la matriz de covarianza $S=X^TX$, que representa las relaciones entre las características de la matriz de datos normalizada X [15]. A continuación, se determinan los valores propios de la matriz S y se ordenan en orden descendente. Se seleccionan los primeros k, correspondientes a las direcciones más importantes y se construye una matriz de proyección W utilizando los valores propios seleccionados. Para fines de este trabajo, en el paso 6, se planteó mantener los primeros k valores para conservar el 95 % de varianza de la información. Finalmente se transforma el conjunto de datos originales en un nuevo subespacio de características de la forma T=XW [15].

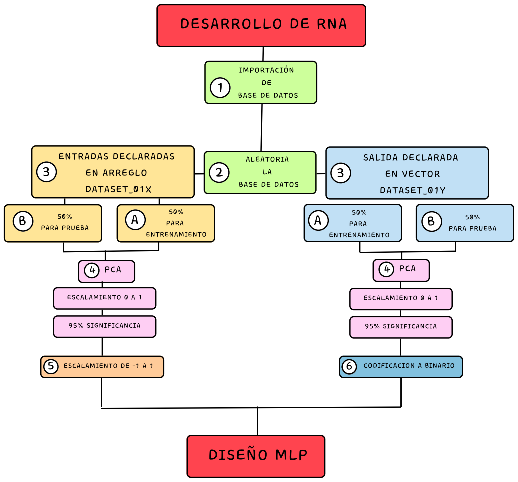


FIGURA 1. **Metodología empleada para el diseño de la RNA.**

Diseño RNA tipo Perceptrón Multicapa

El diseño de la RNA se basa en el perceptrón de capas múltiples (MLP) el cual es un diseño muy versátil y ampliamente utilizado de RNA feedforward que consta de una capa de entrada, una de salida y una o múltiples capas ocultas que pueden implementarse en hardware [18][19][20]. La operación de un MLP con n capas sobre un vector de datos de entrada x se puede representar como una secuencia de funciones $f(x)=f^{(n)}(f^{(n-1)}(\ldots(f^{(2)}(f^{(1)}(x))$. En cada capa n se realiza una transformación y=f(Wx+b), donde W es una matriz de pesos, b es el vector de sesgo y f es la función de activación [18]. La función de activación introduce no linealidades a la red como una tangente hiperbólica (tanh(z)), o una función de activación rectificada RELU = max(0,z). De esta forma para cada vector de entrada x la red calcula la salida de cada neurona de cada capa n hasta llegar a la capa de salida. Al llegar a la capa de salida, se calcula el error respecto a las salidas deseadas (función de pérdida) minimizado mediante un algoritmo de optimización de descenso de gradiente y propagando desde la capa de salida hasta la capa de entrada durante el aprendizaje [18]. De esta forma durante el aprendizaje los pesos y sesgos de la red se ajustan para minimizar la función de pérdida.

Estudios previos han mostrado la utilidad de MLP para la clasificación y evaluación de los síntomas y actividades en sujetos con enfermedad de Parkinson. En específico el uso de MLP en combinación con técnicas de remuestreo ha logrado mejorar la clasificación de la gravedad del temblor de la enfermedad de Parkinson, alcanzando 93.81 % de precisión general [21]. Otro estudio logró detectar la bradicinesia utilizando MLP y relojes inteligentes con una precisión de 0.77 [22]. En otro estudio utilizando un sensor inercial y MLP se logró una precisión del 100 % para distinguir entre estados "encendidos" y "apagados" para terapia de estimulación cerebral profunda [23].

El diseño de la RNA actual considera el diseño de una RNA reportada previamente, la cual fue entrenada utilizando la base de datos completa BD [13]. Así el diseño inicia con el preprocesamiento de la base de datos que incluye el escalamiento de -1 a 1 de las entradas y la codificación a binario One Hot Encoder (OHE) para las etiquetas de salida [13].

Posteriormente se declaró un modelo de red secuencial con 2 capas ocultas; la primera capa utilizó 16 neuronas ejecutadas con la Función de Activación (FA) Rectified Linear Units (ReLu, por sus siglas en inglés) la cual ha demostrado proporcionar mejoras en el entrenamiento de una red neuronal artificial [24]. Para la segunda capa oculta fue declarada con 10 neuronas y FA *Sigmoid*. (tradicionalmente conocida en la teoría e implementación de las RNA [25]); a dicho modelo se le aplicó el optimizador *Adam*. Este modelo requirió 1000 épocas para su procesamiento y un *batch size* de 250.

Comportamiento de red por subconjunto de sujetos

El desempeño de cualquier clasificador se basa en la representatividad de los datos de entrenamiento. Sin embargo, esto puede ser difícil de lograr en casos como la Enfermedad de la Enfermedad de Parkinson, donde la manifestación de síntomas es heterogénea y variante a lo largo del tiempo. Incluso se ha documentado la posibilidad de confundir esta enfermedad con síndromes parkinsonianos, por lo cual se han propuesto subcategorías con compromiso en la especificidad y sensibilidad de su clasificación en base a características clínicas como la presencia de temblor en reposo y afectación asimétrica [26]. Para explorar el desempeño de nuestro clasificador cuando se aplica en un subconjunto de sujetos pertenecientes a la misma población clínica pero tal vez con diferencias no perceptibles en su desempeño motor, se realizó un experimento de elaboración de un desarrollo de un clasificador con entrenamiento y prueba en subconjuntos diferentes de sujetos. El estudio de la precisión de la clasificación entre grupos demográficos ha sido abordado previamente. Un estudio previo mostró una disminución de precisión en la clasificación de felicidad asociada a características demográficas del subconjunto de sujetos evaluados [27].

Para simular condiciones donde la red es entrenada con un conjunto de sujetos y probada en otro conjunto que puede ser diferente, se propuso evaluar el comportamiento de la red con una base de datos que contenga las señales del 80 % de los sujetos de estudio en la fase de entrenamiento; y las señales del 20 % de los sujetos utilizarlas en la etapa de prueba.

Para esta exploración de la base de datos original, se extrajeron las observaciones 1 a 126,073; pertenecientes al 80 % de sujetos [9]. Dichas señales se guardaron en una nueva base de datos denominada “mat01” y las señales del 20 % de sujetos restante (variables 126,074 a 179,755) [9] se guardaron en “mat02”.

Posteriormente, se aleatorizaron las bases de datos y se llevaron a cabo los puntos 2 y 3 de la metodología descrita. No obstante, el modelo de red propuesto para esta exploración se caracteriza por tener 15 neuronas en la primera capa oculta y 11 neuronas en la segunda. En la Tabla 1 se pueden apreciar los parámetros establecidos para los modelos descritos en este trabajo.

El modelo BD65 pertenece al diseño reportado en trabajos anteriores [13] donde no se aplicó PCA a la base de datos. Por otro lado, los modelos BD PCA y BD PCA Sujetos mostrados en la Tabla 1, se ajustan al criterio establecido del 95 % de significancia, por lo que se utilizan 22 y 20 componentes respectivamente obtenidos por el método PCA. Se destaca que las diferencias en los modelos son los parámetros de ajuste de las redes, la cantidad de neuronas y principalmente la distribución de los datos en el entrenamiento del Perceptrón Multicapa.

TABLA 1. **Parámetros de modelos diseñados**

Modelo	Norma lización	Capas	Función Activación	Neuronas
BD65	0 a1	4	ReLU sigmoid sigmoid sigmoid	61 52 40 27
BD PCA 22	-1 a 1	2	ReLU sigmoid	15 11
BD PCA Sujetos 20	-1 a 1	2	ReLU sigmoid	17 11

RESULTADOS Y DISCUSIÓN

El comportamiento de generalización para el modelo de red neuronal artificial es la habilidad que el sistema tiene de reconocer de manera correcta las AVDs que no se le presentaron con anterioridad en la etapa de entrenamiento. Dicho comportamiento para el modelo elaborado en el trabajo de investigación previo se puede apreciar en la Figura 2.

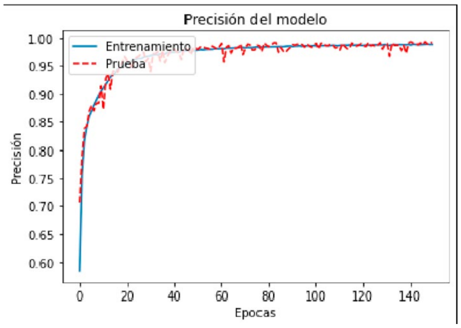


FIGURA 2. **Comportamiento *BD 65.***

Por medio de esta gráfica se comprueba que las redes neuronales artificiales y el aprendizaje profundo permiten la clasificación de actividades físicas capturadas con sensores inerciales y almacenadas en la base de datos con un resultado de exactitud elevado tanto en la fase de entrenamiento como de prueba [13]. Para una mejor visualización del modelo, se llevó a cabo la representación arquitectónica con los detalles de las neuronas y funciones de activación para cada capa. Dicha representación se puede observar en la Figura 3.

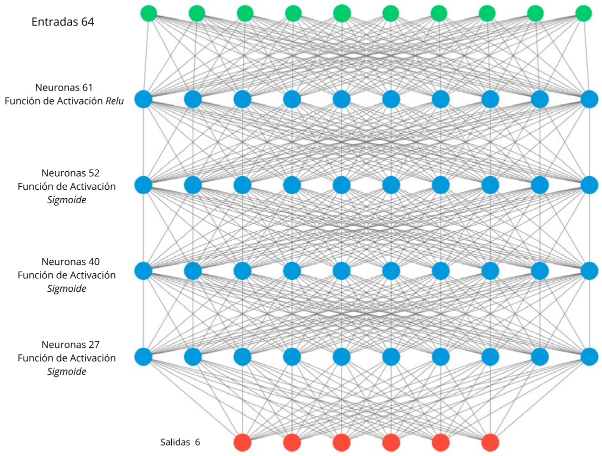


FIGURA 3. Arquitectura de Red Neuronal Artificial *BD65.*

En la Tabla 2, se puede observar la distribución de parámetros que se entrenan al interior de la red de primer modelo BD65 en KB.

TABLA 2. Resumen del modelo *BD65*

Capa (tipo)	Forma de Salida	Parámetros
Densa (Densa)	(Ninguno, 61)	3065
Densa_1 (Densa)	(Ninguno, 52)	3224
Densa_2 (Densa)	(Ninguno, 40)	2120
Densa_3 (Densa)	(Ninguno, 27)	1107
Densa_4 (Densa)	(Ninguno, 6)	168
Total de parámetros: 10584 (41.34 KB) Parámetros entrenables: 10584 (41.34 KB) Parámetros no entrenables: 0 (0 KB)		

Para el segundo modelo el Perceptrón multicapa BD PCA que fue obtenido mediante el análisis de Componentes Principales que permite ordenar la varianza de los datos a partir de aquellas características más relevantes hasta las dimensiones menores que se consideran como ruido aleatorio [28], se alcanzó niveles muy elevados de precisión tanto en el entrenamiento como en prueba, como se puede observar, en la Figura 4.

En esta propuesta, la Red Neuronal Artificial obtuvo un porcentaje de aprendizaje del 95 % en la prueba y entrenamiento. Para lograr este porcentaje de aprendizaje, se tuvo que entrenar durante más épocas y así garantizar que el método de reducción de características funciona de manera similar al comportamiento de la Figura 2, perteneciente al primer modelo.

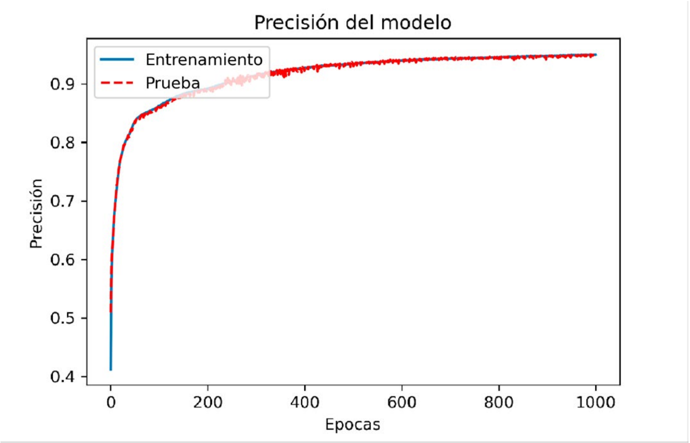


FIGURA 4. Comportamiento de aprendizaje *BD PCA.*

En este segundo modelo de red donde se utilizó una base de datos con 22 componentes y se obtuvo la arquitectura mostrada en la Figura 5.

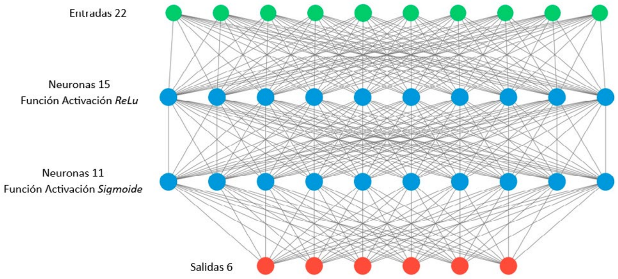


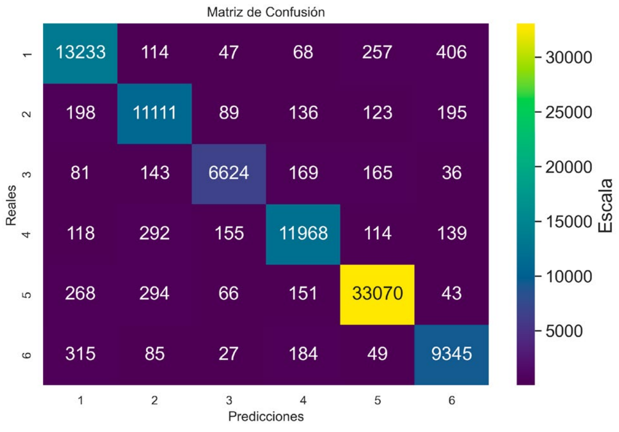
FIGURA 5. Arquitectura MLP *BD PCA.*

En la Tabla 3, se observa el resumen de este segundo modelo y cómo es que el comportamiento y uso de KB cambia con respecto al primer modelo.

TABLA 3. **Arquitectura MLP *BD PCA.***

Capa (tipo)	Forma de Salida	Parámetros
Densa (Densa)	(Ninguno, 15)	345
Densa_1 (Densa)	(Ninguno, 11)	176
Densa_2 (Densa)	(Ninguno, 6)	72
Total de parámetros: 593 (2.32 KB) Parámetros entrenables: 593 (2.32 KB) Parámetros no entrenables: 0 (0 KB)		

En cuanto a la clasificación real de las actividades respecto a sus predicciones, la Figura 6, muestra la matriz de confusión de esta red, que indica la proporción de datos clasificados correctamente con la cantidad total de ellos, es decir, permite ver la distribución de los errores identificando a detalle contra cual actividad se equivoca de las 6 evaluadas, así como la precisión en la clasificación de estas.



	1	2	3	4	5	6
1	13233	114	47	68	257	406
2	198	11111	89	136	123	195
3	81	143	6624	169	165	36
4	118	292	155	11968	114	139
5	268	294	66	151	33070	43
6	315	85	27	184	49	9345

FIGURA 6. **Matriz de confusión Modelo BD PCA.**

Para el tercer modelo, se utilizan las observaciones del 80 % de sujetos en la etapa de entrenamiento, sin presentar las señales restantes hasta la fase de prueba. También se mantuvo una significancia del 95 % en el PCA, de esta manera se requirieron 20 componentes. Lo anterior, sugiere que el quinto sujeto podría aportar información relevante para el aprendizaje del sistema. Este comportamiento se puede observar en la gráfica de precisión mostrada en la Figura 7.

