Lecture Notes in Networks and Systems 1040

Marcelo Zambrano Vizuete ·
Miguel Botto-Tobar ·
Sonia Casillas · Carina Gonzalez ·
Carlos Sánchez · Gabriel Gomes ·
Benjamin Durakovic *Editors*

Innovation and Research – Smart Technologies & Systems

Proceedings of the CI3 2023, Volume 1

Springer

Lecture Notes in Networks and Systems 1040

Series Editor

Janusz Kacprzyk, *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Advisory Editors

Fernando Gomide, *Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil*
Okyay Kaynak, *Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye*
Derong Liu, *Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA*
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Witold Pedrycz, *Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada*
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
Marios M. Polycarpou, *Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus*
Imre J. Rudas, *Óbuda University, Budapest, Hungary*
Jun Wang, *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Marcelo Zambrano Vizuete ·
Miguel Botto-Tobar · Sonia Casillas ·
Carina Gonzalez · Carlos Sánchez ·
Gabriel Gomes · Benjamin Durakovic
Editors

Innovation and Research – Smart Technologies & Systems

Proceedings of the CI3 2023, Volume 1

Springer

Editors
Marcelo Zambrano Vizuete
Instituto Tecnológico Superior Rumiñah
Sangolquí, Ecuador

Miguel Botto-Tobar
Eindhoven University of Technology
Eindhoven, The Netherlands

Sonia Casillas
universidad de salamanca
Salamanca, Spain

Carina Gonzalez
Universidad de La Laguna
Tenerife, Spain

Carlos Sánchez
Universidad de Zaragoza
Zaragoza, Spain

Gabriel Gomes
Universidad Estatal de Campinas
Campinas, Brazil

Benjamin Durakovic
International University of Sarajevo
Sarajevo, Bosnia and Herzegovina

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-63433-8 ISBN 978-3-031-63434-5 (eBook)
https://doi.org/10.1007/978-3-031-63434-5

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The 4th edition of the International Research and Innovation Congress – Smart Technologies and Systems, CI3 2023, took place from August 30 to September 1, 2023, at the facilities of the Instituto Tecnológico Universitario Rumiñahui, located in the city of SangolquÍ, Pichincha, Ecuador.

The conference was organized by the Red de Investigación, Innovación y Transferencia de Tecnología RIT2, made up of the most relevant university institutes in Ecuador, among which are ITCA, BOLIVARIANO, ARGOS, VIDA NUEVA, ESPÍRITU SANTO, SUDAMERICANO, ISMAC, SAN ISIDRO, ARTES GRÁFICAS, ORIENTE, HUMANE, SUCRE, CENTRAL TÉCNICO, POLICÍA NACIONAL and RUMIÑAHUI.

Additionally, the event is sponsored by the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación SENESCYT, Labortorio de Comunicación Visual de la Universidad Estatal de Campinas—Brazil, Universidad Ana G. Méndez—Puerto Rico, Centro de Investigaciones Psicopedagógicas y Sociológicas—Cuba, Instituto Superior de Diseño de la Universidad de La Habana—Cuba, GDEON and the Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia—CEDIA.

The main objective of CI3 2023 is to generate a space for dissemination and collaboration, where academia, industry and government can share their ideas, experiences and results of their projects and research.

"Research as a pillar of higher education and business improvement" is the motto of the Conference and suggests how research, innovation and academia must coincide with the productive sector to leverage social and economic development.

CI3 2023 had 145 papers submitted, of which 52 were accepted for publication and presentation. To guarantee the quality of the publications, the event has a staff of more than 70 experts, from different countries such as Spain, Argentina, Chile, Mexico, Peru, Brazil, Ecuador, among others, who carry out an exhaustive review of each proposal sent.

Likewise, during the event a series of keynote conferences were held, given by both national and international experts, allowing attendees to get in touch with the latest trends and technological advances around the world. Keynote speakers included: Ph.D. Carina González, University of La Laguna, Spain; Ph.D. Gabriel Gómez, State University of Campinas, Brazil; Ph.D. Carlos Sanchez, University of Zaragoza, Spain; Ph.D. Juan Minango, Instituto Tecnológico Superior Rumiñahui; Dr. Iván Cherrez, Universidad Espíritu Santo, Ecuador; and Ph.D. Nela Paustizaca, Escuela Politécnica del Litoral, Ecuador.

The content of this proceeding is related to the following topics:

- Smart Cities
- Innovation and Development
- Applied Technologies
- Economics and Management
- ICT for Educations.

Organization

General Chairs

Marcelo Zambrano Vizuete	Instituto Tecnológico Universitario Rumiñahui, Ecuador

Organizing Committee

Marcelo Zambrano Vizuete	Instituto Tecnológico Universitario Rumiñahui, Ecuador
Miguel Botto-Tobar	Eindhoven University of Technology, The Netherlands
Sonia Casillas	Universidad de Salamanca, Spain
Carina Gonzales	Universidad de La Laguna, Spain
Carlos Sánchez	Universidad de Zaragoza, Spain
Gabriel Gomes	Universidad Estatal de Campinas, Brazil
Benjamin Durakovic	International University of Sarajevo, Bosnia and Herzegovina

Steering Committee

Marcelo Zambrano	Instituto Tecnológico Superior Rumiñahui, Ecuador
Miguel Botto	Eindhoven University of Technology, The Netherlands
Karla Ayala	Instituto Tecnológico Superior Rumiñahui, Ecuador
Xavier Duque	Instituto Tecnológico Superior Rumiñahui, Ecuador
Ernesto Huerta	Instituto Tecnológico Superior Rumiñahui, Ecuador
Carmita Suárez	Instituto Tecnológico Superior Rumiñahui, Ecuador
Maritza Salazar	Instituto Superior Tecnológico Espíritu Santo, Ecuador
Wilfrido Robalino	Instituto Superior Tecnológico Vida Nueva, Ecuador

Roberto Tolozano	Instituto Superior Tecnológico Bolivariano, Ecuador
Carlos Pérez	Instituto Tecnológico Sudamericano, Ecuador
Alicia Soto	Instituto Superior Tecnológico ITCA, Ecuador
Lorena Ávila	Instituto Superior Tecnológico Policía Nacional, Ecuador
Denis Calvache	Instituto Superior Tecnológico ISMAC, Ecuador
Efraín Paredes	Instituto Superior Tecnológico IGAD, Ecuador
Paola Vasconez	Instituto Superior Tecnológico Oriente, Ecuador
Sandra El Khori	Instituto Superior Tecnológico San Isidro, Ecuador
Santiago Illescas	Instituto Superior Tecnológico Sucre, Ecuador
Jorge Calderón	Instituto Superior Tecnológico Argos, Ecuador
Trotsky Corella	Instituto Superior Tecnológico Central Técnico, Ecuador
Javier García	Instituto Superior Tecnológico Humane, Ecuador
Sandra Salazar	Instituto Superior Tecnológico American College, Ecuador
David Flores	Instituto Superior Tecnológico Cordillera, Ecuador
Marcelo Aguilera	Instituto Superior Tecnológico del Azuay, Ecuador

