UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE GUAYAQUIL

CARRERA DE INGENIERÍA DE SISTEMAS

USE OF MACHINE LEARNING ALGORITHMS FOR DIAGNOSIS AND TREATMENT OF PSYCHOLOGICAL PATHOLOGIES

Trabajo de titulación previo a la obtención del
Título de Ingeniero de Sistemas

AUTOR 1: KAREN ESTEFANIA CHARCOPA LAJONES

AUTOR 2: FREDDY JEFFERSON FARIÑO ROSEMBERG

TUTOR: GALO ENRIQUE VALVERDE LANDIVAR

Guayaquil - Ecuador

2022

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Karen Estefania Charcopa Lajones con documento de identificación N° 0931438568 y Freddy Jefferson Fariño Rosemberg con documento de identificación N° 0940729155 manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Guayaquil, 27 de octubre del año 2022

Atentamente,

FREDDY Rosemberg

Karen Estefania Charcopa Lajones

0931438568

Freddy Jefferson Fariño Rosemberg

0940729155

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, Karen Estefania Charcopa Lajones con documento de identificación No.0931438568 y Freddy Jefferson Fariño Rosemberg con documento de identificación No.0940729155 expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autor(es) del Artículo Académico: "Use of Machine learning algorithms for diagnosis and treatment of psychological pathologies", el cual ha sido desarrollado para optar por el título de: Ingeniero de Sistemas, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Guayaquil, 27 de octubre del año 2022

Atentamente,

FREDDY Rosemberg

Karen Estefania Charcopa Lajones

0931438568

Freddy Jefferson Fariño Rosemberg

0940729155

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Galo Enrique Valverde Landívar con documento de identificación N° 0912511532, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: USE OF MACHINE LEARNING ALGORITMS FOR DIAGNOSIS AND TREATMENT OF PSYCHOLOGICAL PATHOLOGIES, realizado por Karen Estefania Charcopa Lajones con documento de identificación No.0931438568 y Freddy Jefferson Fariño Rosemberg con documento de identificación No.0940729155 obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Guayaquil, 27 de octubre del año 2022

Atentamente,

Galo Valverde

Galo Enrique Valverde Landivar

0912511532

DEDICATORIA

A mi madre que siempre ha estado para mí en todo momento, a mi padre que me ha dado su apoyo incondicional, este logro es para ellos por darlo todo por mí.
A mi mejor amiga con la que inicie la carrera y siempre me alentó a que no me rindiera,

Karen Charcopa

A mí madre, por su amor, trabajo y sacrificio en todos estos años, he logrado llegar hasta aquí́ y convertirme en lo que soy.

Freddy Rosemberg

AGRADECIMIENTO

Agradecemos a Dios, por darnos la sabiduría y fuerza en los días que pensábamos desistir, él ilumino cada paso dado en esta etapa de nuestra vida y nos permitió culminar nuestra carrera.

A nuestro Tutor Galo Valverde Landivar que estuvo ahí apoyándonos y proporcionándonos los conocimientos necesarios a lo largo del proceso de investigación y redacción de este artículo, además por su paciencia, comprensión y valiosos consejos.

A todos los docentes que nos impartieron clases a lo largo de nuestra carrera, y nos brindaron sus conocimientos e historias, por sus paciencias y enseñanzas.

RESUMEN

La inteligencia artificial (AI) y el aprendizaje profundo (ML) se han empleado para el entrenamiento y tratamientos de datos masivos, permitiendo la mejora de los sistemas y haciéndolos más inteligentes a la hora de tomar decisiones. El (SER) es un área de investigación de la voz para el reconocimiento de emociones atreves del habla, evalúa la señal de la voz y clasifica las distintas emociones. En los últimos años, los avances tecnológicos del aprendizaje profundo han ayudado al (SER) a detectar y clasificar las emociones con eficacia ya que los métodos de procesamiento de la señal del hablar representan dificultades por la variedad de las frecuencias de las emociones como son la de feliz, enojo, triste, neutral entre otros. En este estudio hemos utilizado una arquitectura de redes convolucionales profundas (DSCNN) para implementar el modelo del (SER). Esta utiliza las redes simples para aprender las características salientes y discriminativas a partir del espectrograma de las señales del habla, generados por medio del conjunto de datos RAVDESS, se consideraron 8 emociones para el análisis y clasificación de emociones, se obtuvo un resultado de predicción del 61%. Posteriormente se propuso una implementación de la (DSCNN) en el área de psicología para determinar los diagnósticos y tratamientos de personas que padecen depresión y ansiedad. Con la ayuda de esta red neuronal profunda en el futuro se podría obtener un diagnóstico eficaz y reducir el tiempo en los tratamientos.