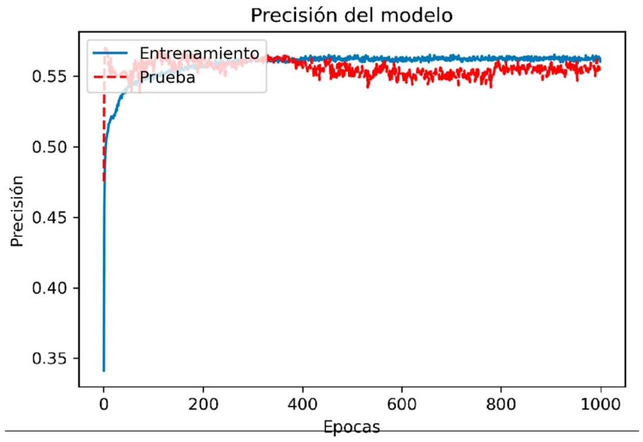


FIGURA 7. **Precisión de modelo *BD PCA Sujetos.***

La Figura 7 es una representación gráfica del porcentaje de exactitud que se obtuvo al entrenar únicamente con las señales de 80 % de sujetos y probar con las señales del porcentaje restante. La precisión de la red en la etapa de prueba únicamente llegó a ser del 51.09 %. Comprobando que la generalización de la red no mostró resultados favorables. La arquitectura de dicho modelo se puede observar en la Figura 8.

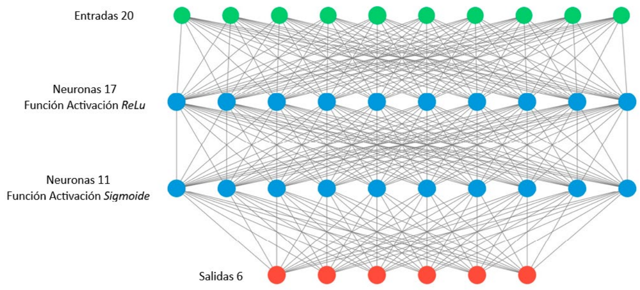


FIGURA 8. **Precisión de modelo *BD PCA Sujetos.***

Finalmente, para el tercer modelo se presenta en la Tabla 4, la reducción en el uso de KB de los parámetros entrenables de la red. La disminución progresiva del número de parámetros en cada modelo, nos indica que el número de operaciones para obtener una inferencia (salida de la red) se realiza con menos tiempo, dando posibilidad de implementarse en dispositivos con recursos más limitados y por tanto económicos.

TABLA 4. **Resumen parámetros *BD PCA Sujetos***

Capa (tipo)	Forma de Salida	Parámetros
Densa (Densa)	(Ninguno, 15)	345
Densa_1 (Densa)	(Ninguno, 11)	176
Densa_2 (Densa)	(Ninguno, 6)	72
Total de parámetros: 593 (2.32 KB) Parámetros entrenables: 593 (2.32 KB) Parámetros no entrenables: 0 (0 KB)		

En cuanto a la matriz de confusión para este último modelo donde se utilizaron 20 % de las señales pertenecientes a un solo sujeto en la fase de prueba, se puede notar en la Figura 9; que la cantidad de error en las predicciones de la clasificación para las 6 clases es elevada. Lo que lleva a comprobar que la información del último sujeto sí es relevante para el adecuado entrenamiento de la red.

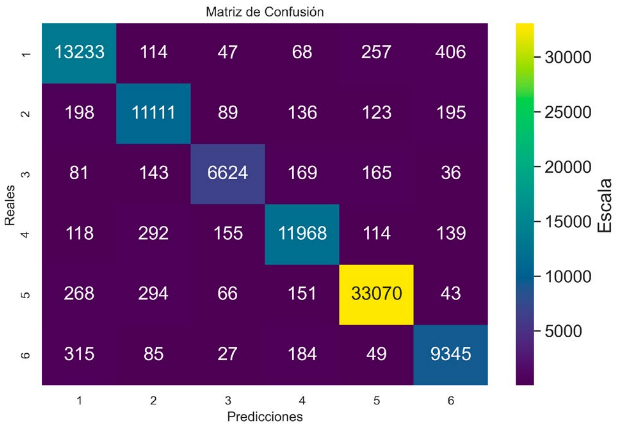


FIGURA 9. **Matriz de confusión para modelo *RNA PCA Sujetos***

Múltiples parámetros permiten evaluar el rendimiento de los modelos. Por ejemplo, la exactitud es la métrica más utilizada para evaluar la exactitud general de las predicciones del clasificador [29][30]. La precisión mide la exactitud de las predicciones positivas [29][30]. La sensibilidad o recall mide la habilidad para identificar la proporción de casos positivos mientras que la especificidad hace lo propio para los casos negativos [29][30]. La medida F1-score es la combinación de la exactitud con la sensibilidad. La medida F1-score es la combinación de la exactitud con la sensibilidad [29][30], por esta razón y debido a que la base está desbalanceada se consideró a F1-score como la métrica apropiada en la propuesta. Para facilitar la comprensión de la calidad de las predicciones de los modelos de RNA PCA 22 y PCA sujetos los parámetros anteriores fueron calculados en Python, ver Tabla 5. Todos los parámetros muestran la superioridad del modelo RNA BD PCA 22 y que el PCA funcionó muy bien ya que todos los parámetros son altos, incluso la medida f1-score.

TABLA 5. **Parámetros de modelos de RNA**

CLASE	BD PCA 22 Exactitud 0.95			BD PCA Sujetos 20 Exactitud 0.55		
	Precisión	Recall	F1 Score	Precisión	Recall	F1 Score
1	0.93	0.94	0.93	0.71	0.51	0.59
2	0.92	0.94	0.93	0.27	0.12	0.16
3	0.95	0.92	0.93	0.27	0.12	0.17
4	0.94	0.94	0.94	0.60	0.61	0.60
5	0.98	0.98	0.98	0.57	0.92	0.70
6	0.92	0.93	0.93	0.2	0.07	0.10

Cabe destacar que no es definitiva la conclusión respecto a la aportación del último sujeto debido a que es necesario realizar un estudio más detallado que involucre una mayor cantidad de participantes en el protocolo y analizar la distribución de la colocación de los sensores e identificar la mejor frecuencia de muestreo.

Al comparar el comportamiento de las gráficas de precisión se destaca que quitarle las señales de un sujeto de prueba a la red en la fase de entrenamiento, perjudica su aprendizaje. Esto puede deberse a diversos factores. Son pocos los sujetos de investigación que participaron en este proyecto por lo que la red no cuenta con diversas señales. Igualmente puede que la colocación de los sensores haya influenciado en la toma de señales.

Como resultado de comparar los porcentajes de precisión de las redes neuronales artificiales evaluadas tanto en trabajos previos como los reportados en este escrito, se obtuvo la Figura 10.

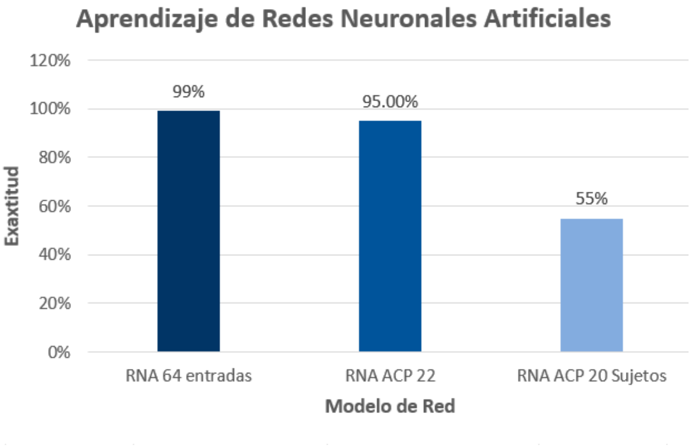


FIGURA 10. **Exactitud en clasificación de AVDs en RNA.**

La reducción de la dimensionalidad a través de PCA permite mejorar la eficacia del sistema manteniendo la información de los componentes más relevantes y con esto descartar aquellas variables que no proporcionan información relevante. Logrando mantener porcentajes elevados de precisión.

La clasificación de actividades realizadas por pacientes con enfermedad de Parkinson ha sido estudiada mediante otras técnicas y algoritmos con tasas de exactitud consideradas aceptables a través de diferentes lenguajes de programación como en Matlab. El aprendizaje profundo ha irrumpido con fuerza en este problema y son cada vez más las redes neuronales artificiales diseñadas para la resolución de distintos tipos de clasificación, permitiendo obtener buenos resultados de manera sencilla. El uso de Análisis de Componentes Principales permite mantener altos porcentajes de exactitud sin necesidad de utilizar toda la información de la base de datos; por lo que se considera una herramienta importante para trabajos futuros; como lo es la implementación de algún diseño de aplicación móvil que permita monitorear las señales del paciente e incluso queda la posibilidad de estudiar las señales a otro rango de frecuencia adquirido con la finalidad de ver si esto proporciona un cambio en el registro de los movimientos para cada actividad.

CONCLUSIONES

En las últimas décadas, la aplicación del *Deep Learning* ha ido creciendo debido a las características que ofrece trabajar con diversos modelos de Redes Neuronales Artificiales. El presente estudio demuestra la versatilidad de las RNA basadas en MLP combinadas con la técnica de PCA, pues incluso en una base de datos desbalanceada como la utilizada permite alcanzar medidas F1-score mayores al 93 %. Emplear una base de datos reducida utilizando aproximadamente el 34 % de la información original, permite tener resultados elevados en la exactitud utilizando herramientas de PCA. El estudio desarrollado ha demostrado el potencial de implementar una de las amplias ramas que la Inteligencia Artificial brinda para la resolución de los problemas en cuanto a la clasificación; para posteriormente ser integrado a un sistema que apoye al personal de medicina en el monitoreo de actividades realizadas por el paciente dando seguimiento de la enfermedad de Párkinson.

AGRADECIMIENTOS

Los autores quieren agradecer al CONACyT por su patrocinio para la creación de la base de datos original.

DECLARACIÓN ÉTICA

Por este medio se declara que ninguno de los autores participantes en este trabajo cuenta con algún conflicto de interés. Así mismo en este estudio no se realizó investigación en humanos directamente pues se utilizó una base de datos, conformada de un estudio previo con aprobación de los comités de ética e investigación institucionales.

CONTRIBUCIONES DE LOS AUTORES

L.R.M. participó en la curación de datos, realizó investigación para el proyecto, desarrolló la metodología e implementó software especializado y contribuyó en la preparación y visualización de resultados. J.A.B. Conceptualizó el proyecto, participó en la curación de datos, desarrolló la metodología e implementó software especializado, contribuyó en la validación e implementación de resultados, supervisó y obtuvo recursos para el desarrollo del proyecto. A.I.P.S. Conceptualizó el proyecto, participó en la curación de datos, desarrolló los análisis, realizó investigación para el proyecto, supervisó y obtuvo recursos para el desarrollo del proyecto. Todos los autores participaron en todas las etapas del desarrollo del manuscrito, y aprobaron la versión final del mismo.

REFERENCIAS

[1] A.-E. Gómez Ayala, “Enfermedad de Parkinson,” Offarm, vol. 26, no. 5, pp. 70-78, 2007. [En línea]. Recuperado de: https://www.elsevier.es/es-revista-offarm-4-articulo-enfermedad-parkinson-13102417

[2] R. Martínez-Fernández, C. Gasca-Salas, A. Sánchez-Ferro, J. A. Obeso, “Actualización en la Enfermedad de Parkinson,” Rev. Med. Clin. Las Condes, vol. 27, no. 3, pp. 363-379, 2016.

[3] J. Quinzaños-Fresnedo y A. I. Pérez San Pablo, “Rehabilitación del Paciente Con Enfermedad de Parkinson,” Bol. Méd. Inform. Inst. Nac. Rehab., no. 79, pp. 2-25, 2021. [En línea]. Disponible en: https://www.inr.gob.mx/Descargas/boletin/079Boletin.pdf

[4] L. Borzì, M. Varrecchia, G. Olmo, C. A. Artusi, et al., “Home monitoring of motor fluctuations in Parkinson’s disease patients,” J. Reliable Intell. Environ., vol. 5, no. 3, pp. 145-162, 2019, doi: https://doi.org/10.1007/s40860-019-00086-x

[5] M. Barrachina-Fernández, A. M. Maitín, C. Sánchez-Ávila, J. P. Romero, “Wearable Technology to Detect Motor Fluctuations in Parkinson’s Disease Patients: Current State and Challenges,” Sensors, vol. 21, no. 12, art. no. 4188, 2021, doi: https://doi.org/10.3390/s21124188

[6] M. A. Hobert, W. Maetzler, K. Aminian, L. Chiari, “Technical and clinical view on ambulatory assessment in Parkinson's disease,” Acta Neurol. Scand., vol. 130, no. 3, pp. 139-147, 2014, doi: https://doi.org/10.1111/ane.12248

[7] F. M. Rast y R. Labruyère, “Systematic review on the application of wearable inertial sensors to quantify everyday life motor activity in people with mobility impairments,” J. NeuroEngineering Rehabil., vol. 17, no. 1, art. no. 148, 2020, doi: https://doi.org/10.1186/s12984-020-00779-y

[8] S. Jung, M. Michaud, L. Oudre, E. Dorveaux, L. Gorintin, N. Vayatis, D. Ricard, “The Use of Inertial Measurement Units for the Study of Free Living Environment Activity Assessment: A Literature Review,” Sensors, vol. 20, no. 19, art. no. 5625, 2020, doi: https://doi.org/10.3390/s20195625

[9] A. I. Pérez Sanpablo, A. Meneses Peñaloza, J. Quinzaños Fresnedo, V. Bueyes Roiz, I. Quiñones Uriostegui, C. Hernandez Arenas, “Accuracy of discriminant analysis methods to classify activities of subjects with Parkinson’s Disease using wearable sensors,” Arch. Phys. Med. Rehab., vol. 102, no. 10, art. no. e103, 2021, doi: https://doi.org/10.1016/j.apmr.2021.07.797

[10] H. Nguyen, K. Lebel, S. Bogard, E. Goubault, P. Boissy, C. Duval, “Using Inertial Sensors to Automatically Detect and Segment Activities of Daily Living in People With Parkinson’s Disease,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 1, pp. 197-204, 2018, doi: https://doi.org/10.1109/TNSRE.2017.2745418

[11] A. Rana, A. Dumka, R. Singh, M. K. Panda, N. Priyadarshi, B. Twala, “Imperative Role of Machine Learning Algorithm for Detection of Parkinson’s Disease: Review, Challenges and Recommendations,” Diagnostics, vol. 12, no. 8, art. no. 2003, 2022, doi: https://doi.org/10.3390/diagnostics12082003

[12] M. Awais, L. Chiari, E. A. F. Ihlen, J. L. Helbostad, L. Palmerini, “Classical Machine Learning Versus Deep Learning for the Older Adults Free-Living Activity Classification,” Sensors, vol. 21, no. 14, art. no. 4669, 2021, doi: https://doi.org/10.3390/s21144669

[13] L. R. Montero, J. A. Bastian, A. I. P. SanPablo, “Classification of Activities of Daily Living in Subjects with Parkinson’s Disease using Artificial Neural Networks,” en 2023 Global Medical Engineering Physics Exchanges/Pacific Health Care Engineering (GMEPE/PAHCE), Songdo, Korea, Republic of, 2023, pp. 1-5, doi: https://doi.org/10.1109/GMEPE/PAHCE58559.2023.10226479

[14] N. González García y A. Taborda Londoño, “Análisis de Componentes Principales Sparse, Formulación, algoritmos e implicaciones en análisis de datos,” Tesis de maestría, Univ. Sal., Salamanca, España, 2015. [En línea]. Disponible en: https://gredos.usal.es/bitstream/handle/10366/126046/TFM_MAADM_Gonz%C3%A1lez_Taborda.pdf?sequence=4

[15] I. T. Jollife y J. Cadima, “Principal component analysis: A review and recent developments,” Philos. Trans. A Math. Phys. Eng. Sci., vol. 374, art. no. 2065, 2016, doi: https://doi.org/10.1098/rsta.2015.0202