Program Committee

Carina González	Universidad de La Laguna, Spain
Carlos Sánchez	Universidad de Zaragoza, Spain
Sergey Balandin	Universidad de Helsinki, Finland
Silva Ganesh Malla	SKR Engineering College, India
Alberto Ochoa	Universidad Autónoma Ciudad de Juárez, Mexico
Antonio Orizaga	Universidad de Guadalajara, Mexico
Rocío Maciel	Universidad de Guadalajara, Mexico
Víctor Larios	Universidad de Guadalajara, Mexico
Yorberth Montes	Universidad de Zulia, Venezuela
Juan Minango	Universidad Estatal de Campinas, Brazil
Francisco Pérez	Universidad Politécnica de Valencia, Spain
Víctor Garrido	Universidad Politécnica de Valencia, Spain
Alberto García	Universidad Politécnica de Valencia, Spain
Javier Hingant	Universidad Politécnica de Valencia, Spain
Javier Prado	Universidad Técnica Federico Santa María, Chile
Marco Heredia	Universidad Politécnica de Madrid, Spain

Diego Paredes	Universidad de Zaragoza, Spain
Angela Cadena	Universitat de Valencia, Spain
Sonia Casillas	Universidad de Salamanca, Spain
Marco Cabezas	Universidad de Salamanca, Spain
Francesc Wilhelmi	Centre Tecnològic de Telecomunicacions de Catalunya, Spain
Santiago Vidal	Universidad Nacional Del Centro de la Provincia de Buenos Aires, Argentina
Adailton Antônio Galiza Nunes	Universidad Estatal de Campinas, Brazil
Leandro Bezerra De Lima	Universidad Federal de Mato Grosso Do Sul, Brazil
Andrea Carolina Flores	Embraer Defensa y Seguridad, Brazil
Elmer Levano Huamacto	Universidad Federal Do Parana, Brazil
Ari Lazzarotti	Universidad Federal de Goiás, Brazil
Ellen Synthia Fernandes de Oliveira	Universidad Federal de Goiás, Brazil
Fernanda Cruvinel Pimentel	Universidad Federal de Goiás, Brazil
Yuzo Lanoi	Universidad Estatal de Campinas, Brazil
Gabriel Gómez	Universidad Estatal de Campinas, Brazil
Pablo Minango	Universidad Estatal de Campinas, Brazil
Benjamin Durakovic	International University of Sarajevo, Bosnia and Herzegovina
Jhonny Barrera	Universidad Nacional de La Plata, Argentina
Daniel Ripalda	Universidad Nacional de La Plata, Argentina
Alonso Estrada Cuzcano	Universidad Nacional Mayor de San Marcos, Peru
Marcelo Zambrano	Instituto Tecnológico Superior Rumiñahui, Ecuador
Ana Zambrano	Escuela Politécnica Nacional, Ecuador
Angel Jaramillo	Universidad de las Américas, Ecuador
Ángela Cadena	Universidad de Guayaquil, Ecuador
Oscar Leon	Universidad de Guayaquil, Ecuador
Wladimir Paredes	Consejo de Aseguramiento de la Calidad de la Educación Superior, Ecuador
Miguel Angel Zúñiga	Universidad de Cuenca, Ecuador
Alex Santamaría	Universidad Laica Eloy Alfaro, Ecuador
Fabián Sáenz	Universidad de las Fuerzas Armadas ESPE, Ecuador
Darwin Aguilar	Universidad de las Fuerzas Armadas ESPE, Ecuador
Sergio Montes	Universidad de Las Fuerzas Armadas Espe, Ecuador
Jhenny Cayambe	Pontificia Universidad Católica del Ecuador Sede Ibarra, Ecuador

Santiago Ushiña	Likatelec, Ecuador
Raisa Bernal	Instituto Tecnológico Superior Rumiñahui, Ecuador
Aníbal Altamirano	Universidad de las Fuerzas Armadas ESPE, Ecuador
Moisés Toapanta	Universidad Católica Santiago de Guayaquil, Ecuador
Ricardo Rosero	Instituto Tecnológico Superior Rumiñahui, Ecuador
Danny De La Cruz	Universidad de las Fuerzas Armadas ESPE, Ecuador
Johana Tobar Quevedo	Universidad de las Fuerzas Armadas ESPE, Ecuador
Jhony Caucha	Universidad Nacional de Tumbes, Peru
Mariana Lima Bandeira	Universidad Andina Simón Bolívar, Ecuador
Norma Molina	Universidad Israel, Ecuador
Elfio Pérez	Universidad Indoamérica, Ecuador
Cristian Tasiguano	Instituto Superior San Antonio, Ecuador
Lizbeth Suarez Morales	Instituto Superior Vida Nueva, Ecuador
Carlos Ruiz	Instituto Superior Vida Nueva, Ecuador
Wilfrido Robalino	Instituto Superior Vida Nueva, Ecuador
William Venegas	Escuela Politécnica Nacional, Ecuador
César Ayabaca	Escuela Politécnica Nacional, Ecuador
Dario Morales	Instituto Superior Vida Nueva, Ecuador
Pamela Villareal	Instituto Superior Vida Nueva, Ecuador
Darwin Tituaña	Instituto Superior Vida Nueva, Ecuador
Ismenia Araujo	Instituto Superior José Chiriboga Grijalva, Ecuador
Luz Marina Rodríguez	Instituto Superior José Chiriboga Grijalva, Ecuador
Luis Alzate	Instituto Superior Bolivariano, Ecuador
Noemi Delgado	Instituto Superior Bolivariano, Ecuador
Marítza Salazar	Instituto Superior Espíritu Santo, Ecuador
Martha Fernández	Instituto Superior Espíritu Santo, Ecuador
Néstor Xavier Maya Izurieta	Instituto Superior Central Técnico, Ecuador
Michael Enrique Carrión Garzón	Instituto Superior Central Técnico, Ecuador
Nela Pastuizaca	Escuela Politécnica del Litoral, Ecuador
Marcelo Flores	Universidad Politécnica Salesiana, Ecuador
Diego Vizuete	Instituto Superior Sucre, Ecuador
Edgar Maya	Universidad Técnica del Norte, Ecuador
Mauricio Domínguez	Universidad Técnica del Norte, Ecuador
Fabián Cuzme	Universidad Técnica del Norte, Ecuador

Contents

A Voice-Based Emotion Recognition System Using Deep Learning Techniques

Carlos Guerrón Pantoja[1], Edgar Maya-Olalla[1](✉),
Hernán M. Domínguez-Limaico[1], Marcelo Zambrano[1,2],
Carlos Vásquez Ayala[1], and Marco Gordillo Pasquel[1]

[1] Universidad Técnica del Norte, Avenida 17 de Julio 5-21 y General María Córdoba, 100105 Ibarra, Ecuador
eamaya@utn.edu.ec

[2] Instituto Superior Tecnológico Rumiñahui, Avenida Atahualpa 1701 y 8 de Febrero, 171103 Sangolquí, Ecuador

Abstract. The project aims to create an emotion recognition system based on voice using deep learning techniques. The system is based on supervised learning with artificial neural networks, enabling it to accurately predict emotions. The system's potential usage in detecting depression pathologies in the psychological area gives rise to its development. The system is designed using the KDD (Knowledge Discovery in Database) methodology and utilizes an existing database containing audio with various emotions. These audios are subjected to Multilevel Wavelet transform, decomposing the original signal into sub-signals with specific characteristics for each audio to form a training data set that is subsequently normalized, followed by the generation of the LSTM neural network architecture. Performance tests are eventually conducted on patients with depressive pathology, involving the application of the "Beck test", which indicates the severity of depression experienced by the patient. As a result, the individual reads a text that is recorded, followed by the process of feature extraction and emotion recognition performed with the pre-trained neural network. The outcome indicates that 50% of the patients exhibit severe depression, while the remainder displays milder symptoms, which is supported by the emotions detected by the system alongside the administered test.