Palabras claves: aprendizaje profundo, inteligencia artificial, reconocimiento de emociones del habla, psicología, espectrogramas del habla.

ABSTRACT

Artificial intelligence (AI) and deep learning (ML) have been used for training and processing of massive data, allowing the improvement of systems, and making them more intelligent when making decisions. Speech Emotion Recognition (SER) is an area of voice research for speech emotion recognition, evaluating the voice signal and classifying different emotions. In recent years, technological advances in deep learning have helped (SER) to detect and classify emotions effectively, as speech signal processing methods are difficult due to the variety of emotion frequencies such as happy, angry, sad, neutral and others. In this study we have used a deep convolutional network architecture (DSCNN) to implement the (SER) model. This uses simple networks to learn salient and discriminative features from the spectrogram of speech signals, generated through the RAVDESS dataset, 8 emotions were considered for the analysis and classification of emotions, a prediction result of 61% was obtained. Subsequently, an implementation of the (DSCNN) was proposed in psychology to determine the diagnoses and treatments of people suffering from depression and anxiety. With the help of thisdeep neural network, an effective diagnosis could be obtained in the future and treatment timecould be reduced.

Key words: deep learning, artificial intelligence, speech emotion recognition, psychology, speech spectrograms.

TABLE OF CONTENTS

1. INTRODUCTION

Information technologies continue to evolve as time goes by, emerging new research on these technologies, such as the internet of things, big data, deep learning, and machine learning. (Ramos, 2014).

With the help of technologies, the field of psychology has taken a great step forward in understanding the pathologies of patients, because its goal is to achieve a combination of parameters that best fits a given problem.

Mental disorders are associated with external factors such as social environment, culture, and life experiences. This makes the study of psychiatry intricate since the great number of variables depend on and interact due to these factors creating a complexity to obtain an adequate treatment.Computational psychiatry provides us with solving complexity in 2 ways, theory-driven computational approaches employ mechanistic models to make explicit hypotheses at multiplelevels of analysis and data-driven machine learning approaches can make predictions from high-dimensional data and are generally agnostic as to the underlying mechanisms (Rutledge et al., 2019).

Mental health problems have been increasing in recent years, and current treatments have not made any change, patients are still years in therapy and taking medication without any improvement in most cases, as mentioned before the large number of variables make it so complex to obtain a specific, effective, and short-lived treatment. (Benhamou, 2020).

The development of applications that simulate a psychotherapist or provide psychological care with machine learning algorithms and deep learning are booming because now people pay more attention to their mental health, and not all people have money or time to go to the psychologist. These applications allow online care and can keep track of their daily states and at the end of the month they can get a report, generating a data and associating the stored data with existing pathologies, they manage to give a diagnosis.

2. LITERATURE REVIEW

2.1 Study of dyadic social interaction using a simulator.

The dyadic interaction is fundamental to determine the behavior of the people involved in the study, and thus be able to develop a program with the relevant algorithms, predicting how the interaction between patient and psychotherapist could be.

There are several types of research that allow the study of the social interaction of people in a dyadic interactive encounter by means of simulated programs. One of these is the creation of an (SBS) called "Eliza", the peculiarity of this model, with respect to other programs, is that it uses data obtained from a real person by means of a survey. (Tardón, 2008).

How the program works

A possible systematization of the interactive sequence, a set of relevant categories in a social situation and a probabilistic decision-making model based on internal variables and the sex of the person with whom the simulator interacts are introduced, the simulator was applied to several groups of participants and finally the data resulting from the interactions were analyzed to assess the validity of the model.

This consisted of a reactive verbal interaction structure, the "Eliza" program, and a series of personalities or scripts, which were a set of keywords and their associated transformations, for a particular type of communication.

Any interaction model must necessarily have two elements defined: what is the interaction sequence and what system of action and emotion categories is used. The results indicate that the key factor in the interaction with simulated beings is immersion and that immersion depends mainly on familiarity and the program interface.