[16] U. Dillmann, C. Holzhoffer, Y. Johann, S. Bechtel, et al., “Principal Component Analysis of gait in Parkinson’s disease: relevance of gait velocity,” Gait Posture, vol. 39, no. 3, pp. 882-887, 2014, doi: https://doi.org/10.1016/j.gaitpost.2013.11.021

[17] T. Varrecchia, S. F. Castiglia, A. Ranavolo, C. Conte, et al., “An artificial neural network approach to detect presence and severity of Parkinson’s disease via gait parameters,” PLoS ONE, vol. 16, no. 2, 2021, art. no. e0244396, 2021, doi: https://doi.org/10.1371/journal.pone.0244396

[18] A. Rana, A. S. Rawat, A. Bijalwan, H. Bahuguna, “Application of Multi Layer (Perceptron) Artificial Neural Network in the Diagnosis System: A Systematic Review,” 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), San Salvador, El Salvador, 2018, pp. 1-6, doi: https://doi.org/10.1109/RICE.2018.8509069

[19] A. D. Pano-Azucena, E. Tlelo-Cuautle, S. X.-D. Tan, B. Ovilla-Martinez, L. G. De la Fraga, “FPGA-Based Implementation of a Multilayer Perceptron Suitable for Chaotic Time Series Prediction,” Technologies, vol. 6, no. 4, art. no. 90, 2018, doi: https://doi.org/10.3390/technologies6040090

[20] J. Zou, Y. Han, S. S. So, "Overview of artificial neural networks," Methods Mol. Biol., vol. 458, pp. 15-23, 2008, doi: https://doi.org/10.1007/978-1-60327-101-1_2

[21] G. AlMahadin, A. Lotfi, M. M. Carthy, P. Breedon, "Enhanced Parkinson's Disease Tremor Severity Classification by Combining Signal Processing with Resampling Techniques," SN Comp. Sci., vol. 3, art. no. 63, 2022, doi: https://doi.org/10.1007/s42979-021-00953-6

[22] L. Sigcha, B. Domínguez, L. Borzì, N. Costa, et al., "Bradykinesia Detection in Parkinson's Disease Using Smartwatches' Inertial Sensors and Deep Learning Methods," Electronics, vol. 11, no. 23, art. no. 3879, 2022, doi: https://doi.org/10.3390/electronics11233879

[23] R. LeMoyne, T. Mastroianni, D. Whiting, N. Tomycz, "Preliminary Network Centric Therapy for Machine Learning Classification of Deep Brain Stimulation Status for the Treatment of Parkinson's Disease with a Conformal Wearable and Wireless Inertial Sensor," Adv. Parkinson's Dis., vol. 8, no. 4, pp. 75-91, 2019, doi: https://doi.org/10.4236/apd.2019.84007

[24] A. Fred Agarap, "Deep Learning using Rectified Linear Units (ReLU)," 2018, arXiv: 1803.08375, doi: https://doi.org/10.48550/arXiv.1803.08375

[25] J. Lederer, "Activation Functions in Artificial Neural Networks: A Systematic Overview," 2021, arXiv: 2101.09957, doi: https://doi.org/10.48550/arXiv.2101.09957

[26] J. P. Larsen, E. Dupont, E. Tandberg, "Clinical diagnosis of Parkinson's disease. Proposal of diagnostic subgroups classified at different levels of confidence," Acta Neurol. Scand., vol. 89, no. 4, pp. 242-251, 2009, doi: https://doi.org/10.1111/j.1600-0404.1994.tb01674.x

[27] C. Sweeney, E. Ennis, M. Mulvenna, R. Bond, S. O'Neill, "How Machine Learning Classification Accuracy Changes in a Happiness Dataset with Different Demographic Groups," Computers, vol. 11, no. 5, art. no. 83, 2022, doi: https://doi.org/10.3390/computers11050083

[28] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, E. Tuzhilina, "Principal component analysis," Nat. Rev. Methods Primers, vol. 2, art. no. 100, 2022, doi: https://doi.org/10.1038/s43586-022-00184-w

[29] N. Japkowicz, "Assessment Metrics for Imbalanced Learning," en Imbalanced Learning: Foundations, Algorithms, and Applications, H. He y Y. Ma, Eds., Hoboken, New Jersey, Estados Unidos: Wiley, 2013, cap. 8, pp. 187-206, doi: https://doi.org/10.1002/9781118646106.ch8

[30] M. Bekkar, H. K. Djemaa, T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," J. Inf. Eng. Appl., vol. 3, no. 10, pp. 27-38, 2013, [Online]. Disponible en: https://www.iiste.org/Journals/index.php/JIEA/article/view/7633/8051

RESEARCH ARTICLE

VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 140 - 151

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.10

E-LOCATION ID: **1395**

A U-Net with Statistical Shape Restrictions Applied to the Segmentation of the Left Ventricle in Echocardiographic Images

U-Net con Restricciones Estadísticas de Forma, Aplicada a la Segmentación del Ventrículo Izquierdo en Imágenes de Ecocardiograma

Eduardo Galicia-Gómez[1], *Fabián Torres-Robles*[2], *Jorge Pérez-González*[3], *Boris Escalante-Ramirez*[4]
Fernando Arámbula Cosío[3] ✉

[1] Universidad Nacional Autónoma de México, Posgrado en Ciencia e Ingeniería de la Computación, Ciudad de México - México

[2] Universidad Nacional Autónoma de México, Laboratorio de Física Medica, Instituto de Física, Ciudad de México - México

[3] Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Yucatán - México

[4] Universidad Nacional Autónoma de México, Facultad de ingeniería, Ciudad de México - México

ABSTRACT

This paper aims to introduce an innovative approach to semantic segmentation by leveraging a convolutional neural network (CNN) for predicting the shape and pose parameters of the left ventricle (LV). Our approach involves a modified U-Net architecture with a regression layer as the final stage, as opposed to the traditional classification layer. This modification allows us to predict all the shape and pose parameters of a statistical shape model, including rotation, translation, scale, and deformation. The adapted U-Net is trained using data from a point distribution model (PDM) of the LV. The experimental results demonstrate a mean Dice coefficient of 0.82 on good quality images, and 0.66 including mean and low-quality images. Our approach successfully overcomes a common issue encountered in CNN-based semantic segmentation. Unlike the inaccurate pixel classification that often leads to unwanted blobs, our CNN generates statistically valid shapes. These shapes hold significant potential in initializing other methods, such as active shape models (ASMs). Our novel CNN-based approach provides a novel solution for semantic segmentation, offering shapes and pose parameters that can enhance the accuracy and reliability of subsequent medical image analysis methods.

KEYWORDS: convolutional neural networks, echocardiography, left ventricle segmentation, statistical shape analysis

RESUMEN

Este artículo tiene como objetivo introducir un enfoque innovador para la segmentación semántica utilizando una red neuronal convolucional (CNN) para predecir los parámetros de forma y posición del ventrículo izquierdo (VI). Nuestro enfoque implica una arquitectura U-Net modificada con una capa de regresión como etapa final, en contraposición a la capa de clasificación tradicional. Esta modificación nos permite predecir todos los parámetros de un modelo estadístico de formas que incluyen rotación, traslación, escala y deformación. La red convolucional se entrena utilizando datos de un modelo de distribución de puntos (PDM) del VI. Los resultados experimentales muestran un coeficiente Dice promedio de 0.82 para imágenes de buena calidad y de 0.66 cuando se incluyen imágenes de calidad media y baja. Nuestro enfoque supera con éxito un problema común en la segmentación semántica basada en CNNs. A diferencia de la clasificación inexacta de píxeles que a menudo conduce a elementos no deseados (blobs), nuestra CNN genera formas estadísticamente válidas. Estas formas tienen un gran potencial para inicializar otros métodos, como los modelos de forma activa (ASMs). En resumen, nuestro enfoque basado en CNN proporciona una solución innovadora para la segmentación semántica, ofreciendo formas y parámetros de posición que pueden mejorar la precisión y confiabilidad de otros métodos de análisis del VI.

PALABRAS CLAVE: análisis estadístico de forma, ecocardiografía, redes neuronales convolucionales, segmentación del ventrículo izquierdo

Corresponding author

TO: **Fernando Arámbula Cosío**
INSTITUTION: **Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Unidad Yucatán, Universidad Nacional Autónoma de México**
ADRESS: **Carretara Mérida-Tetiz Km. 4.5, Ucú, Yucatán, México C.P. 97357**
EMAIL: **fernando.arambula@iimas.unam.mx**

Received:

31 Oct 2023

Accepted:

10 Dec 2023

INTRODUCTION

Left ventricle (LV) segmentation in ultrasound images is a crucial process for evaluating the ejection fraction and assessing heart condition. Ultrasound offers real-time imaging with moderate costs and no ionizing radiation. However, image quality is affected by speckle noise and acoustic shadows of adjacent structures. For these reasons, several automatic and non-automatic methods have been developed to perform the segmentation of the LV in ultrasound images.

Some of the main approaches to LV segmentation have been active contours, such as geodesic models [1] and level sets [2], deformable templates, such as shape models [3] registration-based methods [4], and supervised learning models such as data base guided segmentation [5], and hybrid active appearance models [6]. International challenges have contributed to the publication of large image databases that enable the training and testing of supervised learning models [7]. Machine learning models have shown high performance in the segmentation of the LV in echocardiography in 2D and 3D, taking advantage of the public image databases. An example of a machine learning method for LV segmentation is reported in [8]. A structured random forest was developed for automatic segmentation of the myocardium and the LV on an echocardiographic data set of 250 patients, including healthy athletes and cardiac patients. The random forest showed improved segmentation results when compared against an active appearance model.

Statistical shape models (SSMs) and active shape models (ASMs) [9] have been extensively used, with continuous improvements. An enhanced ASM, as reported in [10], incorporates an adaptive strategy to construct appearance models of each landmark point, optimizing the number of principal components (PCs) using mean squared eigenvalue error (MSEE). Preprocessing with a Nakagami filter further improved segmentation results on good quality images from the CAMUS dataset. However, details on the ASM's initialization method were not provided in the study.

Recently deep learning methods have been applied with good results for the segmentation of the LV. In [11] is presented a review of the application of convolutional neural networks (CNNs) to the segmentation of the LV in ultrasound and MRI. Fully convolutional [12] and the U-Net architecture [13][14][15] have been successfully applied to the classification of the pixels corresponding to the LV; Different SSMs such as morphological models, snake models, and ASMs have been combined with CNNs to create hybrid architectures in different studies. In [16] is reported a hybrid method with a fast region-based CNN and an ASM for LV segmentation in ultrasound images where adaptive anisotropic diffusion filtering is applied to all images, the fast CNN detects the bounding box around the LV for ASM initialization and the ASM finds the final boundary of the LV. On the other hand, new CNN architectures have been developed for the adjustment of a shape contour. In [17] is reported a new approach for left ventricle segmentation based on a CNN that first, detects 3 landmarks: apex, starting and, end points of the endocardium; Afterwards, a triangle (start-apex-end point) is used as initialization for a deep-snake [18] which is adjusted to the endocardium using circular convolution, good results were obtained in the HMC-QU echocardiography data set. Also, convolutional neural networks based on autoencoders such as U-Net [19], have shown good results when performing segmentation with reduced training sets; however, this type of network has a great disadvantage as they are based on semantic segmentation, they produce areas (blobs) with an inaccurate classification that are sometimes far from the organ to be segmented (see Figure 1).

In this work, is reported a modified U-Net based architecture, which incorporates expert shape knowledge of the LV in a point distribution model (PDM) [9], this approach diverges from the conventional pixel by pixel classification method discussed in [19]. Our model focuses on generating statistically valid shapes for the left ventricle (LV). This alternative segmentation strategy avoids the production of blobs due to pixel classification errors. Our work contributes to improve the reliability of LV segmentation. The training error of the proposed CNN was calculated as the RSME between the expert pose and shape values and the CNN output at each epoch. During the test stage our U-Net model optimizes the shape and pose parameters corresponding to a non-training ultrasound image of the LV. In the following sections are presented all the details of the PDM of the LV, the U-Net architecture developed and its training parameters. In section III are reported the tests and results on the CAMUS [20] and EchoNet-Dynamic [21] databases. In section IV are presented the discussion and conclusions.

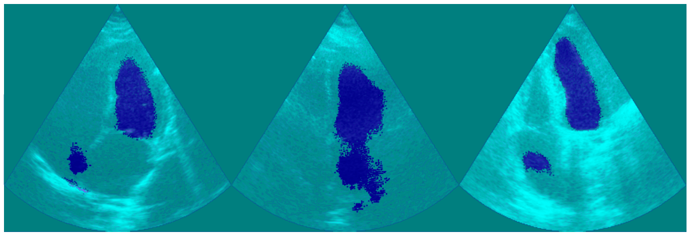


FIGURE 1. **Example of inaccurate classification by the U-Net generating blobs that are not part of the left ventricle.**

MATERIALS AND METHODS

This section presents the proposed methodology. As can be observed in Figure 2, the methodology starts with the training of a U-Net convolutional network, which is fed with a set of echocardiography images (Figure 2a-2b). The target is to estimate the parameters of translation, rotation, scale, as well as a deformation vector (Figure 2c), which are useful to adjust the mean shape obtained from the PDM, and thus perform the segmentation of the left ventricle in the systole and diastole phases (Figure 2d). Finally, the proposed methodology is validated using the Dice coefficient and the Haussdorff distance. The details of each stage are explained in the following subsections.

Point distribution models

Point distribution models (PDMs) have been widely used for modeling complex structures, such as the organs of the human body [9][22] and specifically in this paper the left ventricle. The model is based on a set of points known as *"landmarks"*, which must be a fixed number and should correspond to the same position along the contour of each example in the training set, as shown in Figure 3.

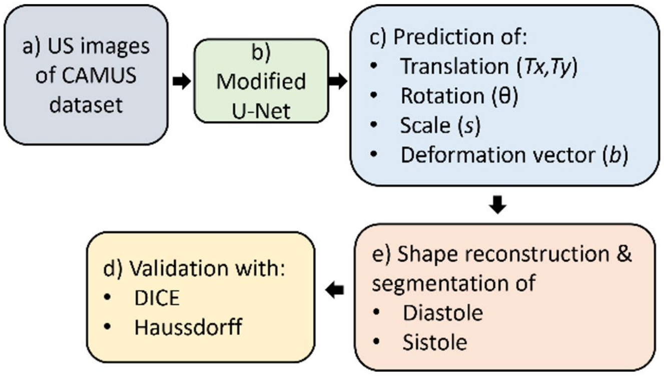


FIGURE 2. **Workflow of the proposed method.**

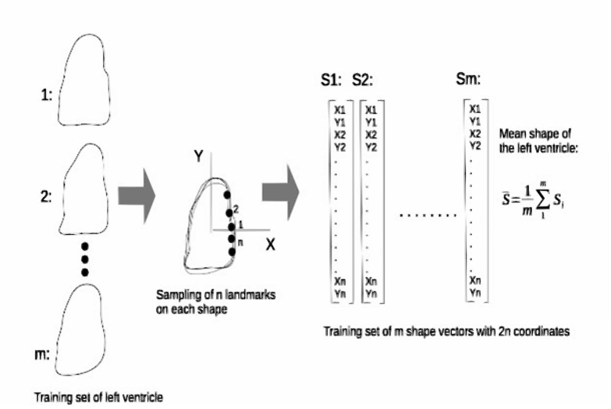


FIGURE 3. **Landmark sampling of each left ventricle shape in the training set.**

A shape vector is constructed by concatenation of all the $[x,y]$ coordinates of the landmarks of each training shape. Subsequently, with this set of landmarks, the mean shape $(\bar{x})$ is calculated as proposed by Cootes *et al.*, [9], and the training set is aligned following the Procrustes algorithm [23]. Finally, a principal component analysis (PCA) is performed on the covariance matrix of this set, obtaining the shape parameters (P,b) according to the 90 % of the explained variance, thus enabling the generation of new instances of the left ventricle, as described in Equation (1).

$$\hat{x} = \overline{x} + Pb \qquad (1)$$

Where:

$\hat{x} = new\ shape.$

$\overline{x} = mean\ shape.$

$P = principal\ eigenvector\ matrix.$

b = *deformation parameter vector.*

Shape and pose parameters.