Keywords: Emotion Recognition · Wavelet Transform · deep learning · emotion recognition

1 Introduction

Emotion recognition can determine human attitude through speech-based emotional cues. Therefore, detecting human emotions is crucial for personal healthcare and mental well-being [1]. Emotions are an innate aspect of human beings, and they reflect a range of moods including sadness, anger, happiness, and neutrality. However, issues may arise due to emotional imbalance, including depression which significantly impacts a person's usual functioning, well-being, and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
M. Z. Vizuete et al. (Eds.): CI3 2023, LNNS 1040, pp. 155–172, 2024.
https://doi.org/10.1007/978-3-031-63434-5_12

quality of life [2]. Artificial intelligence (AI) is currently a popular trend, with many research projects utilizing this technique to gain new insights and strategies that enhance people's lives. The research aims to study how people's emotions are expressed through their speech patterns and to create a machine-learning reference criterion that is inspired by the biological neural networks of the human brain. These elements are composed in a way that mimics the biological neurons in their most common functions. These elements are organized in a manner that resembles that of the human brain [3].

To conduct the research, it is important to discuss various concepts that aid in the development of the system. This starts with the voice, which is a sound produced by the vibration of vocal cords in living organisms and has qualities such as timbre, intensity, and quality [4]. The process of extracting acoustic parameters of vocal utterances is based on the idea that variations in speech are caused by different states of excitation or valence in the speaker. These variations can be estimated by different wave parameters [5]. The approach for Speech Emotion Recognition (SER) incorporates two main phases: feature extraction and feature classification [6]. One method of feature extraction is the multi-level modeling method. This method uses the Continuous Wavelet Transform (CWT) to model hierarchical prosodic features, such as an initial signal and energy contour [7].

A neural network is a computational model that emulates the behavior of the human brain. It consists of interconnected artificial neurons, represented by nodes or vertices, and connections, represented by edges. In a neural network, the fundamental element is the artificial neuron. It emulates its natural counterpart found in the brain [8].

The main focus of this research is to create an emotion recognition system using deep learning techniques. This system will detect emotions in patients who are attending consultations. The goal is to implement this system in patient care to provide better support to psychologists and offer improved diagnoses to the patient based on professional judgment.

The cascade model will facilitate the selection of system requirements during the system development process, which comprises several successive phases. For instance, the analysis phase addresses the concepts covered in the project. Another phase of the project is the design and specifications where the programming, training based on the Knowledge Discovery in Databases (KDD) methodology, implementation, functional testing, and finally a verification of the results obtained will be made.

Finally, we obtain the results of the tests conducted with the system on various individuals. We classify the effectiveness of measuring emotions and determine the percentage of emotions measured. We identify the emotion(s) with the lowest percentage of accuracy, precision, and sensitivity, enabling the system administrator to consider these values and use this tool effectively.

2 Development

The creation process of the speech emotion recognition system includes multiple stages such as emotion audio acquisition, signal preprocessing, feature extraction, training data preparation, and neural network training. The entire process is aligned with the Knowledge Discovery in Databases (KDD) methodology, represented in Fig. 1

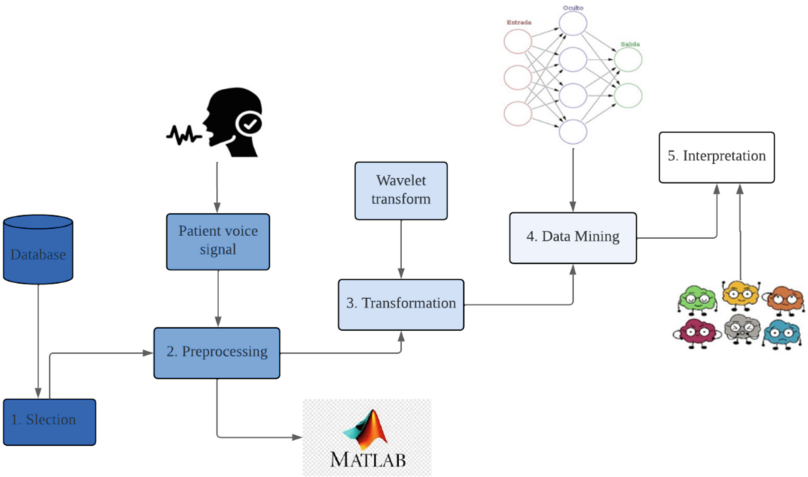


Fig. 1. Generalized KDD process.

2.1 Obtaining Audio of Emotions

Classical data mining tasks require two types of data: a training dataset that trains the algorithm on the problem-specific data and a test dataset that evaluates the quality of the algorithm. Thus, it was necessary to generate at least one audio recording for each human emotion during the data collection phase. However, due to the lack of Spanish audio data, a Mexican audio dataset known as the Mexican Emotion Speech Database (MESD) is utilized.

MESD provides single-word utterances for the affective prosodies of anger, disgust, fear, happiness, neutrality, and sadness with Mexican cultural conformation. The database was retrieved from the Kaggle repository, a scientific community focused on machine learning. The database is licensed for free use, however, proper credit should be given to its author, Saurabh Shahane.

The database consists of single-word utterances expressing affective prosodies of anger, disgust, fear, happiness, neutrality, and sadness. The affective prosodies are selected according to Mexican cultural adaptation. The Multilingual Everyday Speech Dataset (MESD) was recorded by adult actors and non-professional children. Three female, two male, and six child voices are included in the dataset. In this study, a database of 864 voice recordings is evaluated (see Table 1).

Table 1. Database prosodies

Database	MESD
Interpreter	Men, Women, Children
Size	864 recordings
Prosodias	Anger, disgust, fear, happiness, neutral, sadness
License	Free
Sampling frequency	48 kHz

The audio signals corresponding to different emotions are analyzed. Each emotion has specific characteristics that can be reflected by analyzing the corresponding graph of the audio signal. The important characteristics of each emotion can be highlighted by analyzing the obtained raw data. Accordingly, the audio of "Up", which was expressed in the emotion of Anger, was selected. Figure 2 displays the signal corresponding to anger. The first part of the figure shows the signal over time, while the second part shows its frequency spectrum. In Fig. 3, the amplitude of the signal starts with short peaks that are segmented over some samples. This is followed by a gradual increase in amplitude, culminating in very high peaks in the positive and negative axes. This signal is compared to the way anger is expressed vocally, as it begins with a soft tone that ends in a high tone producing a signal as seen in the graph. The signal's energy reaches up to -20dB, as illustrated in the spectrum graph.