The future application can be open to many fields, from the simulation of real patients to test possible interventions, to models to teach psychologists what can be the "anomalous" social behavior of certain pathologies, being able to experiment with it in an interactive way.

2.1.1 Artificial intelligence (AI) as it could help in medical treatments.

The field of artificial intelligence (AI) has evolved from its humble beginnings to a field with global impact. The definition of (AI) varies according to the thinking of each expert or technology insider for (Kaplan and Haenlein 2019) define "AI as the ability of a system to correctly interpret obtained data, learn from that data, and use those learnings to achieve specific goals and tasks through flexible adaptation." (Bartneck et al., 2021).

AI initiated the creation of machine learning and deep learning algorithms which allow to collect macro data, study it and learn from it to then give a prediction of a situation or area of study. It is currently the most widely used technology because it is versatile and can be applied to any area such as business, finance, entertainment, among others. But the area that interests us in this article is in the medical and more specifically in psychology.

2.1.2 Machine learning: brief description

Machine learning uses quantitative models to induce general principles underlying a set of observations without explicit instructions. Such algorithmic methods are characterized by

1) Make few a-priori assumptions
2) Let the data "speak for itself".
3) The ability to extract structured knowledge from large data. Its members include ***supervised methods***, such as performing vector machines and neural network algorithms, specialized to obtain the best possible prediction result, as well as ***unsupervised methods***, such as algorithms for data clustering and dimensionality reduction, effective for discovering new statistical configurations in data. (Bzdok et al., n.d.).

2.1.3 Types of ML applications focused on mental health

1. *Main data domains for ML in mental health***.** Four main types of data that serve to extract information to be analyzed and studied by ML algorithms can be identified: sensors, text, structured data, and multimodal technological interactions.

   Sensors: the use of sensors in smartphones allows data collection, audio signal analysis. (Van Den Broek et al., 2013).

   Text: social media posts are a great source of data, text message content can also be extracted, as well as clinical suicide notes (Nobles et al., 2018).

Structured Data: these can be clinical records, questionnaire evaluation. (Jain & Agarwal, 2017).

Complex multimodal systems: They are based on the everyday technology of a robot or virtual agent with respect to its interactions with real people (Rastogi et al., 2018).

2. *Type of ML applications for mental health:* Models that have been successfully developed in the main areas of mental health:
   Symptom compression, detection, and diagnosis or in mental health outcome reports. ML algorithms provide early diagnosis and monitoring of pathological conditions through speech which analyzes their characteristics (Chang et al., 2011) as well as mood detection through data obtained from cell phone applications. (Morshed et al., 2019).
   Dyadic interaction: Allows assessment of the patient and his or her relationship with the physician and improvements in the treatment (Rastogi et al., 2018).

3. *Mental health behaviors or conditions according to the objective:* the aim is to improve mental health treatments by determining which2 are the behaviors or conditions, for example: depression that leads to suicide in certain cases, anxiety, stress and how it is reflected in moods, as well as post-traumatic stress disorders and schizophrenia that lead to drug abuse.

2.1.4 Use of machine learning algorithms to determine psychological pathologies

Machine learning allows us to obtain data on different types of approaches to a patient's mentalhealth and on their relationship with the psychotherapist. As well as to assess psychotherapy, the following table will show the applications of machine learning algorithms.

Table 1. Use of ML algorithms to identify psychological pathologies.

Data domain (sensors)	Target	ML Approach	What for?	Pathology	Reference
Audio	Detect symptoms	Development of a weak supervision of learning framework for social detection of anxiety and depression by audio by long clips including a novel feature model the technique is (NN2Vec).	To detect high-level social speakers and their symptoms that do not require extensive equipment or clinical training.	Anxiety	(Salekin et al., 2018)