In this work we constructed a PDM of the left ventricle with the purpose of modeling the deformation of each shape in the training set. The deformation of the PDM for each left ventricle shape can be calculated by solving for b in Equation 1, which yields Equation 2, where *b* contains the *shape parameters* of each training shape. The *pose parameters* (rotation, translation in X, translation in Y and scale) were obtained as follows: for each of the contours in the training set the rotation and scale were computed as suggested in [9] while translation in X and Y were calculated as the average value in each axis, on each training example.

$$b = (\hat{x} - \overline{x}) * P' \qquad (2)$$

U-Net with statistical shape restrictions

We adapted a convolutional neural network architecture that uses the features extracted by the convolutional filters to predict the values of *shape and pose* of the PDM of left ventricle in this way, all segmentations are *valid shapes* restricted by the training of the network, thus avoiding the blobs shown in Figure 1. The ultimate goal of this CNN is to predict the values of *rotation, translation,* scale and the deformation vector b corresponding to a given ultrasound image of the left ventricle.

Network architecture

To provide a better understanding of the proposed U-Net architecture (see Figure 4), it will be divided into 4 blocks:

1. **Input block:** Is an image input layer of size 256x256x1 of the form: (width, height, channels).

2. **Encoder - Decoder block:** The encoder block is formed by convolutional filters of size 3x3 and its goal is to extract relevant features of the image, also the encoder path reduces the spatial resolution of the extracted feature maps applying maxpooling operations. On the other hand, the decoder path upsamples the features maps and preserves the spatial resolution of the input while also performs convolutional operations. By employing the skip connections from the encoder, the decoder layers enhance their ability to detect and fine-tune the features within the image as explained in [19].

3. **Fully connected block:** It consists of a flatten layer followed by a set of fully connected layers. The purpose of this block is to link the features extracted by the *encoder-decoder* block with the shape and pose parameters related to the input image.

4. **Output block:** Finally at the end of the fully connected block is a regression layer that allows to predict the desired parameter: rotation, translation, scale, or *b*-vector.

Network Training

Using the pose and shape parameters described above, an input vector was constructed for each training image I_i as follows in Equation (3):

$$V_i = [I_i\ \theta_i, x_i, y_i, s_i, b_i] \qquad (3)$$

Where:

$I_i = the\ i - th\ image\ of\ the\ training\ set.$

θ_i= the i-th rotation value.

x_i = the i-th x translation value.

y_i= the i-th y translation value.

s_i= the i-th scale value.

b_i= the i-th deformation vector.

The vector V_i then becomes the input of the proposed CNN, the intention of training a U-Net architecture is to take full advantage of its ability to extract features throughout the encoder stage, the idea is then to use these features to predict the values of pose and shape of the left ventricle, assuming that each LV image corresponds to a single contour and these are never the same between patients. Also the network *loss* during the training stage is calculated as the RMSE between the values of shape and pose predicted by the network (as output from the regression layer) and the *ground truth* contained in each training vector V_i through an iterative process called: stochastic gradient descent (SGD) the network adjusts its weights according to the values of the loss trying to decrease it in each iteration, the lower the loss the better the prediction of the pose and shape values.

Predicted shape reconstruction

Once the network has been trained, it is possible to predict the values of θ,Tx,Ty,s and b following the diagram in Figure 2. These values allow reconstructing the LV contour corresponding to the input image, first by applying b and P to the mean shape $(\bar{x})$ to obtain the deformation of the left ventricle (see Equation 2) and then by multiplying the resulting shape $(\hat{x})$ by the transformation matrix in Equation 4, thus obtaining the reconstructed LV shape *(S)*. Finally, S is placed in the input image to obtain the corresponding LV segmentation.

$$S = \hat{x}s \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} + [Tx, Ty] \quad (4)$$

Where:

$\theta = the\ rotation\ predicted\ by\ theCNN.$

$s = the\ scale\ predicted\ by\ theCNN.$

$[Tx, Ty] = two - column\ translation$

$values.$

Validation

To validate the proposed method, we used two measures:

1. **Dice Coefficient:** This method allows us to determine the area of overlap between the segmentation mask of the expert and the segmentation mask of our method. A higher overlap area corresponds to a higher Dice coefficient value. Therefore, the larger the coefficient value, the better the segmentation. This coefficient is within the range of 0 to 1.

2. **Haussdorff Distance:** It measures the distance between two sets of points. In this case, it measures the distance between the contour marked by the expert and the contour reconstructed with the network values. A smaller Haussdorff distance indicates a greater similarity between the two contours.

RESULTS AND DISCUSSION

This section presents how the training and test dataset were created, the CNN training parameters and the segmentation results obtained by the CNN reported in this paper. It is also important to mention that the data extraction and image preprocessing were performed in Python 3.7, while the development, training and testing of the proposed network were done using MATLAB 2022b.

Training dataset

A set of 800 images (400 systole and 400 diastole) were randomly selected from the CAMUS database [20], this set includes images of good, regular, and poor quality, then, data augmentation modifying rotation, translation, scale, brightness, and contrast was applied to achieve a total of 4800 training images an example of these images is shown in Figure 5. With the use of data augmentation, we managed to attain greater diversity in the training set and, consequently, enhance the generalization achieved by the network.

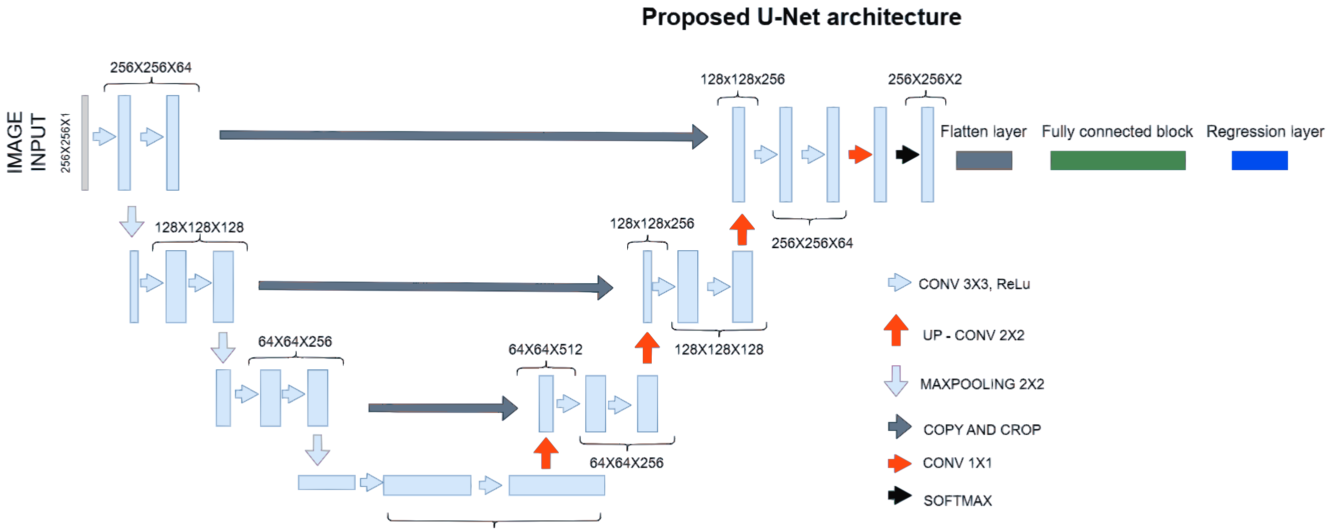


FIGURE 4. **Proposed U-Net architecture.**

Test dataset

The test set used to evaluate our method was divided into 2 parts: the first consists of 98 images from the CAMUS database, comprising 49 systole images and 49 diastole images (hold-out set). The second part consists of non-training images from the EchoNet-dynamic database [21], these images were extracted from .AVI videos at the end of the systolic and diastolic cycles. The task was done by Cervantes - Guzmán in [24]. The above-mentioned images were subsequently annotated by an expert and the segmentation masks were obtained. The total number of systole images is 207, while for diastole, it is 211. It is worth noting that the resolution of the images from the EchoNet-dynamic database is 112x112 pixels, so they were scaled to 256x256, the resolution at which the proposed network is trained.

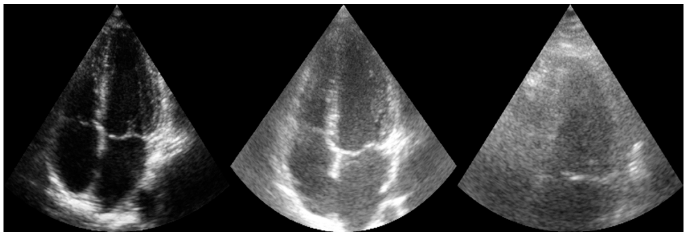


FIGURE 5. **Examples of good, regular and bad quality images from the CAMUS database.**

CNN training parameters

The training parameters for the CNN were defined as follows:

- Training Images = 4560 images.
- Validation images = 240 images.
- Number of epochs = 20.
- Learning rate = 0.0001.
- Batch size = 32.
- Encoder depth = 3.
- Start filters = 64.

A batch size of 32 images is generally manageable on most modern GPUs, ensuring efficient use of hardware resources without causing memory issues. Additionally, this batch size introduces a moderate amount of noise in the gradient updates, providing some regularization effects without being too small. Empirical tests were conducted, and the training time was found to be acceptable, with the model converging effectively without overfitting. Additionally, the mean squared error (RMSE) was used as the loss function since the final layer of the network is a linear regression layer. Therefore, the RMSE tends to penalize larger errors more significantly, leading to better weight adjustments in the network compared to accuracy. The average RSME in the training stage was: 9.86.

This architecture was implemented on a PC with 32 GB of RAM, GPU NVIDIA Tesla K40c and a NVIDIA Tesla T4 working in parallel into an Ubuntu server environment.

PDM training parameters

- Training Images = 4800 images.
- Number of landmarks per example = 64.
- Explained variance = 0.9.

These parameters were selected as follows, the number of landmarks (64) are a good point density number and can represent the LV contour with good quality as depicted in Figure 6; normally the explained variance as proved in [9] is about 90 %, thus characterizing the most significant shape information provided by the PCA. Finally, the training set was the same as the one selected for the CNN training.

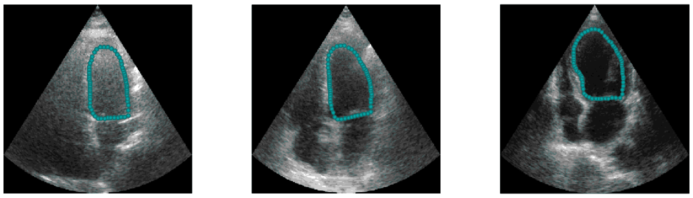


FIGURE 6. **Example of landmarks sampled in the LV training images.**

Segmentation results

In this section, we present the results obtained by our method on the test dataset. In Figures 7 and 8, are shown examples of image segmentation results for the CAMUS database during systole and diastole. Figures 9 and 10 illustrate the segmentation for both phases using images from the EchoNet Dynamic database. Additionally, Table 1 displays the average Dice coefficients for both test sets. Subsequently, Table 2 compares our method against others described in [20] and [22]. The average Dice value in this table represents the average obtained from the CNN on the CAMUS dataset only. Furthermore, we decided to categorize the obtained segmentations based on their Dice coefficient into three categories: good, regular, and bad. The threshold used for this categorization is shown in Table 3. Meanwhile, in Figure 11a and 11b, the median for each category is observed for systole images from the CAMUS and EchoNet-Dynamic databases. Figures 11c and 11d display the median for each category for diastole images in the mentioned databases.

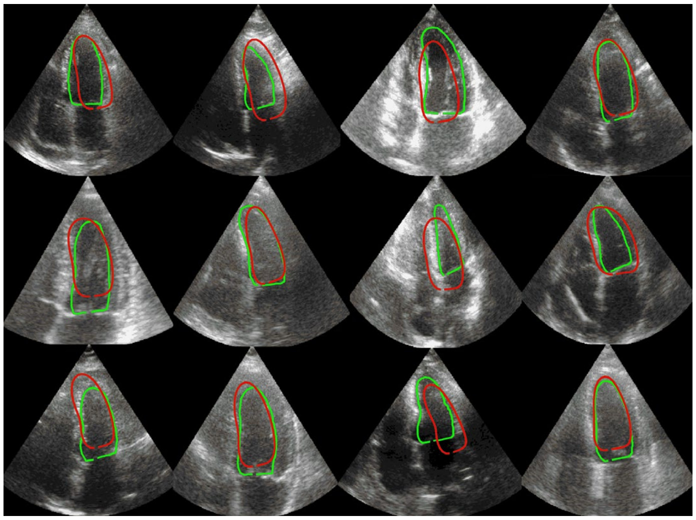


FIGURE 7. **Systole CNN LV shape segmentations (red) vs. ground truth (green) for CAMUS database.**

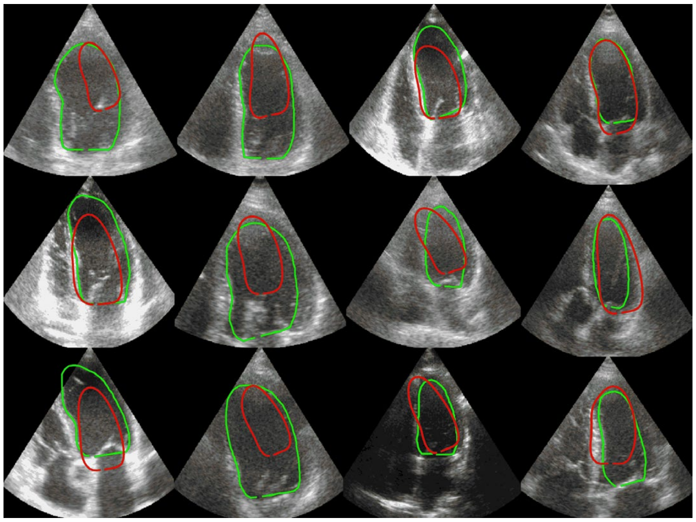


FIGURE 8. **Diastole CNN LV shape segmentations (red) vs. ground truth (green) for CAMUS database.**

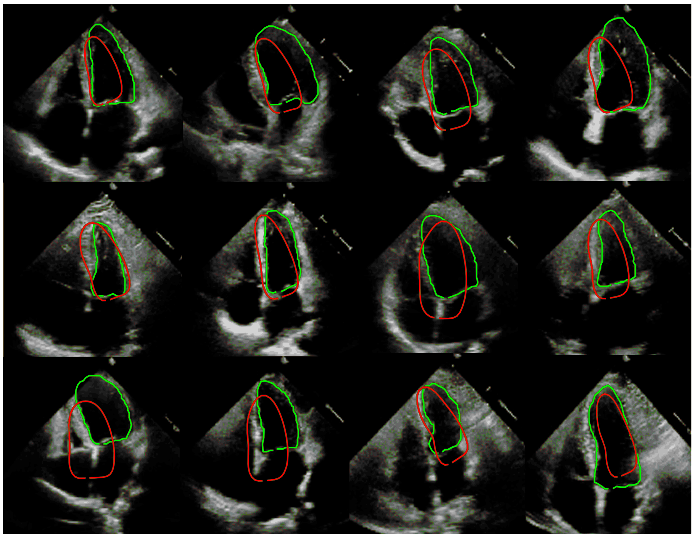


FIGURE 9. **Systole CNN LV shape segmentations (red) vs. ground truth (green) for EchoNet-dynamic database.**