2.2 Signal Preprocessing

Preprocessing is a procedure performed on the audio files before analysis, which may include resizing, changing the sample rate, enhancing quality, and extracting features.

It is important to note that the audio files in the database have a bandwidth of 48000 Hz, but their bandwidth is reduced to 16000 Hz after resizing. Thus, all the audio files in the database have a bandwidth of 16000 Hz.

For the audio files obtained in the previous section, the sampling rate of each element is reviewed. This is done to ensure that all the audio files have standardized parameters. To achieve this, the resampling function in MATLAB is utilized. This function alters the sample rate of an audio file to a desired rate. This process is demonstrated in Fig. 4.

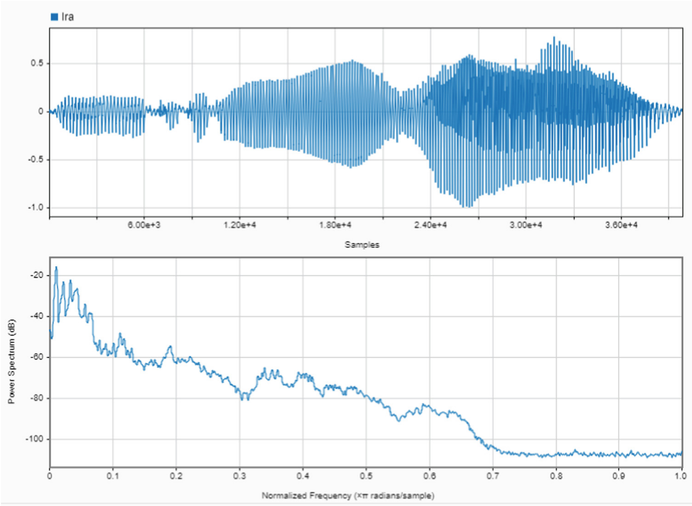


Fig. 2. Plot of anger signal versus time.

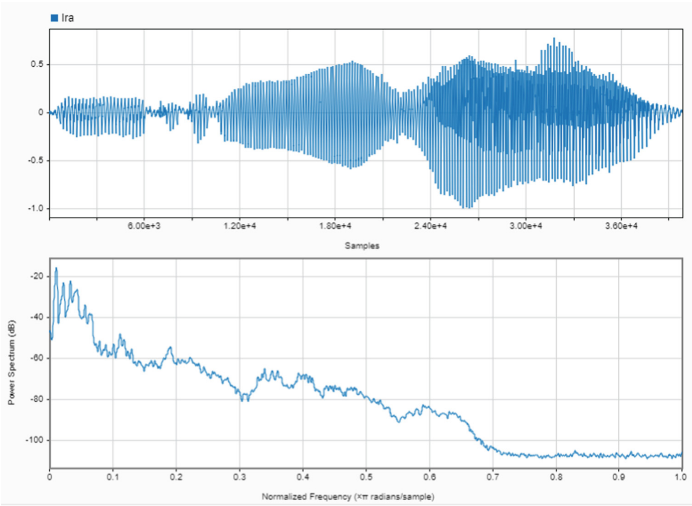


Fig. 3. Plot of the spectrum of the anger signal.

The audio to be processed is read first. Then, it is approximated by rational fractions using the rat function. Next, the sample rate is changed from the current one to the desired one by applying the resample function. Finally, the resulting audio is played back and saved in the desired direction. This process is repeated for each audio file in the folder.

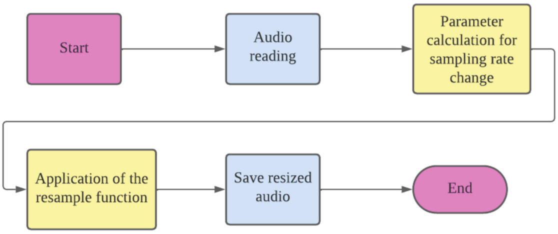


Fig. 4. Process for the audio sampling rate change.

After changing the sample rate, a data store is created with the addresses and names of each audio file. This is done to control variable generation, especially concerning their size in megabytes. The process shown in Fig. 5 is used for this purpose.

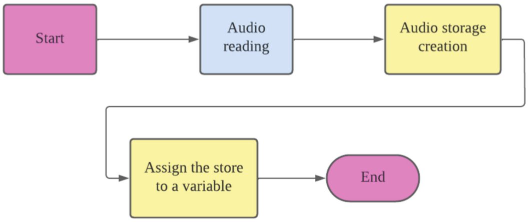


Fig. 5. Process for creating the data warehouse.

2.3 Feature Extraction

This process involves identifying and extracting unique features numerically from emotional audio recordings. The extraction of significant features is performed using wavelet transform in Matlab. To do so, one must follow a logical sequence of five steps, involving reading audio, calculating multilevel wavelet decomposition, quantifying signal entropy, and associating the features with corresponding emotions.

The main concept is to decompose the signal into a sequence of subspaces using chosen wavelet functions (Symmlet and Daubechies). This decomposition is usually accomplished by shifting (right or left) and scaling (amplifying or attenuating) the chosen wavelet function. By doing so, the signal is projected into the area represented by the particular scaling and translation.

In the issue of emotion recognition, the signal energy of the audio signal is recognized as one of the most crucial characteristics. A straightforward approach to extract features would be to employ the signal energy that is available in each subspace or node. Once you acquire each node's feature value, assemble them side by side, and that constitutes your feature vector (refer to Fig. 6).

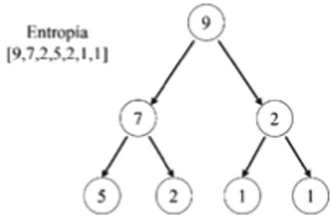


Fig. 6. Example of feature extraction for three levels of decomposition.

To illustrate the process described above, we proceed to calculate the wavelet transform for each considered emotion type. The anger signal is subjected to a wavelet decomposition of the Daubechies-4 family with seven levels of decomposition. As a consequence of using seven levels of decomposition, 255 features are obtained ($2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 + 2^7 = 255$), as illustrated in Fig. 7.

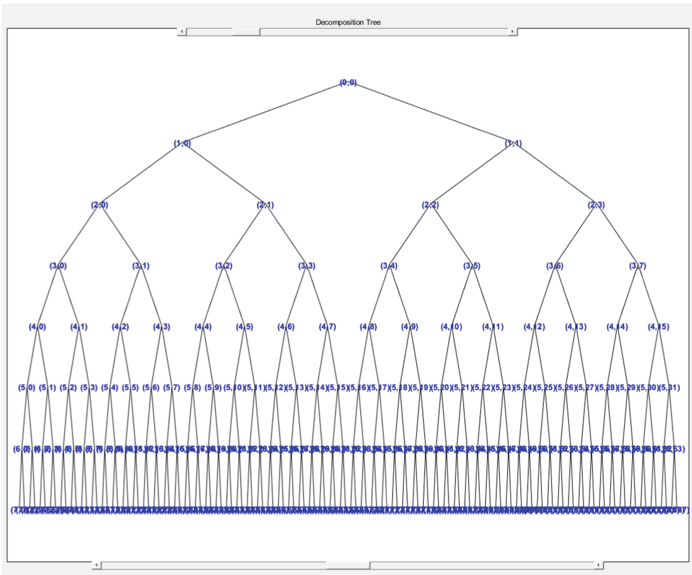


Fig. 7. 7-level wavelet decomposition tree for the emotion of anger.