Accelerometer detection	Symptoms condition	Development of multiple ML models to detect the presence and level of depression in motor activity recordings.	To accurately detect the depression and obtain an engine activity	Depression	(Frogner et al., 2019)
Multiple detection	Symptoms condition	Development of a continuous method for stressful event detection using a commercial wrist device.	To assist in self-management of wellness to develop a stress detection through an application that can be used through a cell phone.	Stress	(Gjoreski et al., 2016)
Multiple Understanding	(body) Predicting risks	Development of a system based on a cloud architecture to collect and process body sensor data in real time, as well as additional patient information to assess suicide risks.	It effectively prevents atypical and suicidal mental states in patients.	Suicide	(Rabiul Alam et al., 2014)
Messages (Chat application)	Treatment Improvement	Use of ML tools to assist potential supporters of text-based mental health issues allows for real-time evaluation of the quality of training.	Evaluate and improve the quality of responses Silby provides	Mental Health (Generic)	(Wilbourne et al., 2018)
Questionnaire (Mental Health Registry)	Mental health compression	Development of Bayesian models to improve and understand the most significant symptoms in patients with depression.	To better understand factors influencing mental health in patients with thoughts of death.	Depression	(Galiatsatos et al., 2015)
Questionnaire detection	Symptoms condition	Development of a multi-task RNN encoder-decoder to learn patterns of different users and serve to predict their moods.	To provide a useful tool for use in the early diagnosis of mood problems.	State of mind	(Spathis et al., 2019)

2.1.5 Deep Learning: brief description

Machine learning provides more viable solutions to complex real-world problems with the ability to improve through experience and data. Although machine learning can extract patterns from data, there are limitations in processing raw data, which relies heavily on hand-designed features. For this reason, deep learning is more promising as it manages to discover effective features as well as assignments from data for specific tasks.

Deep learning can learn complex features by combining simpler features learned from the data (LeCun Y, Ranzato M).

2.1.6 Types of deep learning algorithms

Deep Neural Networks (DNN)

Consisting of an input layer, multiple hidden layers, and an output layer, they can be classified as MLP, SAE or DBN. Among all the algorithms are the most recognized and successful because they allow to analyze a large amount of data as hierarchical representation learning methods could discover unknown patterns.

Convolutional Neural Networks (CNN)

They are mostly used for image recognition, as their architecture consists of nonlinear layers, seizure layers, and clustering layers. CNNs are designed to process multiple types of data, especially in two-dimensional images, in the case of the study of mental health could be used to analyze the cerebral cortex.

Recurrent Neural Networks (RNN)

They are designed to use the sequential information of the input data by means of cyclic connections between building blocks of perceptrons. These networks are not as deep as DNNs and CNNs, in terms of numbers of layers, because of their capabilities RNNs allow mapping avariable length between one sequence and another or a fixed size prediction.

3. METHODOLOGY

3.1 Proposed algorithm (CNN) based on Deep Stride Neural Network Architecture (DSCNN) for the identification of psychological pathologies.

The area of voice research in the field of (AI) is booming due to the use of voice assistants that are increasingly immersed in the daily lives of people, however it is not the only use, there are also call systems such as those used by call centers that can identify customer emotions through the voice and thus to serve them in an appropriate manner, But what is behind these, the technology used is not only the hardware, the key is the machine learning algorithm or rather the deep learning which allows a more complex processing of voice synthesis allowing the improvement of human-machine interaction.

We have already seen in which aspects of psychology machine learning algorithms can be used, but the real focus of study will be on neural networks (CNN) and on an architecture that further deepens speech recognition by emotions which is the (DSCNN).

3.2 Speech Emotion Recognition (SER)

In the last decade there have been many studies and technologies that enable human-computer communication and one of them has been the research area of speech emotion recognition (SER), this technology is developed to recognize and identify the different emotions that a human being can transmit through his voice, it also plays an important role when it comes to human-machine interaction in real time. Researchers are working in this field to find certain characteristics ranging from the most effective to the most discriminative of speech signals. This allows the process of classifying speech signals to become more concrete, and thus to determine the emotional state of the speaker.

The process of selection and extraction of the most outstanding and discriminative characteristics has a high level of difficulty, therefore the implementation of artificial intelligence focused on deep learning is needed to find the most robust and outstanding characteristics of the SER.

3.3 How would (SER) help us when determining a psychological diagnosis?

Implementing this technology combined with a deep neural network in a medical system would allow us to recognize the emotional state of the patient through their voice, by analyzing,

identifying and classifying emotions. Most patients do not say how they feel when they are in front of the medical specialist and sometimes, they lie about their states so that they will never have to attend a consultation again or the psychologist will stop asking questions.