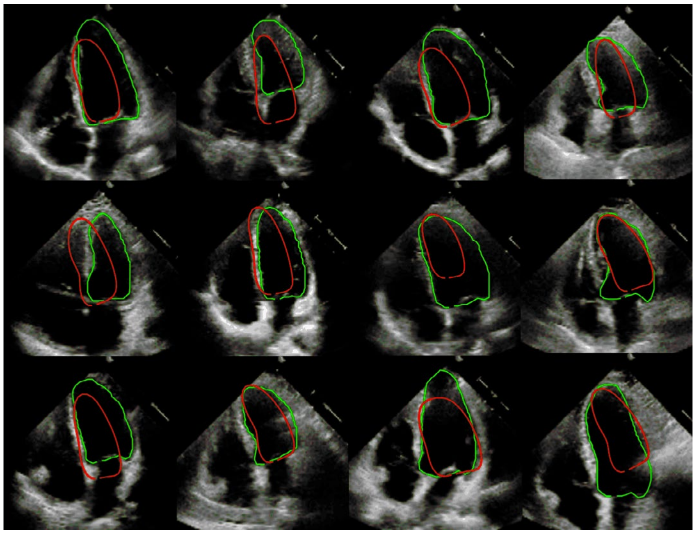


FIGURE 10. **Diastole CNN LV shape segmentations (red) vs. ground truth (green) for EchoNet-dynamic database.**

TABLE 1. **Mean Haussdorff distance and mean Dice results for CAMUS and EchoNet-dynamic database**

	CAMUS database		EchoNet-dynamic database	
Metric	**Systole**	**Diastole**	**Systole**	**Diastole**
Haussdorff distance (px)	25.92	37.55	28.95	31.20
Dice coefficient	0.71	0.66	0.64	0.65

TABLE 2. **Comparison between the proposed U-Net versus other shape-based LV segmentation works**

Method	Mean Dice coefficient systole	Mean Dice coefficient diastole
BEASM-SEMI [20]	0.86	0.92
BEASM-FULLY [20]	0.82	0.87
ASM [22]	0.82	0.81
FMR-ASM [22]	0.83	0.80
Proposed U-Net	**0.71**	**0.66**

TABLE 3. **Category classification by Dice coefficient**

Category	Threshold
Good	Dice coefficient > = 0.80
Regular	0.70 < = Dice coefficient < = 0.79

Results analysis

In this section, we discuss the results presented in the previous section. The first point to address is the performance of our U-Net, both for the CAMUS database it was trained on and the EchoNet-dynamic database, which contains completely new images for the network to segment. As observed in Figures 11a - 11d, the performance is very similar, which suggests that there is no significant overfitting due to data augmentation. We can also understand that the performance of the "good" category is competent in both cases (CAMUS and EchoNet-dynamic) with respect to the results presented in [20]. On the other hand, in Figures 7 to 10, it is evident that all segmentations produced by our method show shape characteristics very similar to those of the left ventricle (LV). They exhibit smoothed shapes located on the correct side of the image, with scale and rotation close to the expert annotation. Even in the case of poor segmentations (see Figure 12), the obtained contours preserve the shape qualities of the left ventricle. This is the effect achieved by the statistical shape constraint arising from training the network with parameters derived from a PDM of the LV. This represents an advantage over convolutional networks based on semantic segmentation, which, when they fail, produce classification errors like the one shown in Figure 1. Such misclassifications are challenging to correct since contour extraction yields noisy shapes and sometimes contours are located far from the LV, a situation in which our method proves to be more robust. Another point to consider, which impacts our method, is the quality of the images. If ultrasound already poses a significant challenge due to speckle noise, the images acquired in both databases are not of the highest quality. This affects the segmentations obtained by our method and makes it difficult to find the optimal features corresponding to the pose and shape values during the network training hence the low Dice average in the tests performed. However, unlike semantic segmentation, these segmentations can be corrected by some other method that performs a LV contour fine-tuning.

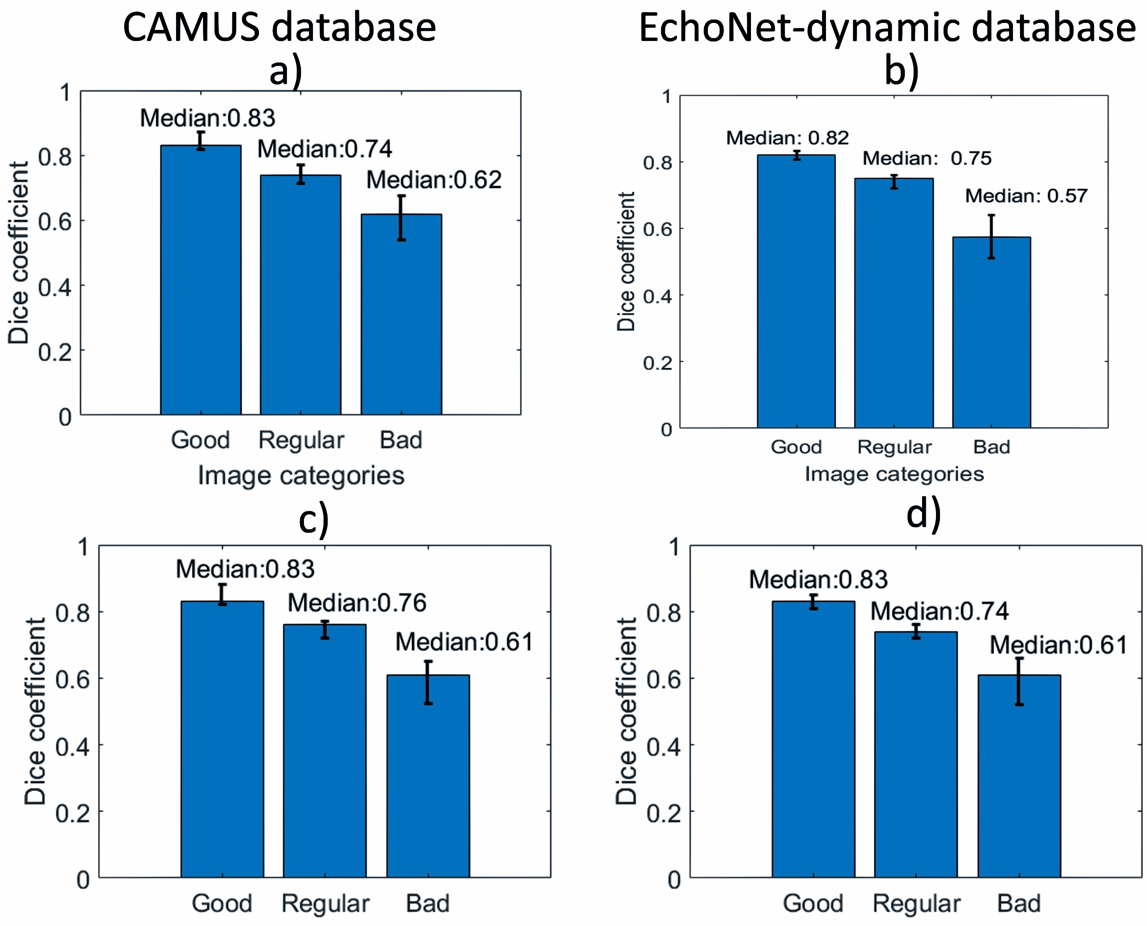


FIGURE 11. Median of each category for systole (11a, 11b) and diastole segmentations (11c, 11d) in CAMUS and EchoNet-dynamic database.

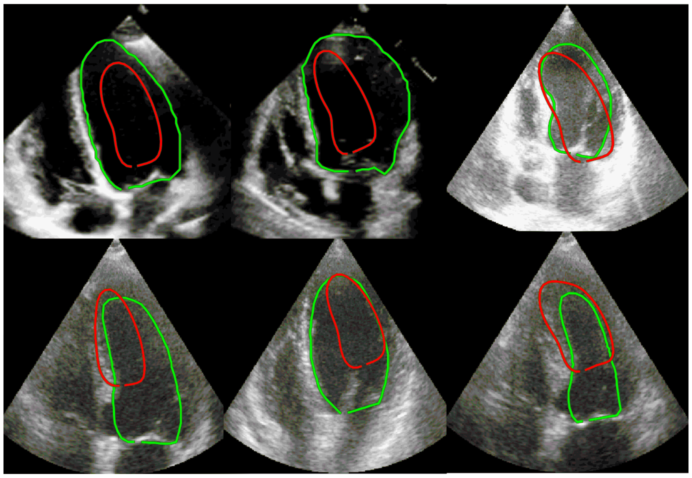


FIGURE 12. **Diastole CNN LV shape segmentations (red) vs. ground truth (green) for EchoNet-dynamic database.**

CONCLUSIONS

In this paper we presented an alternative way to the use of CNNs for LV shape and pose parameter prediction, taking advantage of convolution layers to find features extracted from the ultrasound image, resulting in segmentations with statistically valid shapes as the network is trained with the shape parameters of a PDM. Although, currently, the overall segmentation accuracy is not high, when high quality images are selected the accuracy of our method compares favorably to other previously published work. Our method avoids the appearance of blobs, and the fact that the segmentation is always a statistically valid shape allows it to be used as an initialization mechanism for another method to perform an adjustment and improve the segmentation starting from a smoothed shape and not from a noisy shape and in some cases far from the ventricle, as it would be the case with the extraction of the contour of a semantic segmentation CNN. In conclusion, this method explores the possibility of generating segmentations with statistical shape restrictions using the power of CNNs and can also be used as an automatic initialization method to later fine tune the predicted LV segmentation.

ACKNOWLEDGMENT

This work is supported by UNAM under projects: PAPIIT IV100420, PAPIIT IA104622 and “Programa de Becas Posdoctorales de DGAPA”. Fabian Torres acknowledges the support of the Postdoctoral Fellowship granted by CONAHCYT (CVU 298645).

AUTHOR CONTRIBUTIONS

E.G.G. conceptualized the project, participated in the data curation, carried out formal analyses, contributed to investigation and the development of methodology and the use of specialized software, validated results and wrote the manuscript in their different stages of development. F.T.R. carried out formal analyses, participated in the development of methodology and in all the stages of the writing of the manuscript. J.L.P.G. carried out formal analyses, investigation, participated in the development of methodology and the writing, reviewing and editing of the manuscript. B.E.R. conceptualized and oversaw the project, carried out formal analyses, obtained resources for the project, and participated in the writing, reviewing and editing of the manuscript. F.A.C. conceptualized and oversaw the project, carried out formal analyses, obtained resources for the project, and participated in the writing, reviewing and editing of the manuscript. All Both authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] N. Paragios and R. Deriche, “Geodesic active regions and level set methods for supervised texture segmentation,” Int. J. Comput. Vis., vol. 46, no. 3, pp. 223-247, Feb. 2002, doi: https://doi.org/10.1023/A:1014080923068

[2] N. Paragios, “A level set approach for shape-driven segmentation and tracking of the left ventricle,” IEEE Trans. Med. Imag., vol. 22, no. 6, pp. 773-776, Jun. 2003, doi: https://doi.org/10.1109/TMI.2003.814785

[3] J. C. Nascimento and J. S. Marques, “Robust shape tracking with multiple models in ultrasound images,” IEEE Trans. Image Process., vol. 17, no. 3, pp. 392-406, Mar. 2008, doi: https://doi.org/10.1109/TIP.2007.915552

[4] V. Zagrodsky, V. Walimbe, C. R. Castro-Pareja, J. X. Qin, J.-M. Song, and R. Shekhar, “Registration-assisted segmentation of real-time 3-D echocardiographic data using deformable models,” IEEE Trans. Med. Imaging., vol. 24, no. 9, pp. 1089-1099, Sep. 2005, doi: https://doi.org/10.1109/tmi.2005.852057

[5] B. Georgescu, X. S. Zhou, D. Comaniciu, and A. Gupta, “Database-guided segmentation of anatomical structures with complex appearance,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 429-436 vol. 2, doi: https://doi.org/10.1109/CVPR.2005.119

[6] S. C. Mitchell, B. P. F. Lelieveldt, R. J. van der Geest, H. G. Bosch, J. H. C. Reiber, and M. Sonka, “Multistage hybrid active appearance model matching: Segmentation of left and right ventricles in cardiac MR images,” IEEE Trans. Med. Imag., vol. 20, no. 5, pp. 415-423, May 2001, doi: https://doi.org/10.1109/42.925294

[7] O. Bernard, J. G. Bosch, B. Heyde, M. Alessandrini, et al., “Standardized Evaluation System for Left Ventricular Segmentation Algorithms in 3D Echocardiography,” IEEE Trans. Med. Imaging, vol. 35, no. 4, pp. 967-977, Apr. 2016, doi: https://doi.org/10.1109/TMI.2015.2503890

[8] S. Leclerc, T. Grenier, F. Espinosa and O. Bernard, “A fully automatic and multi-structural segmentation of the left ventricle and the myocardium on highly heterogeneous 2D echocardiographic data,” 2017 IEEE International Ultrasonics Symposium (IUS), Washington, DC, USA, 2017, pp. 1-4, doi: https://doi.org/10.1109/ULTSYM.2017.8092797

[9] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, “Active shape models their training and application,” Comput. Vis. Image Underst., vol. 61, no. 1, pp. 38-59, Jan. 1995, doi: https://doi.org/10.1006/cviu.1995.1004

[10] Y. Ali, S. Beheshti, F. Janabi-Sharifi, “Echocardiogram segmentation using active shape model and mean squared eigenvalue error,” Biomed. Signal Process. Control, vol. 69, art. no. 102807, Aug. 2021, doi: https://doi.org/10.1016/j.bspc.2021.102807

[11] M. A. Shoaib, J. Huang Chuah, R. Ali, K. Hasikin, et al., “An Overview of Deep Learning Methods for Left Ventricle Segmentation,” Comput. Intell. Neurosci., vol. 2023, art. no. 4208231, 2023, doi: https://doi.org/10.1155/2023/4208231

[12] N. Savioli, M. S. Vieira, P. Lamata, G. Montana, “Automated Segmentation on the Entire Cardiac Cycle Using a Deep Learning Work - Flow,” in 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 2018, pp. 153-158, doi: https://doi.org/10.1109/SNAMS.2018.8554962

[13] X. Zou, Q. Wang, T. Luo, “A novel approach for left ventricle segmentation in tagged MRI,” Comput. Electr. Eng., vol. 95, art. no. 107416, Oct. 2021, doi: https://doi.org/10.1016/j.compeleceng.2021.107416

[14] T. Wech, M. J. Ankenbrand, T. A. Bley, J. F. Heidenreich, “A data-driven semantic segmentation model for direct cardiac functional analysis based on undersampled radial MR cine series,” Magn. Reason. Med., vol. 87, no. 2, pp. 972-983, Feb. 2022, doi: https://doi.org/10.1002/mrm.29017

[15] G. Veni, M. Moradi, H. Bulu, G. Narayan, and T. Syeda-Mahmood, “Echocardiography segmentation based on a shape-guided deformable model driven by a fully convolutional network prior,” 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 2018, pp. 898-902, doi: https://doi.org/10.1109/ISBI.2018.8363716

[16] W.-Y. Hsu, “Automatic Left Ventricle Recognition, Segmentation and Tracking in Cardiac Ultrasound Image Sequences,” IEEE Access, vol. 7, pp. 140524-140533, 2019, doi: https://doi.org/10.1109/ACCESS.2019.2920957

[17] Y. Li, W. Lu, P. Monkam, Z. Zhu, W. Wu, M. Liu, “LVSnake: Accurate and robust left ventricle contour localization for myocardial infarction detection,” Biomed. Signal Process. Control, vol. 85, art. no. 105076, Aug. 2023, doi: https://doi.org/10.1016/j.bspc.2023.105076

[18] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, “Deep Snake for Real-Time Instance Segmentation,” 2020, arXiv: 2001.01629, doi:https://doi.org/10.48550/arXiv.2001.01629