The first two elements of the final decomposition level are presented, which correspond to the approximation coefficients displayed in Fig. 8. The approximation coefficient exhibits a much clearer signal when compared to the original audio signal, where the amplitude varies constantly.

Emotion audio labeling is the process of assigning emotions to audio. After obtaining the characteristics of the emotion audio, the data is associated with the relevant emotion through a process known as audio labeling. To achieve this, a binary logic, shown in Table 2, is used. Each vector of extracted characteristics is associated with a corresponding binary vector, as shown in Table 2, to label the audio. This labeling process is performed for all available audio in the database.

Table 2. The logic for emotion detection

Ira	Disgust	Fear	Happiness	Neutral	Sadness
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1

2.4 Preparation of Training Data

After processing emotional audio and identifying and extracting signal characteristics, three matrices are created. To accomplish this, the matrix resulting from identifying and extracting characteristics is divided into three groups based on the percentages 80%, 10%, and 10%, which correspond to the training, validation, and testing data, respectively.

The normalization process adjusts the data set to scale values that differ from the original ones. In the obtained data set, the normalization process is performed. This process establishes a certain range for the data set so that it enters the neural network and is trained in a manner that helps the network comprehend the incoming data. The input dataset contains both maximum and minimum values. The data undergoes normalization, which establishes a range of [0,1]. The normalization process takes the negative data as the minimum and the maximum value as the highest. Finally, with the help of Eq. (1), the data can be normalized within the desired range.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where:

- X' represents the normalized value of each datum in the dataset.

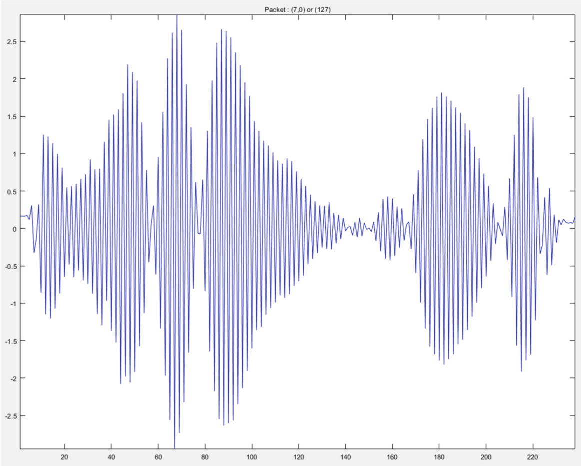


Fig. 8. Approximation coefficient, level 7 associated with the emotion of anger.

In contrast, the detail coefficient indicates that the signal is predominantly present during speech pauses (refer to Fig. 9).

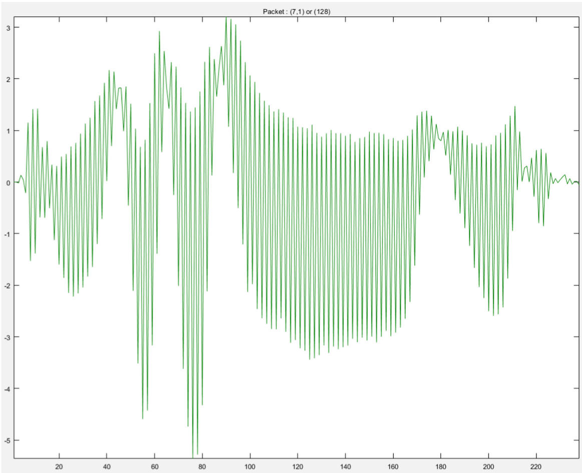


Fig. 9. Detail coefficient, level 7 associated with the emotion of anger.

- The value we wish to normalize is represented by X.
- The value of the minimum data found in the data set is denoted by $X_{\min}$.
- The maximum value found in the data set is represented by $X_{\max}$.

The prepared data set has been normalized, as shown in Fig. 10.

carNorma

868x255 double

	3	244	245	246	247	248	249	250	251	252	253	254	255
856	).2143	0.2117	0.2118	0.2149	0.2132	0.2169	0.2180	0.2168	0.2171	0.2114	0.2126	0.2176	0.2154
857	).2262	0.2241	0.2277	0.2239	0.2264	0.2272	0.2307	0.2316	0.2328	0.2251	0.2288	0.2338	0.2321
858	).2221	0.2240	0.2254	0.2197	0.2229	0.2295	0.2319	0.2311	0.2319	0.2270	0.2288	0.2318	0.2304
859	).1875	0.1845	0.1859	0.1861	0.1872	0.1896	0.1931	0.1942	0.1936	0.1862	0.1893	0.1969	0.1931
860	).1820	0.1870	0.1841	0.1840	0.1847	0.1767	0.1804	0.1816	0.1802	0.1861	0.1839	0.1881	0.1843
861	).2228	0.2219	0.2229	0.2229	0.2217	0.2275	0.2299	0.2305	0.2295	0.2232	0.2252	0.2319	0.2282
862	).2870	0.2895	0.2883	0.2851	0.2863	0.2947	0.2967	0.2967	0.2961	0.2905	0.2912	0.2959	0.2933
863	).3837	0.3755	0.3814	0.3776	0.3806	0.3818	0.3864	0.3840	0.3901	0.3798	0.3845	0.3831	0.3880
864	).0720	0.0712	0.0725	0.0716	0.0721	0.0771	0.0791	0.0806	0.0794	0.0732	0.0758	0.0825	0.0785
865	).2429	0.2444	0.2440	0.2434	0.2434	0.2478	0.2508	0.2512	0.2501	0.2465	0.2476	0.2542	0.2495
866	).2374	0.2370	0.2361	0.2366	0.2347	0.2404	0.2428	0.2436	0.2421	0.2394	0.2407	0.2455	0.2421
867	).2784	0.2788	0.2809	0.2754	0.2781	0.2894	0.2922	0.2918	0.2919	0.2806	0.2831	0.2893	0.2865
868	).2085	0.2020	0.2065	0.2062	0.2084	0.1988	0.2072	0.2079	0.2106	0.2072	0.2115	0.2114	0.2143
869													

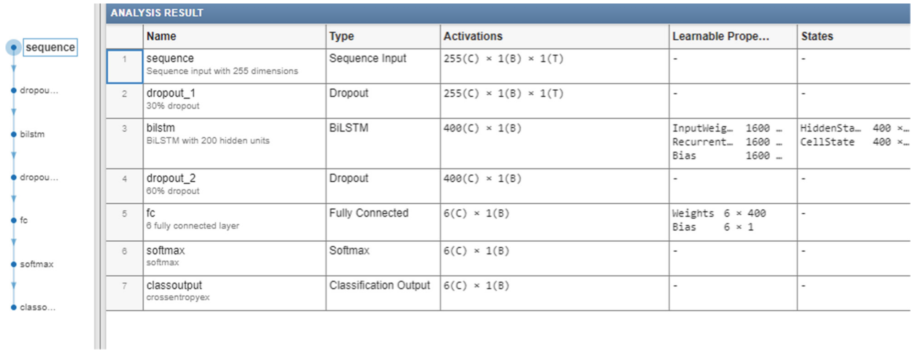
Fig. 10. Normalized database with range [0,1].