The (SER) works in real time so it would decrease the time of diagnosis of the patient and the doctor could treat the patient more effectively and reducing the time of treatment. When one goes to a session we are asked about a lived experience and thus have a record in which we can find the cause of the psychological condition of the patient, and therefore this leads to many sessions, however with the help of the (SER) we could detect their emotions when talking about the lived experience and denote what is the level of affectation of that experience.

3.4 SER operation

This system starts by extracting the characteristics of a speaker's speech, then classifies them depending on the different existing emotions and predicts the type of emotion of the speaker. There are many ways to use speech recognition among them we have deep learning algorithms such as deep neural networks (RNN), long term memory (LTSM) and finally convolutional networks (CNN) that we will use in this study, among others.

The use of neural networks increases the computational complexity and therefore many problems arise when running the algorithms and give a result with greater assertiveness, among these cases we have the use of the (RNN) together with the short-term memory (LTSM) these allow sequential analysis to train the data, but due to its level of computational complexity is difficult to train. For that and other reasons it is much more effective to use (CNN) but with a different structure such as (DSCNN) which uses special strides to detect hidden patterns in speech signals for top-down sampling of the feature map. Speech Emotion Recognition (SER) is the natural and fastest form of human-computer exchange and communication and plays an important role in real-time applications. (Mustaqeem & Kwon, 2020)

3.5 Proposed Model

The proposed algorithm uses a discriminative convolutional neural network for the classification of emotional states, which allows the learning of discriminative features of the voice by using spectrograms. The CNN architecture will have input layers, convolutional layers, a flat layer, and fully connected layers, and at the end the use of a classifier such as SoftMax. For this neural network model, it was necessary to use spectrograms, because theycontain relevant information.

Table 2. Study model of a system based on the recognition of emotions of the individual assisted by (CNN).

What type of system?	What kind of algorithm?	What architecture do you use?	Through what?	Which study do you use?	How is the data processed?
Intelligent affective computing monitoring system	Convolutional Neural Networks Algorithm	The Deep Stride CNN Neural Network Architecture (DSCNN)	They use a sensor in the microphone which allows digital voice signals to be analyzed and processed.	They use the strategy of simple networks to learn the salient and discriminative features of convolutional layers.	Spectrograms of voice signals are sampled, as the network learns it could more assertively identify the condition of the individual.

3.6 Data Collection Methods and Techniques Used

3.6.1 Spectrograms

They are a 2D visual representation of speech data of the being due to this allows us to identify the different variations of the frequency and intensity of the sound over a period, in short allows us to determine the quality of the voice signal. By means of the Short- Term Fourier Transform (STFT) the speech signal can be calculated, which is represented bytime-frequency.

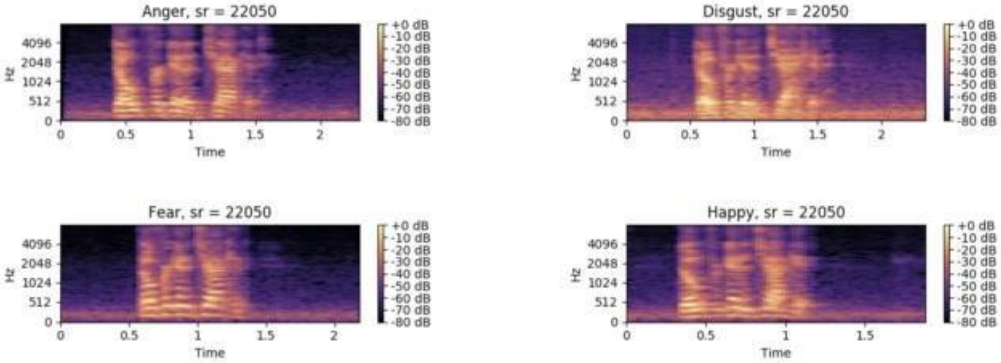


Figure 1. Spectrograms of different emotions

Spectrogram preprocessing

For the results of the algorithm to be more accurate and efficient, the audio signals must be cleaned, this is a relevant part of the data preparation process. Any type of noise and any other not so important information of the voice signal is eliminated, for this process the adaptive preprocessing based on thresholds is used.

3.6.2 Detailed description of the DSCNN creation process

For this study we used a functional model of the DSCNN neural network hosted in the GitHub repository, which we downloaded and tested to verify the functionality of the code.