[19] O. Ronneberger, P. Fischer, T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, Munich, Germany, 2015, pp. 234-241, doi: https://doi.org/10.1007/978-3-319-24574-4_28

[20] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, et al., “Deep Learning for Segmentation using an Open Large-Scale Dataset in 2D Echocardiography,” IEEE Trans. Med. Imaging, vol. 38, no. 9, pp. 2198-2210, Sep. 2019, doi: https://doi.org/10.1109/TMI.2019.2900516

[21] EchoNet-Dynamic: a Large New Cardiac Motion Video Data Resource for Medical Machine Learning, github, 2019. [Online]. Available: https://echonet.github.io/dynamic/index.html#motivation

[22] E. Galicia Gomez, F. Torres Robles, B. Escalante Rmirez, J. Olveres, F. Arámbula Cosío, “Full multi resolution active shape model for left ventricle segmentation,” in 17th International Symposium on Medical Information Processing and Analysis, Campinas, Brazil, 2021, doi: https://doi.org/10.1117/12.2606252

[23] M. B. Stegmann, D. Delgado Gomez. (2002). A Brief Introduction to Statistical Shape Analysis [Online]. Available: http://www2.imm.dtu.dk/pubdb/edoc/imm403.pdf

[24] A. Cervantes-Guzmán, K. McPherson, J. Olveres, C. F. Moreno-García, F. Torres Robles, E. Elyan, B. Escalante-Ramírez, “Robust cardiac segmentation corrected with heuristics,” PLoS One, vol. 18, no. 10, art. no. E0293560, Oct. 2023, doi: https://doi.org/10.1371/journal.pone.0293560

RESEARCH ARTICLE
VOL. 44 | NO. 4 | **SPECIAL ISSUE 2023** | PP 152 - 164

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.11

E-LOCATION ID: 1390

Evaluación de la Calidad de los Agentes Conversacionales para la Creación de Instrumentos de Evaluación en Medición de Señales Bioeléctricas

Evaluation of the Quality of Conversational Agents for the Creation of Evaluation Instruments in Bioelectric Signals Measurement

Alberto Isaac Pérez-Sanpablo[1,2], *Marcela D. Rodriguez-Urrea*[3], *María del Carmen Arquer-Ruíz*[2]
Adrian Octavio Ramirez-Morales[2], *Alicia Meneses-Peñaloza*[1]

[1] Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra - México
[2] Universidad La Salle - México
[3] Universidad Autónoma de Baja California - México

RESUMEN

El objetivo de esta investigación es evaluar la calidad de agentes conversacionales basados en Modelos de Lenguaje Grandes, para la evaluación de aplicación de conocimiento en Ingeniería Biomédica. Se desarrolló un instrumento de evaluación sobre seis temas de medición de señales bioeléctricas elaborado por un agente humano y los agentes conversacionales Chat-GPT y Bard. Se evaluó la calidad del instrumento en términos de nivel de pensamiento, validez, relevancia, claridad, dificultad y capacidad de discriminación, mediante índice kappa (k) del acuerdo de dos expertos y análisis Rasch de resultados de treinta y ocho estudiantes. Tras eliminar siete preguntas de los agentes conversacionales por problemas de validez y originalidad se integró un instrumento de seis preguntas. Las preguntas fueron válidas y relevantes, claras (>0.95, k=1.0), con dificultad baja a alta (0.61-0.87, k=0.83), índice de discriminación adecuado (0.11-0.47), a nivel de pensamiento de análisis (k=0.22). El promedio de los estudiantes fue de 7.24±2.40. Este es el primer análisis crítico de la calidad de los agentes conversacionales a un nivel de pensamiento superior al de comprensión. Los agentes conversacionales presentaron limitaciones en términos de validez, originalidad, dificultad y discriminación en comparación con el experto humano lo que resalta la necesidad aún de su supervisión.

PALABRAS CLAVE: Bard, Chat-GPT, evaluación educativa, ingeniería biomédica, inteligencia artificial

ABSTRACT

This research aims to evaluate the quality of conversational agents based on Large Language Models for evaluating the application of knowledge in Biomedical Engineering. An evaluation instrument was developed on six topics for measuring bioelectrical signals prepared by a human agent and the conversational agents Chat-GPT and Bard. The quality of the instrument was evaluated in terms of level of thinking, validity, relevance, clarity, difficulty, and discrimination capacity, using the kappa (k) index of the agreement of two experts and Rasch analysis of results from thirty-eight students. After eliminating seven questions from the conversational agents due to validity and originality problems, a 6-question instrument was integrated. The questions were valid and relevant, clear (>0.95, k=1.0), with low to high difficulty (0.61-0.87, k=0.83), adequate discrimination index (0.11-0.47), at the analysis level of thinking (k =0.22). The average score of the students was 7.24±2.40. This is the first critical analysis of the quality of conversational agents at a level of thinking higher than comprehension. The conversational agents presented limitations in terms of validity, originality, difficulty, and discrimination compared to the human expert, which highlights the need for their supervision.

KEYWORDS: artificial intelligence, Bard, biomedical engineering, Chat-GPT, educational measurement

Autor de correspondencia

DESTINATARIO: **Alberto Isaac Pérez-Sanpablo**
INSTITUCIÓN: **Institution Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra**
DOMICILIO: **Calz México-Xochimilco 289, Coapa, Col. Arenal de Guadalupe, Tlalpan, 14389 Ciudad de México, México**
CORREO ELECTRÓNICO: **albperez@inr.gob.mx**

Recibido:

31 oct 2023

Aceptado:

18 dic 2023

INTRODUCCIÓN

Los avances recientes en inteligencia artificial (IA) y Modelos de Lenguaje Grandes (LLMs, por sus siglas en inglés) [1], están permitiendo desarrollar agentes conversacionales capaces de asistir al humano en tareas que requieren de conocimiento especializado. Ejemplos de estos agentes son los modelos de lenguaje ChatGPT y Bard [2], reemplazado por Gemini en diciembre del 2023, que están siendo altamente adoptados para apoyar actividades de educación, investigación y la práctica de las Ciencias Médicas [3].

Dado que los modelos de lenguaje de estos agentes fueron generados mediante un entrenamiento con una gran cantidad de datos, son capaces de generar respuestas coherentes, pero no necesariamente con un alto grado de precisión para todos los dominios de aplicación. Por lo anterior, es importante evaluar su desempeño para asistir en tareas académicas en el ámbito de la Ingeniería Biomédica. Lo anterior ayudaría a establecer sus alcances y limitaciones para integrarlos a la práctica docente en este campo, y, por otro lado, permitiría identificar oportunidades para mejorar su desempeño. En este artículo presentamos un procedimiento novedoso para evaluar la calidad de preguntas generadas a través de los modelos de lenguaje ChatGPT y Bard, que conformaron un instrumento de evaluación de 6 preguntas, y que fue aplicado a estudiantes de la carrera de Ingeniería Biomédica.

Modelos de Lenguaje

Los LLMs han permitido generar agentes que proporcionan un estilo de interacción conversacional, similar al de las personas [4][5] además de que no limitan su respuesta a dominios específicos de conocimiento.

Un modelo de lenguaje o LLM, tal como el de los agentes ChatGPT y Bard, resultan de pre-entrenar una red neuronal con una gran cantidad de datos públicos [6][7][8]. El modelo resultante es una función matemática que representa la distribución de probabilidad de secuencias de palabras. Entonces, dada una secuencia de palabras, los modelos de lenguaje de estos agentes, generan el siguiente token de palabra dependiendo de la probabilidad estimada para las palabras anteriores. Este mecanismo subyacente de los LLMs de predecir la siguiente palabra es uno de los factores importantes que explican por qué los agentes basados en LLMs pueden generar respuestas diferentes para la misma pregunta, que suenan plausibles, pero que, además, podrían contener errores fácticos [6]. Tales modelos son perfeccionados mediante un proceso conocido como aprendizaje por refuerzo a partir de la retroalimentación humana (RLHF), que consiste en entrenadores de IA humanos que brindan recompensas al modelo para ayudarlo a corregir sus errores [6][7][8].

Adopción de los Modelos de Lenguaje

El ChatGPT (del inglés, *"Generative Pre-trained Transformer"*), fue lanzado por el laboratorio de investigación de inteligencia artificial, OpenAI, en noviembre de 2022. Tal evento incrementó rápidamente el interés de la sociedad por la Inteligencia Artificial ya que el volumen de tráfico en el Internet por buscar la palabra "IA" y "ChatGPT" se triplicó, pasando de 7.9 millones de búsquedas en noviembre del 2022 a más de 30.4 millones a principios del 2023 [9]. La adopción del ChatGPT ha sido principalmente por la población adulta joven de 18 a 44 años [10], lo que incluye estudiantes. Por ejemplo, dentro de la enseñanza en el área del cuidado de la salud, se ha propuesto su uso para mejorar el aprendizaje personalizado como tutor mediante la resolución de dudas, y elaboración de resúmenes [3][11]. Recientemente, en marzo del 2023, Google lanzó el agente conversacional Bard, considerado el competidor directo de ChatGPT. Aun con el poco tiempo que tiene disponible para accederse por los usuarios, se puede identificar el interés por estudiar su utilidad en el campo de la salud [12]. Este interés radica principalmente, por sus similitudes en las tecnologías de IA que utilizan, pero son modelos que resultaron de un entrenamiento con diferentes fuentes de datos públicos, pero de diferentes tamaños. El modelo de ChatGPT fue entrenado en 2021 con una fuente de datos de texto masiva, mayor que la usada para entrenar el modelo de Bard, mientras que Bard fue entrenado con datos recientes obtenidos exclusivamente de la Web y es capaz de buscar respuestas en Internet en tiempo real.

El interés en estos agentes conversacionales incluye consultarles sobre temas especializados del mundo académico para facilitar el proceso de enseñanza-aprendizaje [3][12][13], ya que son capaces de dialogar de forma escrita, dando respuestas que potencialmente son consistentes y contextualmente apropiadas a las preguntas del usuario. Sin embargo, es necesario tener en cuenta que estas herramientas no pueden interpretar ni comprender a profundidad el contenido de un diálogo, por lo que podrían generar información incorrecta [14]. Un problema reconocido de la tecnología LLM es su tendencia a producir información que no se basa en sus datos de entrenamiento, conocida como "alucinación" [15]. En consecuencia, la información que generan podría erróneamente percibirse como confiable por lectores no especializados, tal como estudiantes.

Trabajos previos

Para utilizar de forma confiable a los agentes conversacionales basados en LLMs, se han formulado guías para integrarlos a la educación superior, que sugieren verificar los hechos o conocimiento que proporcionan, utilizar tal conocimiento con un juicio crítico, verificando fuentes alternas de información [16]. Otros estudios, han evaluado su utilidad para asistir en actividades y áreas académicas específicas. En esta dirección, algunos estudios han evaluado el desempeño del ChatGPT para responder preguntas de exámenes profesionales de diversas especialidades, tal como el examen de la Barra de Abogados de EE. UU [17][18] y de Licencia Médica de EE. UU. (USMLE) [19][20] con una exactitud superior al 57.6 %. Por otro lado, los resultados en términos de la precisión de ChatGPT puede variar entre disciplinas o temas de una misma disciplina. Por ejemplo, mientras que el ChatGPT resultó eficaz para responder preguntas de conocimiento de primer y segundo orden sobre microbiología con una precisión del 80% [4], su rendimiento para responder preguntas clínicas en el campo de la enfermedad glomerular es deficiente (<=60 %)[21]. Lo anterior puede deberse a que los datos de entrenamiento para generar el modelo ChatGPT carecieron de datos ejemplos suficientes en esos temas [3].

Algunos estudios han investigado el uso de agentes conversacionales basados en LLM para generar instrumentos de evaluación, ver Tabla 1. Cheung utilizó ChatGPT Plus para generar 50 preguntas de opción múltiple sobre temas de posgrado de medicina interna y cirugía [22]. Su estudio evaluó la idoneidad, claridad, especificidad, relevancia y poder discriminativo de las preguntas, comparándolas con 50 preguntas generadas por expertos humanos. Los resultados mostraron que ChatGPT Plus generó preguntas con calidad comparable a las del experto humano [22]. Kumar usó ChatGPT para generar seis preguntas de opción múltiple sobre otosclerosis para evaluar diferentes niveles de aprendizaje con diferentes grados de dificultad, pero sin realizar una evaluación formal [23]. Elkins utilizó InstructGPT para generar 612 preguntas abiertas de nivel secundaria sobre temas de biología y aprendizaje automático, utilizando la taxonomía de *Bloom* para evaluar diferentes niveles de aprendizaje [24]. Once expertos evaluaron la calidad de las preguntas en términos de relevancia, gramática y utilidad, con resultados favorables [24]. Yuan utilizó GPT-3 para crear cinco preguntas abiertas para estudiantes de primaria y secundaria sobre diversos temas [25]. Su objetivo era evaluar un método para seleccionar preguntas de calidad entre un conjunto de preguntas generadas con IA. Se realizó evaluación automática contra preguntas de referencia y evaluación por 87 expertos demostrando la eficacia de su método [25]. Finalmente, Xiao utilizó ChatGPT y GPT-2 para realizar 20 preguntas de opción múltiple sobre inglés nivel secundaria [26]. Se realizó evaluación automática y por 373 estudiantes. Sus resultados indicaron que el material generado por IA puede superar en calidad al material escrito por humanos [26].

Este trabajo considera que agentes conversacionales, como ChatGPT y Bard, son una poderosa herramienta para simplificar y mejorar la eficiencia de tareas de los educadores, como las relacionadas con resumir y traducir información [13], organizar y generar material y corregir textos [27][3]. Sin embargo, no hay evidencia de su desempeño para generar exámenes de conocimiento en temas de Ingeniería Biomédica. Consecuentemente, es importante determinar cuáles son las limitaciones y posibles sesgos de utilizar agentes conversacionales basados en LLMs para apoyar tales tareas académicas. Sin embargo, las investigaciones son escasas y limitadas a dominios educativos más generales que requieren niveles de pensamiento básicos para su resolución, tal como usar los LLMs para generar exámenes sobre la comprensión de textos [28]. La adquisición de conocimiento puede evaluarse a diferentes niveles, desde recordar y comprender hasta evaluar y crear de acuerdo con la taxonomía de Bloom [29]. Aquí nos enfocamos en la aplicación de conocimiento que se encuentra en un nivel superior a los estudios previos. Evaluar la aplicación de conocimiento de los estudiantes permite evaluar su capacidad no solo de comprender información sino de usar el conocimiento aprendido en situaciones nuevas [29]. Este nivel de aprendizaje es fundamental para garantizar que los estudiantes podrán utilizar su conocimiento para la solución de problemas reales. Lo anterior, nos motivó a realizar esta investigación con el siguiente propósito.

Objetivo

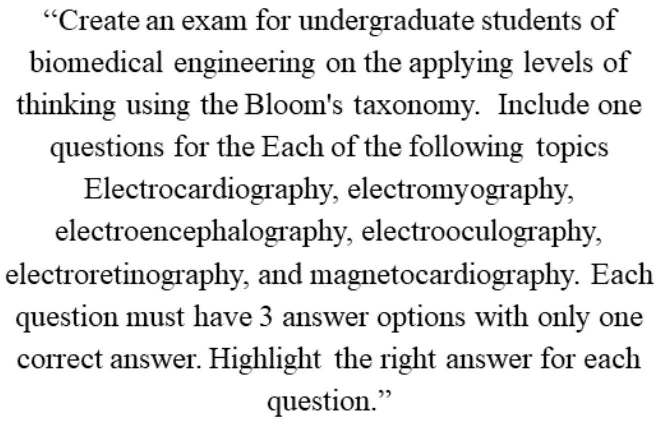
El objetivo de esta investigación fue analizar la calidad de ChatGPT y Bard para crear instrumentos de evaluación en temas de la Ingeniería Biomédica. Debido a las limitaciones de la tecnología LLM, en este trabajo se utiliza un enfoque crítico para evaluar y comparar la calidad en términos de validez, relevancia, claridad, dificultad y capacidad de discriminación de las preguntas generadas por los agentes conversacionales para la evaluación de aplicación de conocimiento sobre medición de señales bioeléctricas en comparación con las de un experto humano.