2.5 Neural Network Training

This is a three-step process: building the LSTM model, parameterizing, and training the neural network. The architecture of the LSTM model is created using MATLAB's Deep Network Designer application, in which two models are built and tested to determine the most optimal model for emotion prediction. The general architecture adopted for the neural network starts with 255 input layer neurons and 6 output layer neurons. In Fig. 11 shows the LSTM architecture used to train the network.

Building the LSTM Model. The neural network models presented in this study are composed of multiple layers, including the input layer, intermediate layers, and output layers. An aspect of particular significance is the activation of the input layer with a data set characterized by dimensions of $255(C) \times 1(B) \times 1(T)$. These dimensions correspond to the features derived from the wavelet transform, signifying that the application of dissimilar multilevel wavelet transforms on an audio input could potentially lead to inaccurate classification outcomes.

Parameterization and Training Process. Upon defining the neural network model, the subsequent phase involves the meticulous configuration of training parameters. While there is variability in the selection of the algorithm and learning rate, certain factors such as the number of epochs and the minimum batch size remain consistent throughout this phase.



ANALYSIS RESULT

	Name	Type	Activations	Learnable Prope...	States
1	sequence Sequence input with 255 dimensions	Sequence Input	255(C) × 1(B) × 1(T)	-	-
2	dropout_1 30% dropout	Dropout	255(C) × 1(B) × 1(T)	-	-
3	bilstm BiLSTM with 200 hidden units	BiLSTM	400(C) × 1(B)	InputWeig... 1600 ... Recurrent... 1600 ... Bias 1600 ...	HiddenSta... 400 ×... CellState 400 ×...
4	dropout_2 60% dropout	Dropout	400(C) × 1(B)	-	-
5	fc 6 fully connected layer	Fully Connected	6(C) × 1(B)	Weights 6 × 400 Bias 6 × 1	-
6	softmax softmax	Softmax	6(C) × 1(B)	-	-
7	classoutput crossentropyex	Classification Output	6(C) × 1(B)	-	-

Fig. 11. BiLSTM neural network architecture with 7 layers for emotion classification.

The chosen algorithm for this study is the Gradient Descent method. This strategic choice stems from its ability to progressively enhance training efficacy over successive epochs, thereby facilitating the minimization of loss or error functions. The determination of the appropriate number of epochs, as depicted in Table 3, commences from the third epoch onward. This decision is informed by meticulous testing, which indicates that optimal training outcomes are achieved within this timeframe. As such, a total of 50 epochs are judiciously employed for conducting the comprehensive tests mandated by this research.

In establishing the optimal minimum number of batches, a comprehensive consideration of the overall data set is undertaken. In this instance, the dataset comprises a total of 868 instances, representing the maximum number of emotion-laden audio samples available. To facilitate a meticulous division of this dataset, an iterative approach is undertaken. This entails partitioning the data into smaller, manageable batches, each tailored to ensure effective training.

The initial batch size selected is 62, which is subsequently reiterated across 14 epochs. Additionally, a batch size of 217 is implemented, repeated four times throughout the epochs. Finally, the total batch size of 868 is engaged and deployed once across the epochs to encompass the entirety of the dataset.

Regarding the learning rate, a judiciously standardized value of 0.005 is employed. This selection is rooted in its intermediate characteristic, rendering it both compatible and recommended for integration with the Gradient Descent algorithm. Notably, a higher learning rate would impede algorithm convergence, while an excessively low value would significantly prolong the algorithm's weight optimization process.

The graphical representation shown in Fig. 12 corresponds to the results obtained from the first test performed, under the parameterization parameters meticulously outlined in Table 3. A noticeable feature within the graph is the manifestation of overtraining within the neural network architecture. This conspicuous overtraining phenomenon can be attributed to the adoption of remarkably small batch sizes that are insufficient for the effective operation of the system under the given parameter set.

Table 3. Parameters for testing with the LSTM neural network model

Parameter	Test 1	Essay 2	Essay 3
Algorithm	Gradient decrease	Gradient decrease	Gradient decrease
Number of epochs	50	50	50
Minimum lot size	62	217	868
Learning rate	0.005	0.005	0.005

The consequences of using such parameter values go beyond overtraining. They culminate in the prolonged period required for the training process, which significantly lengthens the path to achieving the desired results. Furthermore, the confluence of markedly elevated loss values and concurrently diminished accuracy levels substantiates the limited efficacy of the chosen parameterization. Consequently, the pursuit of optimal results necessitates a recalibration of parameter selection, striking a judicious balance between the architectural nuances of the network and the complexity of the task at hand.

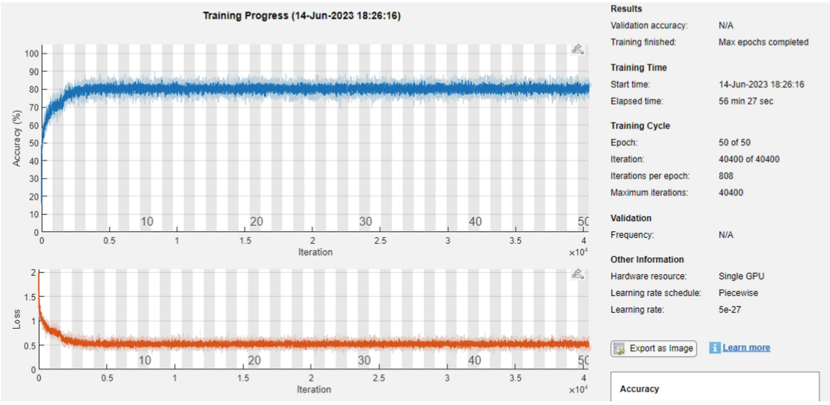


Fig. 12. Accuracy and error result of the first trial applying the LSTM model.

In Fig. 13 gives a visual representation of the training process of the network carried out according to the parameters prescribed in the second test, as described in Table 3. It can be seen that the application of the parameters set for this test leads to consistent results.

In Table 3, the redundancy of the fourth epoch becomes conspicuous, as an observed trend reveals the stability of the accuracy and the parameters from the second epoch onwards. In addition, there is a noticeable reduction in the execution time of the training, depending on the inherent characteristics of the

CPU, which have a significant impact on the temporal dynamics of the training. This correlation confirms the central role of the CPU in influencing the efficiency and duration of the training process.

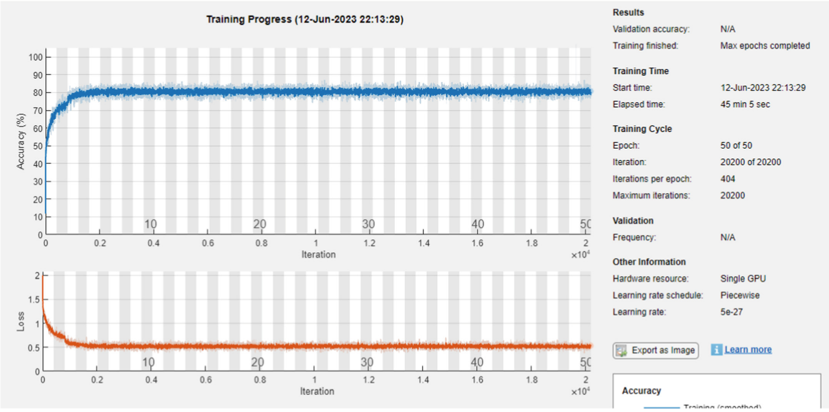


Fig. 13. Accuracy and error result of the second test applying the LSTM model.