One of the obstacles to implement the SER is the dimensioning of the signal using 2D CNN. Because of this, the representation is modified from 1D to 2D to fit the 2D CNN model perfectly. Using this model will allow us to learn more in depth the high-level characteristics of the speech signal. The spectrograms we are going to use are from the RAVDESS database which allows us to represent the audio file.

3.6.3 Architecture

The CNN neural network consists of several layers in a sequential pattern. This model is composed of several convolution layers, clustering layers and fully connected layers and SoftMax unit. The CNN model with deep stride is based on the simple neural networks, adapted for the (SER) and using a modification in the clustering layers. This neural network is the most suitable for the implementation of (SER) as they are specially designed for computer vision recognition, thus giving us a more accurate and efficient result than other deep learning algorithm.

We used a simple CNN model to extract high-level salient features using deep stride after obtaining speech spectrograms.

The model consists of five convolution layers with a linear unit (ReLU), three layers of maximum clustering followed by batch normalization (BN) and three layers fully connected to a SoftMax, which allows us to classify the probabilities of the speaker's emotions. The generated spectrograms are taken as input and convolutional filters are applied to them, which are learned while training allowing to obtain feature maps. The probing layers gather the maximum activation functions of the feature maps, thus reducing the dimensionalities. In the fully connected layers, all input layers are related to the other neurons in the layer. And to finish the process the SoftMax does its job of classifying the emotions.

The spectrograms generated by the RAVDESS database are taken as input. The first convolution layer consists of 96 filters with a kernel size (11x11) and the incoming spectrogram size is (224x224x3) with padding and Stride (4x4) pixels to effectively adjust the dimensionality of the data. Similarly, the second convolutional layer has 256 filters with a kernelsize (5x5) and Stride of (1x1).

Convolutional layers 3 and 4 have the same amount of filter which is 384 with a size of (3x3) and Stride of (1x1). The last layer has a filter of 256 with a size of (3x3) with a Stride setting of (1x1) pixels. A rectified linear unit (ReLU) which is an activation function is used. In addition, batch normalization is performed to improve the stability and performance of the model.

Finally, a flattening layer that converts the shape of the data into a vector is used after the last convolution layer. The output of the flattening layer is sent to the fully connected layers. In the fully connected layers, we have the first and second layers have 4096 and the third layer has 1000 neurons. Finally, for classification we have the SoftMax, which are fed with the extracted features.

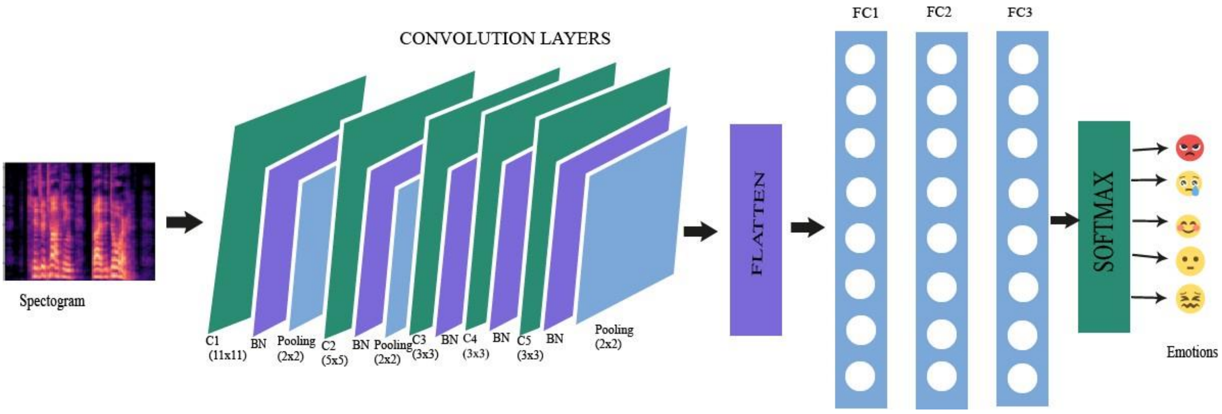


Figure 2. Neural network architecture (DSCNN) for emotion recognition.