MATERIALES Y MÉTODOS

Desarrollo del Instrumento de Evaluación

Un experto humano (agente E) y dos agentes conversacionales (agentes B y C) generaron seis potenciales preguntas cada quién para integrar un instrumento de evaluación para estudiantes de octavo semestre de ingeniería biomédica al final de un curso sobre mediciones biomédicas impartido por el agente E. Previo a la generación de todas las preguntas, el agente E definió que el instrumento de evaluación estuviera integrado por preguntas de forma equitativa tanto de agentes como de temas a un nivel de aplicación de conocimiento dentro de la taxonomía de Bloom [29]. Los temas evaluados fueron electrocardiografía, electromiografía, electroencefalografía, electro-oculografía, electrorretinografía y magneto cardiografía.

El agente E utilizó Bard como agente B y ChatGPT-3 como agente C debido a que Bard y ChatGPT-3 son los dos agentes más utilizados en mayo de 2023 con acceso gratuito, facilitando su uso para esta investigación y la generalización de nuestros resultados. Bard y ChatGPT-3 son agentes ampliamente investigados, con buena reputación, y capacidades comparables debido a la similitud de sus tecnologías.



"Create an exam for undergraduate students of biomedical engineering on the applying levels of thinking using the Bloom's taxonomy. Include one questions for the Each of the following topics Electrocardiography, electromyography, electroencephalography, electrooculography, electroretinography, and magnetocardiography. Each question must have 3 answer options with only one correct answer. Highlight the right answer for each question."

FIGURA 1. **Ejemplo de un mensaje de ChatGPT-3 para generar preguntas a nivel de aplicación sobre la interfaz electrodo-electrolito en señales bioeléctricas.**

TABLA 1. **Trabajos previos sobre LLMs para generación de instrumentos de evaluación de conocimiento.**

Dominio / Artículo	Cheung 2023[22]	Kumar 2023[23]	Elkins 2023[24]	Yuan 2023[25]	Xiao 2023[26]
Modelo de Grandes Lenguajes	ChatGPT plus	ChatGPT	InstructGPT	GPT-3	ChatGPT y GPT-2
Tipo de preguntas	Opción múltiple	Opción múltiple	Abiertas	Abiertas	Opción múltiple
Número de preguntas	50 generadas por IA y 50 por dos expertos humanos	6 preguntas generadas por IA	612 preguntas generadas por IA	5 preguntas	20 preguntas
Nivel y disciplina	Posgrado en Medicina	Medicina	secundaria	Nivel primaria y secundaria	Inglés nivel secundaria
Temas específicos	Medicina interna y cirugía	Otoesclerosis	Machine Learning y Biología	Temás diversos y de literatura infantil	Inglés
Nivel de aprendizaje evaluado	No especificado	Memoria, comprensión, afectivo y diferente nivel de dificultad.	Recordar, comprender, aplicar, analizar, evaluar y crear según taxonomía de Bloom en niveles principiantes, intermedios y avanzados	Comprensión de lectura	Comprensión de lectura
Evaluación	Comparación IA contra experto humano por cinco expertos mediante un formulario estandarizado sobre los siguientes dominios: Idoneidad, claridad y especificidad, relevancia, poder discriminativo, idoneidad para posgrado	Ninguna	Evaluación por 11 expertos sobre relevancia, gramática, responsabilidad, adherencia y utilidad	Evaluación automática contra referencias en términos de similitud y equivalencia semántica. Evaluación por 87 expertos de corrección gramatical, ofensividad, claridad, relevancia, importancia, especificidad, y capacidad de respuesta	Evaluación automática de diversidad y legibilidad. Evaluación por 373 estudiantes de legibilidad, corrección, coherencia, compromiso, calidad general, calidad relativa, probabilidad de generación por IA, coherencia temática, idoneidad, coincidencia de contenido, utilidad, idoneidad, similitud,
Resultados	La IA puede generar preguntas de posgrado en medicina con calidad comparable a las de un experto. La relevancia de las preguntas de la IA fue menor (p = 0.04).	La IA puede generar preguntas de opción múltiple con diferentes niveles de aprendizaje, y niveles de dificultad.	Las preguntas generadas con IA son de alta calidad (>67 %, k Cohen >0.53) en términos de relevancia, gramática, responsabilidad, adherencia, y suficientemente útiles.	Se demostró la efectividad del método para seleccionar preguntas de alta calidad entre las generadas por la IA.	El material generado por IA puede superar la calidad del material escrito por humanos.

Como se mencionó ambos fueron entrenados con un conjunto de datos público masivo, ambos tienen capacidad para escribir contenido y responder preguntas. Estos dos agentes fueron elegidos debido a su representatividad en la generación de lenguaje natural con el objetivo de simplificar el proceso de comparación y hacerlo manejable para realizar un análisis profundo y detallado tomando en consideración limitaciones de tiempo y presupuesto. En la Figura 1 se muestra un ejemplo de solicitud para que ChatGPT-3 genere sus preguntas a nivel de aplicación de conocimiento sobre los temas seleccionados. Se realizó una solicitud similar para Bard.

El agente E analizó la calidad de cada pregunta para seleccionar aquellas que pudieran integrar el instrumento de evaluación de final respectando el principio de equidad y en orden aleatorio.

Aplicación y Evaluación del Instrumento

Posteriormente un grupo de expertos evaluaron atributos de calidad de las seis preguntas respecto a claridad, validez, nivel de pensamiento de acuerdo con la Taxonomía de Bloom, dificultad, validez y relevancia. Este grupo de expertos estaba compuesto por el agente E y dos expertos adicionales con al menos cinco años de experiencia docente en los temas del curso. Para realizar esta evaluación utilizaron un formulario estandarizado en Word. El formulario contenía una explicación de apoyo para cada atributo a evaluar. La claridad, relevancia y validez de las preguntas se evaluaron mediante una variable dicotómica (sí/no). Además, clasificaron cada pregunta en un nivel de pensamiento (aplicación, análisis o evaluación) y especificaron un grado de dificultad (es decir, bajo, medio o alto). La claridad, validez, nivel de pensamiento, dificultad, validez y relevancia se evaluaron midiendo la consistencia en las respuestas de los expertos humanos a estos aspectos.

El instrumento de evaluación se tradujo y aplicó en español. Las preguntas fueron registradas en la plataforma de aprendizaje Moodle para crear el banco de preguntas del instrumento. Se reclutó a todos los estudiantes (es decir, 39) del último año de Ingeniería Biomédica de la Universidad, hispanohablantes que tomaron el curso sobre Mediciones Biomédicas. Sólo un estudiante no aceptó participar, quedando treinta y ocho participantes que dieron su consentimiento para utilizar sus resultados. Estos participantes accedieron a la plataforma para contestar el examen en línea al mismo tiempo y desde las instalaciones de la Universidad. El instrumento fue configurado para que cada pregunta se presentara de forma individual y secuencial acompañada de una pregunta dicotómica para que los estudiantes calificaran la claridad de las preguntas. El tiempo máximo para la resolución del instrumento de evaluación fue de 2 minutos por pregunta.

Análisis de Estadístico

Se realizó estadística descriptiva general de los resultados obtenidos de los participantes. Se estimó la consistencia entre los evaluadores utilizando el índice kappa de Fleiss (k), ver ecuación (1).

$$k = \frac{P_o - P_e}{1 - P_e} \qquad (1)$$

Donde P_o es la proporción observada de acuerdo entre los evaluadores y P_e es la proporción esperada de acuerdo debido al azar. Para cada pregunta se calcularon índices de dificultad y discriminación. La claridad evaluada por los estudiantes (c) se calculó como la proporción de estudiantes (x) que consideraron clara la pregunta, ver ecuación (2).

$$c = \frac{n}{N} \qquad (2)$$

Donde N es el número total de respuestas.

Para evaluar el índice de discriminación, la muestra de estudiantes se dividió en tres grupos equitativos (33 %)

en función de su desempeño para comparar los resultados entre el grupo con mayor y menor desempeño. Se realizó un análisis Rasch con el software gratuito Jamovi [30], ver ecuación (3).

$$P(X_{ni}=1)=\frac{e^{(d_i-h_n)}}{1+e^{(d_i-h_n)}} \tag{3}$$

Donde P es la probabilidad de que un individuo *n* responda correctamente la pregunta *i*, d_i es la dificultad de la pregunta i y h_n es la habilidad del individuo *n*.

Esto se utilizó para calcular el valor p para el ajuste del modelo, la matriz de correlación de todas las preguntas del examen y el mapa de habilidades de *Wright*, ver ecuación (4).

$$d_i-h_n=ln\left(\frac{P(X_{ni}=1)}{1-P(X_{ni}=1)}\right) \tag{4}$$

Las asociaciones error-claridad y agente generador con dificultad y claridad se evaluaron mediante Chi-cuadrado. Las diferencias entre agentes en cuanto a claridad, dificultad y discriminación se analizaron mediante pruebas de Kruskal Wallis (H), ver ecuación (5).

$$H=\frac{12}{N(N+1)}\sum\frac{R_j^2}{n_j}-3(N+1) \tag{5}$$

Donde *N* es el número total de observaciones, R_j es la suma de rangos del grupo *j* y n_j es el tamaño del grupo.

RESULTADOS Y DISCUSIÓN

El agente E eliminó cuatro preguntas generadas por el agente C por problemas de validez y tres preguntas del agente B por problemas de repetición. De esta forma, se seleccionaron dos preguntas por agente, utilizando las preguntas del agente E necesarias para completar la evaluación de los seis temas propuestos respetando el principio de equidad. El diseño del instrumento de evaluación final según los agentes quedó BECBCE. Los expertos mostraron un acuerdo perfecto calificando las seis preguntas del instrumento de evaluación como claras, así como válidas y relevantes. Los expertos mostraron un acuerdo razonable (kappa=0.22) y casi perfecto (kappa=0.83) en cuanto a su asignación de nivel de pensamiento y dificultad de las preguntas del instrumento de evaluación, ver Tabla 2. De forma general el instrumento integró dos preguntas de cada uno de los tres niveles de dificultad. La mayoría de las preguntas fueron interpretadas como preguntas de análisis de acuerdo con los expertos a excepción de la pregunta dos que fue considerada como de aplicación por dos expertos. Así, el agente E fue el único en generar una pregunta de acuerdo con el nivel de pensamiento deseado. La calificación promedio obtenida fue de 7.24, con una desviación estándar de 2.40, en un rango de puntuación total de 1.67 a 10 puntos.

TABLA 2. **Análisis de expertos de las preguntas del instrumento de evaluación.**

Pregunta	Agente	Nivel de pensamiento			Dificultad		
		E	E1	E2	E	E1	E2
Q1	B	2	2	2	m	m	m
Q2	E	1	3	1	m	a	m
Q3	C	2	2	2	a	a	a
Q4	B	2	2	2	b	b	b
Q5	C	1	2	2	b	b	b
Q6	E	1	2	2	a	a	a

E: agente E, E1: evaluador 1, E2: evaluador 2. Para nivel de pensamiento: 1 = aplicación, 2 = análisis, 3 = evaluación. Para dificultad: b = baja, m = media, a = alta.

El análisis de Rasch reveló un valor p para el ajuste del modelo de 0.272. La matriz de correlación no mostró correlaciones fuertes entre preguntas, ver Tabla 3. La correlación más alta (r=-0.37) se encontró entre las preguntas Q1, Q5 y Q6 que corresponden a preguntas de electrocardiografía, electroretinografía y magnetocardiografía de tres agentes diferentes.

TABLA 3. **Matriz de correlación del análisis Rasch de preguntas del instrumento de evaluacióninstrumento de evaluación.**

	Q1	Q2	Q3	Q4	Q5
Q2	-0.26	-			
Q3	-0.15	-0.27	-		
Q4	0.16	-0.22	0.08	-	
Q5	-0.37	0.19	-0.14	-0.27	-
Q6	-0.37	0.28	0.06	-0.13	-0.07

Las preguntas del instrumento presentaron un índice de dificultad promedio de 0.72, con una desviación estándar de 0.09, ver Tabla 4. De acuerdo con este índice, la mayoría de las preguntas (N=4) pueden ser consideradas fáciles con un índice de dificultad >0.7. Sin embargo, no siempre el índice de dificultad calculado en base a la respuesta de los alumnos correspondió con la calificación de dificultad asignada por los expertos. Por ejemplo, mientras que las pregunta Q6 calificada como de alta dificultad por los expertos obtuvo uno de los menores índices de dificultad (0.74).

Los estudiantes asignaron una puntuación de claridad a las preguntas en promedio de 97 % con una desviación estándar de 3 %, ver Tabla 4. Dos preguntas obtuvieron una calificación perfecta de claridad correspondientes a las preguntas sobre temas de electrocardiografía y electro oculografía generadas por el agente B.

En la evaluación del índice de discriminación, el grupo de estudiantes con el peor desempeño estuvo conformado por 12 estudiantes que obtuvieron un puntaje menor o igual a 5.00 (32 %), mientras que el grupo con el mejor desempeño por 11 estudiantes (29 %) que obtuvieron una calificación de 10. La pregunta con el índice de dificultad mínimo (0.87) fue generada por el agente C, mientras que la de mayor índice de dificultad fue generada por el agente E. El índice de discriminación promedio fue de 0.31 con una desviación estándar de 0.13. Cinco preguntas (83 %) presentaron un índice de discriminación adecuado (> 0.2). La pregunta con menor índice de dificultad generada por el agente C fue la que también presentó el menor índice de discriminación (0.37).

TABLA 4. **Índice de dificultad y discriminación de las preguntas del instrumento de evaluación.**

Pregunta	Claridad	Índice de dificultad	Índice de discriminación
Q1	1.00	0.74	0.21
Q2	0.97	0.61	0.47
Q3	0.95	0.66	0.37
Q4	1.00	0.74	0.32
Q5	0.92	0.87	0.11
Q6	0.97	0.74	0.37

El mapa de competencias de Wright del análisis Rasch mostró una distribución de competencias de los estudiantes con mayor agrupación por debajo de la media y dos estudiantes sobresalientes, ver Figura 2. El mapa de Wright también mostro una distribución de dificultad de las preguntas con mayor agrupación por encima de la media, con un efecto de techo debido a los once estudiantes que obtuvieron un puntaje perfecto. También se observó redundancia en las preguntas Q1, Q4 y Q6 que corresponden a los temas de electrocardiografía, electro-oculografía y magneto cardiografía. Dos de estas preguntas fueron generadas por el agente B y una por el agente E.

No se encontraron asociaciones significativas (p>0.06) error-claridad, ni agente generador con la dificultad y claridad. No se encontraron diferencias estadísticamente significativas (p>0.08) por agentes en claridad, dificultad y discriminación mediante pruebas de Kruskal Wallis. No se utilizaron pruebas paramétricas debido al número reducido de preguntas que fueron incluidas en el instrumento de evaluación final. La versión final del instrumento de evaluación se muestra en la Tabla 5. Este es el primer trabajo que evalúa mediante un análisis crítico la calidad de los agentes conversacionales en la evaluación de temas relacionados con mediciones biomédicas.