In Fig. 14 summarizes the results derived from the final trial, which mirror the results observed in Trial 2. Notably, this final study includes an expanded range of epochs and batch sizes. It is pertinent to emphasize that a consistent pattern similar to that observed in Study 2 is evident, with the stabilization of training occurring from the third epoch onward. The observed stability of training renders further epochs unnecessary, implying that a saturation point has been reached in terms of refining model performance through additional epochs.

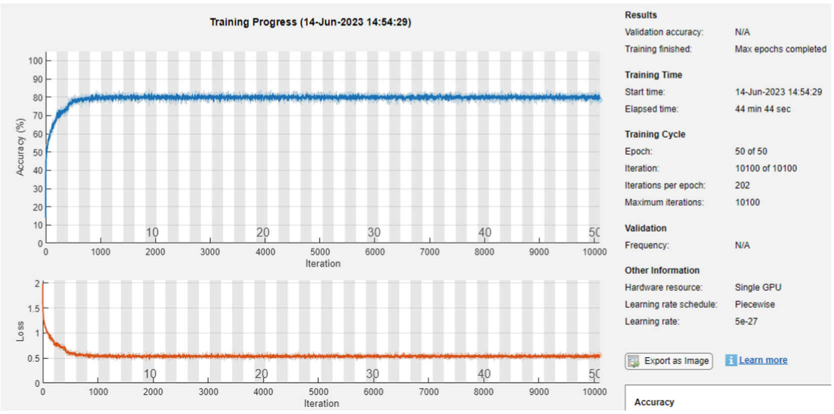


Fig. 14. Accuracy and error result of the third trial applying the LSMT model.

Throughout the neural network training process, careful examination reveals the inherent stability of the learning rate set at 0.005, as opposed to its counterpart at 0.01. As a result, the minimum batch loss exhibits a persistent equilibrium across trials 2 and 3. Significantly, Trial 2 emerges as the optimal training model, characterized by the adoption of a batch size of 217 and the judicious partitioning of the dataset into cohesive segments aligned with the overall size of the dataset. This strategic segmentation avoids the imprudent injection of the entire dataset into the network for training purposes. Conversely, the first batch iteration reveals elevated loss levels, but as the iterations progress, significant loss reduction becomes apparent.

This observed trend bodes well as it anticipates a convergence toward minimal batch losses as the iterations are completed. Finally, it is inferred that after the third epoch, the training reaches a plateau and exhibits consistent performance without further accuracy improvement.

The careful delineation of the parameterization data is comprehensively presented in Table 4. In this tabular format, optimal data parameters are meticulously elucidated, which collectively embody the most favorable configurations for training efforts.

Table 4. Selected training parameters for the neural network

Parameter	Value
Algorithm	Gradient decrease
Number of epochs	3
Minimum lot size	217
Learning rate	0.005

3 Results

This section briefly summarizes the results of the performance of the speech-based emotion recognition system in two different phases. Firstly, the efficiency ratio and the percentage of inaccuracy corresponding to each emotion considered are examined, starting with a comprehensive analysis of the confusion matrix for the trained model. Subsequently, an independent evaluation is carried out with six randomly selected individuals to determine the diagnostic categorization of depression, from mild to severe, for each patient.

The contributions made by this research are multifaceted. First and foremost is the theoretical contribution that underscores the core tenets of this investigation. It encompasses facets ranging from sound representation in speech to sound characterization, the neural networks integral to deep learning and their interconnection with emotion recognition, as well as a juxtaposition with conventional depression detection methods.

Another distinctive contribution is the novel approach to signal characterization. Here, multilevel wavelet decomposition plays a central role. This methodology assigns numerical values to signals depending on the chosen level of decomposition, here at level 7. This complex process results in a comprehensive collection of 255 signal features for each emotional category, meticulously encapsulated in the database.

In addition, this research proposes a linkage between the results derived from the emotion recognition system and the administered depression assessment tests. This synergy complements the psychological domain by providing insights into effective interventions for the treatment of depression.

To critically evaluate the efficiency of the proposed emotion recognition technique, we present the initial model's confusion matrix, as shown in Fig. 15. This analytical lens serves to illuminate the efficiency ratio and the percentage of inaccuracy attributed to each emotion, thus enriching the overall evaluation of the experimental results.

In the first scenario, the accuracy is 73.5%. Although this figure seems relatively modest in the context of the development of neural network systems, it is essential to contextualize this performance. The limited number of audio samples in the database is a prominent factor, limiting the network's ability to reach a superior level of training. In addition, the chosen model and the manner of signal feature extraction inherently influence the accuracy of the results obtained.

The field of emotion recognition is still relatively young. While the earliest investigations date back to 2016, the substantial body of research has proliferated since 2018, highlighting its dynamic and evolving nature. Acknowledging the complex nature of emotion recognition-where variability in emotional expression across individuals poses a challenge literature references are instructive. For example, in [5] reported an effectiveness rate of 56.71% for the neural network used in their research. Similarly, [10] obtained results between 70% and 75%. While the algorithms and databases used are different from the present study, the congruence of the accuracy percentages is noteworthy.

In contrast, the effectiveness of the emotion "anger" is 82.7%, with an associated error rate of 17.3%. Similarly, the emotion "disgust" reaches an effectiveness of 76.1%, paired with a misclassification rate of 23.9%. The emotion "Fear" has a recognition rate of 53.6%, while 46.4% of the cases are misclassified. Similarly, the emotion "happiness" is identified with an efficiency of 50.7% but experiences a misclassification rate of 49.3%. In parallel, the states "neutral" and "sad" manifest themselves as 79.7% and 90.3% correctly classified instances, alongside 20.3% and 9.7% misclassified cases, respectively.

Similarly, the predictive results for the emotion "anger" show an efficiency of 77.6% with an error rate of 22.2%. For the emotion "Fear", the prediction accuracy is 74% with an error rate of 26%. Meanwhile, for the emotion "Disgust", the efficiency rate is 84.4%, coupled with an error rate of 18.6%. "Happiness" has an accuracy of 60%, with an error of 40%. Similarly, "neutral" has an efficiency of 69.2% and an error of 30.8%, while "sadness" has an accuracy of 73.7% and an error of 26.3%.