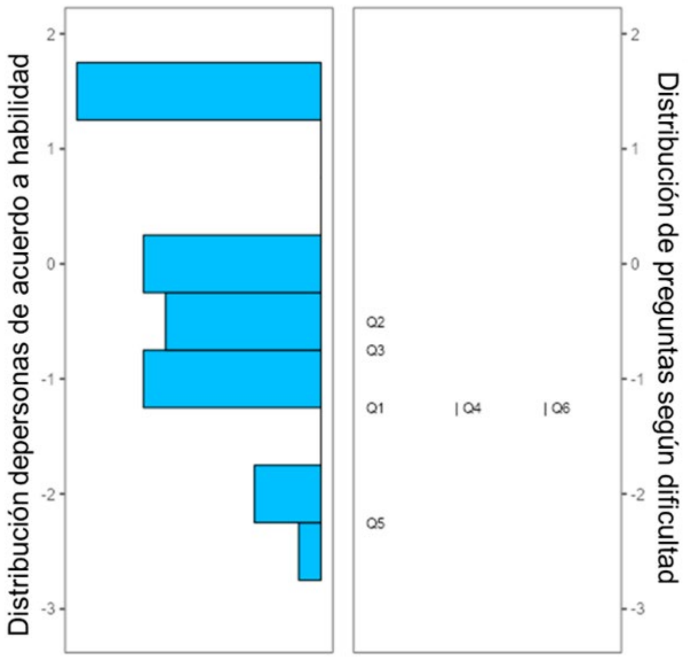


FIGURA 2. **Mapa de Wrigth de análisis Rasch del instrumento de evaluación**

Esto es particularmente destacable para el caso del agente Bard del que no se cuenta con experimentos similares documentados tal vez debido a su más reciente lanzamiento. De forma similar es el primer experimento en realizar una evaluación a un nivel de pensamiento superior al de la simple comprensión de información. La falta de consenso en cuanto a la evaluación de nivel de pensamiento de las preguntas por parte de los expertos puede demostrar la necesidad de revisar el establecimiento de un marco de referencia común entre ellos. Sin embargo, puede identificarse un consenso en la existencia de una diferencia entre el nivel de pensamiento deseado y el generado por los agentes virtuales, tal vez debido a sus limitaciones en interpretación y comprensión [14]. Los problemas de relevancia y originalidad detectados en las preguntas generadas por los agentes virtuales sugieren una limitación para poder utilizar estas herramientas sin supervisión experta.

El análisis de Rasch mostró elevada competencia de los evaluados y baja complejidad de las preguntas. Se puede afirmar que las preguntas del agente E presentaron mayor índice de dificultad (67 %) y discriminación (42 %). Esto pudiera deberse a que este agente fue el que impartió el curso y conocía el contexto de fortalezas y debilidades del grupo. Mientras tanto el agente B generó preguntas redundantes con los menores índices de dificultad (0.74). Por otro lado, el agente C generó preguntas con excelente claridad. De forma interesante, el índice de dificultad del análisis Rasch no siempre guardó relación con el nivel de dificultad asignado por los evaluadores.

Esto revela la importancia del análisis de la dificultad de las preguntas no solo en base al juicio de los expertos sino en el desempeño de los estudiantes. Por otro lado, la agrupación de estudiantes en el extremo superior de desempeño habla de un efecto de techo del instrumento de evaluación.

La ingeniería biomédica es un campo interdisciplinario donde la educación es un componente crucial para la preparación de los futuros ingenieros. De esta manera, los avances en metodologías de aprendizaje y estrategias de evaluación son temas relevantes. Artículos como el presente pueden contribuir a mejorar la calidad de la formación en el área. Por otro lado, la evolución de las formas de aprendizaje ya está impactando la educación en ingeniería biomédica, por lo que los artículos sobre estos temas pueden ayudar a los profesionales a mantenerse actualizados y enfrentar nuevos desafíos. El objetivo principal de nuestra investigación fue evaluar la calidad de agentes conversacionales en el contexto de la medición de señales bioeléctricas, lo cual es fundamental en ingeniería biomédica. Así nuestro artículo es una contribución original a la evaluación de la calidad de los agentes conversacionales a un nivel de pensamiento más significativo que la simple comprensión de conceptos en el área de la medición de señales biomédicas. Creemos importante discutir la necesidad de desarrollar habilidades en estudiantes y profesionales sobre el manejo de agentes conversacionales.

TABLA 5. Instrumento de evaluación final.

Q1: ¿Cuál de las siguientes es la derivación más utilizada en electrocardiografía? •Derivación I •Derivación II •Derivación III
Q2: Está realizando el análisis de un estudio de electromiografía de superficie de un paciente ¿Cuál es el orden correcto para procesar la señal de electromiografía de superficie? •Procesamiento de línea base, rectificación y suavizado •Rectificación, suavizado y procesamiento de línea base •Suavizado, procesamiento de línea base y rectificación
Q3: El registro de EEG de un paciente muestra un ritmo alfa prominente en la región occipital. ¿Qué estado del cerebro es más probable que se asocie con este hallazgo? •Despierto y relajado •Sueño profundo •Alta actividad cognitiva
Q4: ¿Cuál de los siguientes es el electrodo más común utilizado en electrooculografía? •Electrodo de superficie •Electrodo de aguja •Electrodo de alambre
Q5: Está realizando una electrorretinografía en un paciente. ¿Qué configuración de colocación de electrodos se usa comúnmente para registrar las respuestas eléctricas de la retina? •Fpz - Oz •Cz - Pz •Electrodo activo en la córnea y electrodo de referencia en la piel
Q6: Está realizando un estudio de magneto cardiografía a un paciente. ¿Cuáles serían la amplitud y frecuencia máxima de la señal esperados? •5 pT y 100 Hz •5 nT y 1000 Hz •5 fT y 10 Hz

Por otro lado, aunque el número de preguntas incluido en nuestro instrumento de evaluación puede ser una limitante, consideramos que el instrumento actual permitió cumplir con el objetivo de evaluar la calidad del instrumento de evaluación creado con la ayuda de agentes conversacionales en todos los términos propuestos: nivel de pensamiento, validez, relevancia, claridad, dificultad y capacidad de discriminación. Además, los métodos que utilizamos para evaluar la calidad del cuestionario, como la índice kappa (k) para evaluar la concordancia entre los expertos y el análisis de Rasch para determinar la consistencia de los resultados y de los estudiantes, son robustos ya que se centran en analizar la calidad del instrumento en lugar del mero número de pruebas. Sin duda, aumentar el número de preguntas contribuiría a la fiabilidad matemática del estudio. Pero el incremento en el número de preguntas podría afectar las respuestas de los expertos y estudiantes al afectar su calidad y tasa de participación. Por lo tanto, consideramos que, en su versión actual, el instrumento permitió cumplir con los objetivos del estudio y realizar una evaluación eficiente sin comprometer la calidad de los datos.

CONCLUSIONES

Los agentes conversacionales basados en LLMSs tienen capacidad para evaluar la aplicación de conocimiento en medición de señales bioeléctricas. Aunque no se encontraron diferencias de calidad en términos de claridad, dificultad y discriminación en comparación con un experto, se encontraron problemas de validez y originalidad que requieren supervisión de un experto.

Por lo tanto, es necesario realizar más pruebas y estudios para determinar la calidad de la información. Si bien, estas herramientas tienen potencial, la evidencia sustancial que respalda su uso en entornos académicos es fundamental. La alfabetización en estas herramientas en los programas de formación educativa es fundamental. Los docentes deben aprender a usar estas herramientas y comprender claramente sus limitaciones y los objetivos que quieren lograr al usarlas [3]. También es esencial dar a los académicos la autoridad y la independencia para utilizarlos [11]. Sin embargo, es importante que los expertos en el campo correspondiente validen el contenido de las preguntas generadas por estas herramientas para asegurar su calidad y garantizar que cumpla con los estándares académicos requeridos.

DECLARACIÓN ÉTICA

A todos los participantes en este estudio se les solicitó su consentimiento para participar en esta investigación de manera voluntaria previo a la recopilación y análisis de cualquier información. Los participantes que no aceptaron participar fueron excluidos de este estudio.

AGRADECIMIENTOS

Agradecimiento a todos los participantes que ayudaron a esta investigación.

CONTRIBUCIÓN DE AUTORES

A.O.R.M. participó en la curación de datos, realizó investigación y validó resultados, asimismo obtuvo recursos para el desarrollo del proyecto, participó en las diferentes etapas del desarrollo del manuscrito. A.M.P. conceptualizó el proyecto, realizó análisis formales, desarrolló la metodología y validó resultados, participó en la supervisión general del proyecto y participó en las diferentes etapas de la escritura del manuscrito. A.I.P.S. conceptualizó el proyecto, participó en la curación de datos e implementó software especializado, desarrolló análisis formales, validó y visualizó los resultados, realizó investigación y desarrolló la metodología para el proyecto, obtuvo recursos para el desarrollo del proyecto, y se encargó de la administración y supervisión general, participó en las diferentes etapas de la escritura del manuscrito. M.D.R.U. participó en la obtención de fondos y recursos para el desarrollo del proyecto, desarrolló la metodología, la visualización y validación de los resultados, y participó en las diferentes etapas del desarrollo del manuscrito. M.C.A.R. participó en la obtención de recursos para el desarrollo del proyecto, realizó investigación y validó resultados, y participó en las diferentes etapas del desarrollo del manuscrito. Todos los autores aprobaron la versión final del manuscrito.

REFERENCES

[1] OpenAI, “ChatGPT” [Large language model]. OpenAI. https://chat.openai.com/chat (consultado en 2023).

[2] Google, “Bard” [Large Language Model]. 2023. https://bard.google.com/chat (consultado en 2023).

[3] M. Sallam, “ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns,” Healthcare, vol. 11, no. 6, art. no. 887, mar. 2023, doi: https://doi.org/10.3390/healthcare11060887

[4] M. McTear, Z. Callejas, D. Griol, “The conversational interface: Talking to smart devices,” in The Conversational Interface, Switzerland: Springer International Publishing, 2016, doi: https://doi.org/10.1007/978-3-319-32967-3

[5] M. M. E. Van Pinxteren, M. Pluymaekers, J. G. A. M. Lemmink, “Human-like communication in conversational agents: a literature review and research agenda,” J. Serv. Manag., vol. 31, no. 2, pp. 203-225, 2020, doi: https://doi.org/10.1108/JOSM-06-2019-0175

[6] J. Manyika, “An overview of Bard: an early experiment with generative AI,” Google. 2023. [En línea]. Disponible en: https://ai.google/static/documents/google-about-bard.pdf

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al., “Language models are few-shot learners,” 2020, arXiv:2005.14165, doi: https://doi.org/10.48550/arXiv.2005.14165

[8] O. Vinyals, Q. Le, “A Neural Conversational Model,” 2015, arXiv: 1506.05869, doi : https://doi.org/10.48550/arXiv.1506.05869

[9] T. Bianchi, “Global search volume for ‘ChatGPT API’, ‘AI API’ keywords 2022-2023,” Statista. Disponible en: https://www.statista.com/statistics/1398265/chatgpt-ai-api-keywords-search-volume/ (consultado el 26 de octubre de 2023).

[10] A. Petrosyan, “ChatGPT and cyber crime - Statistics & Facts,” Statista. Disponible en: https://www.statista.com/topics/10818/chatgpt-and-cyber-crime/#topicOverview (consultado el 26 de octubre de 2023).

[11] R. S. D'Amico, T. G. White, H. A. Shah, D. J. Langer, “I Asked a ChatGPT to Write an Editorial About How We Can Incorporate Chatbots Into Neurosurgical Research and Patient Care…,” Neurosurgery, vol. 92, no. 4, pp. 663-664, abr. 2023, doi: https://doi.org/10.1227/neu.0000000000002414

[12] F. Y. Al-Ashwal, M. Zawiah, L. Gharaibeh, R. Abu-Farha, A. N. Bitar, “Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools,” Drug Healthc. Patient. Saf., vol. 15, pp. 137-147, sep. 2023, doi: https://doi.org/10.2147/dhps.s425858

[13] H. Yang, “How I use ChatGPT responsibly in my teaching,” Nature, apr. 2023, doi: https://doi.org/10.1038/d41586-023-01026-9

[14] S. Ariyaratne, K. P. Iyengar, N. Nischal, N. Chitti Babu, R. Botchu, “A comparison of ChatGPT-generated articles with human-written articles,” Skeletal Radiol., vol. 52, no. 9, pp. 1755-1758, sep. 2023, doi: https://doi.org/10.1007/s00256-023-04340-5

[15] G. Eysenbach, “The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers,” JMIR Med. Educ., vol. 9, no. 4, art. no. e46885, mar. 2023, doi: https://doi.org/10.2196/46885

[16] D. De Silva, N. Mills, M. El-Ayoubi, M. Manic, D. Alahakoon, "ChatGPT and Generative AI Guidelines for Addressing Academic Integrity and Augmenting Pre-Existing Chatbots," in 2023 IEEE International Conference on Industrial Technology (ICIT), Orlando, FL, USA, 2023, pp. 1-6. doi: **https://doi.org/10.1109/ICIT58465.2023.10143123**

[17] M. J. Bommarito, D. M. Katz, "GPT Takes the Bar Exam," SSRN Electron. J., pp. 1-13, 2023, doi: **https://dx.doi.org/10.2139/ssrn.4314839**

[18] J. Bommarito, M. J. Bommarito, J. Katz, D. M. Katz, "Gpt as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities," SSRN Electron. J., 2023, doi: **https://dx.doi.org/10.2139/ssrn.4322372**

[19] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash, "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment," JMIR Med. Educ., vol. 9, art. no. e45312, 2023, doi: **https://doi.org/10.2196/45312**

[20] T. H. Kung, M. Cheathem, A. Medenilla, C. Sillos, et al., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," PLOS Digit. Health, vol. 2, no. 2, art. no. e0000198, 2023, doi: **https://doi.org/10.1371/journal.pdig.0000198**

[21] J. Miao, C. Thongprayoon, W. Cheungpasitporn, "Assessing the Accuracy of ChatGPT on Core Questions in Glomerular Disease," Kidney Int. Rep., vol. 8, no. 8, pp. 1657-1659, 2023, doi: **https://doi.org/10.1016/j.ekir.2023.05.014**

[22] B. H. H. Cheung, G. K. K. Lau, G. T. C. Wong, E. Y. P. Lee, et al., "ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom)," PLoS One, vol. 18, no. 8, art. no. e0290691, ago. 2023, doi: **https://doi.org/10.1371/journal.pone.0290691**

[23] A. K. Khilnani, "Potential of Large Language Model (ChatGPT) in Constructing Multiple Choice Questions," GAIMS J. Med. Sci., vol. 3, no. 2, pp. 1-3, 2023. [En línea]. Disponible en: **https://gjms.gaims.ac.in/index.php/gjms/article/view/71**

[24] S. Elkins, E. Kochmar, I. Serban, J. C. K. Cheung, "How Useful Are Educational Questions Generated by Large Language Models?," in Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. AIED 2023. Communications in Computer and Information Science, vol 1831, Tokyo, Japón, 2023, pp. 536-542, doi: **https://doi.org/10.1007/978-3-031-36336-8 83**

[25] X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, H. Sauzéon, P.-Y. Oudeyer, "Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation," in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canadá, 2023, pp. 12952-12965, doi: **https://doi.org/10.18653/v1/2023.findings-acl.820**

[26] C. Xiao, S. X. Xu, K. Zhang, Y. Wang, L. Xia, "Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications," in Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Toronto, Canadá, 2023, pp. 610-625, doi: **https://doi.org/10.18653/v1/2023.bea-1.52**

[27] S. Sedaghat, "Early applications of ChatGPT in medical practice, education and research," Clin. Med., vol. 23, no. 3, pp. 278-279, may. 2023, doi: **https://doi.org/10.7861/clinmed.2023-0078**

[28] R. Dijkstra, Z. Genç, S. Kayal, and J. Kamps, "Reading Comprehension Quiz Generation using Generative Pre-trained Transformers," in 4th International Workshop on Intelligent Textbooks, iTextbooks 2022, Durham, Reino Unido, 2022. [En línea]. Disponible en: **https://hdl.handle.net/11245.1/a1109043-92d4-4c63-be33-6e238780d3b7**

[29] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, et al., A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Nueva York: Pearson Education, 2001.

[30] Jamovi. (2023). [En línea]. Disponible en: **https://www.jamovi.org**