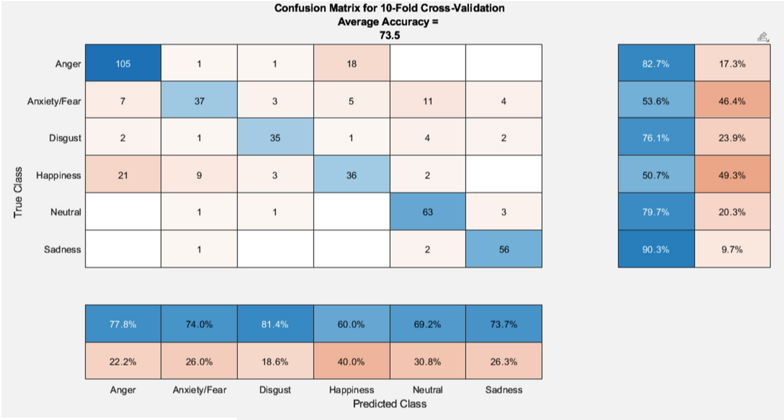


Fig. 15. Confusion matrix of the LSTM model.

The rigorous testing process includes patient consent, audio analysis, feature extraction, and emotion categorization. A systematic 7-step procedure, as illustrated in Fig. 16, is meticulously followed within this testing framework.

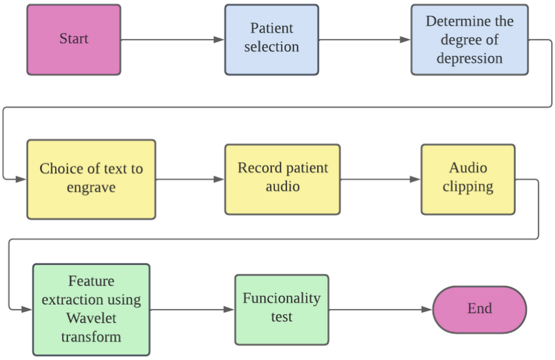


Fig. 16. Confusion matrix of the LSTM model.

The neural network response indicates that the audio input is indicative of the emotion "sadness", as shown in Fig. 17. This result is consistent with the diagnosis of major depression previously confirmed by the Beck test. The schematic shows how the audio labeling was performed. During the prediction phase, the neural network generates values in its six output layers. These values range from 0 to 1, with the value closest to 1 being the expected outcome. To reflect this process, the Simulink diagram uses an adder to effectively round these values to either 1 or 0, thus generating the result graph.

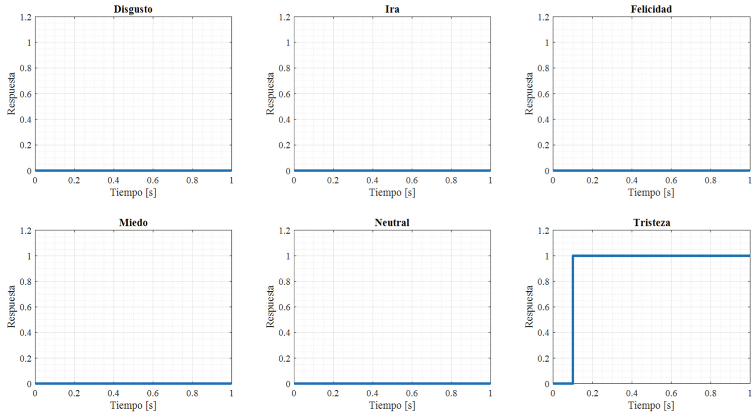


Fig. 17. Response of the emotion recognition system for Individual 1.

Finally, Table 5 is a summary of the results of the emotion detection for each individual. The chart shows the age of the person, the emotion identified by the system, the level of depression determined by the tests administered, and the scores obtained.

Table 5. Summary of the results obtained in the neural network

Individual	Genre	Age	Excitement	Pathology	Test score
Individual 1	Female	18–25	Sadness	Severe Depression	36
Individual 2	Male	18–25	Fear	Mild depression	13
Individual 3	Male	18–25	Fear	Mild depression	16
Individual 4	Female	18–25	Fear	Mild depression	18
Individual 5	Male	18–25	Sadness	Severe Depression	37
Individual 6	Female	18–25	Sadness	Severe Depression	53

4 Conclusions

The voice-based emotion recognition system proves its feasibility by establishing a correlation between the emotions detected by the system and the patient tests administered, thereby facilitating the recognition of the patient's depression level. In this regard, it is estimated that approximately 66.6% of the female participants in the sample are suffering from major depression, while this estimate is 33.3% for the male participants. In addition, the viability of deep learning techniques, particularly neural networks, has been demonstrated.

The Knowledge Discovery in Databases (KDD) architecture was used in the design of the system, including steps such as database acquisition, audio processing, feature extraction, and network training. During neural network training, two models were constructed and analyzed to determine the optimal model for the research. As a result, critical parameters influencing the training of the neural network were identified, including the optimization algorithm (gradient descent), the number of epochs (3), the minimum batch size (217), and the learning rate (0.005).

The overall effectiveness of the neural network was 73.5%. In terms of emotion-specific effectiveness and error rates, the following results were obtained: anger achieved effectiveness of 82.7% with an error rate of 17.3%. Similarly, disgust was identified with 76.1% effectiveness and 23.9% false classifications. Similarly, the network recognized fear with a 53.6% accuracy rate, while 46.4% were misclassified. Happiness was detected with 50.7% effectiveness, accompanied by a 49.3% misclassification rate. Finally, the neutral state and sadness were accurately identified at 79.7% and 90.3%, respectively, with misclassifications at 20.3% and 9.7%.

It is important to note that the effectiveness of the neural network depends significantly on the quality of the extraction of the emotional audio features, as well as the chosen parameterization for the network training. The performance of the network can only be definitively determined through experimentation, using analogous studies from the scientific community as a guide.

References

1. Kaur, G., Pandey, P.S.: Emotion recognition system using IOT and machine learning-a healthcare application. In: Proceeding of The 23rd Conference of Fruct Association, Haryana, 2018, Conference 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016)
2. Orozco, W.N., Baldares, M.J.V.: Addressing depression: intervention in crisis. Dome J. 19–35 (2012)
3. Olabe Basogain, X..: Artificial neuron networks and their applications, Escuela Superior de Ingeniería de Bilbao, p. 79 (2015)
4. Jahangir, R., Teh, Y.W., Hanif, F., Mujtaba, G.: Deep learning approaches for speech emotion recognition: state of the art and research challenges, vol. 80, no. 16. Multimedia Tools and Applications (2021). https://doi.org/10.1007/s11042-020-09874-7
5. Zloteanu, M., et al.: Speech emotion recognition with deep learning. Sensors **11**(2), 1–18 (2022). https://doi.org/10.1016/j.specom.2021.11.006
6. Hansen, L., Zhang, Y.P., Wolf, D., Sechidis, K., Ladegaard, N., Fusaroli, R.: A general- izable speech emotion recognition model reveals depression and remission. Acta Psychiatr. Scand. **145**(2), 186–199 (2022). https://doi.org/10.1111/acps.13388
7. Sisman, B.: Machine Learning for Limited Data Voice Conversion, National University of Singapore, Singapore (2019)
8. Toshinori, M.: Fundamentals of the new Artificial Intelligence, 2a edn. Springer, Cleveland (2008)
9. Ganapathy, A.: Speech emotion recognition using deep learning techniques. ABC J. Adv. Res. **5**(2), 113–122 (2016). https://doi.org/10.18034/abcjar.v5i2.550
10. Sandoval, V.: Emotion recognition by voice, p. 60 (2019)