3.7 Data analysis methods and techniques

3.7.1 RAVDESS data set

For this neural network model, we used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) which contains a complete dataset of 7356 files, with a weight of (24.8 GB) of audio video, voice, and song.

It is a dynamic and multimodal set consisting of facial and vocalized expressions in American English. The database has 24 actors who vocalize lexically making their neutral American accent noticeable by showing emotions such as neutral, sad, calm, angry, happy, disgusted, fearful, and surprised are the 8 emotions of self-talk used in this model. Only 1440 audio files were used for this study.

Table 3. *Detailed description of the data set on the number of audios for each emotion*

Emotion	Speech Sample Count
Neutral	
Calm	192
Happy	192
Sad	192
Angry	192
Fearful	192
Disgust	192
Suprised	192
Total	**1440**

Table 4. Number of actors contained in the DSCNN model system

Genre	No. of Actors	Number of audios
Woman	12	1440
Man	12	

3.7.2 Configuration of the proposed model

Python and its libraries such as librosa, numpy, matplolib, tensorflow, keras and pandas, among others, were used to develop the code.

To treat the input audio sets we used the librosa library which allows the analysis and processing of the time frequency and the extraction of audio features resulting in the spectrogram. Before modeling, we structured the data architecture with the panda library, creating a directory of files and creating a function for the extraction of the emotion and labeling it with a genre depending on the actor. For the neural network model, the keras library was used since it is designed to build each neural network in blocks. The number of emotions used was 8 emotions, we trained the model for 100 epochs with a learning rate of 0.000001.

The training was performed on a single Ryzen 5 GPU with 12 GB of RAM, only a single experiment of the DSCNN model was done was trained with the processed spectrograms and the model accuracy, recognition or prediction rate was evaluated

4 RESULTS

4.1 Experimental results

In this study, a deep convolutional neural network (DSCNN) was trained on the RAVDESS database, and its performance was evaluated. The model was trained with 100 epochs and its prediction performance was tested, respectively as shown in Table 5, as well as the recall, f1 score, weighted and unweighted accuracy of the model on clean spectrograms and using the RAVDESS dataset.

Table 5. Performance of the DSCNN model with training of 100 EPOCHS

	Precision	recall	F1-score	Support
Angry	0.94	0.74	0.83	
Calm	0.67	0.87	0.76	
Disgust	0.80	0.65	0.72	
Fearful	0.79	0.50	0.61	
Happy	0.53	0.54	0.54	
Neutral	0.68	0.62	0.65	
Sad	0.28	0.55	0.37	
Suprised	0.85	0.44	0.58	
Accuracy			**0.61**	288
macro avg	0.69	0.61	0.63	288
weighted avg	0.70	0.61	0.63	288

4.1.1 Matrix of Confusion

For the proposed DSCNN architecture the numerical confusion matrix plot with 100 training epochs. It is observed that the prediction performance of anger, calm, disgust, happiness, neutral, sad, surprised are less than 50%. However, the overall accuracy of the model is 61%. This matrix allows us to obtain the number of correct and incorrect predictions showing the actual predicted values diagonally and the confusion between them in the corresponding rows.

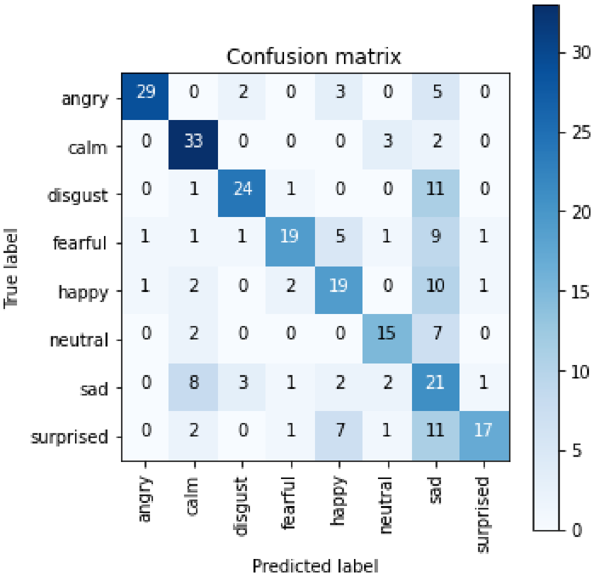


Figure 3. Confusion matrix of the RAVDESS data set with 100 epochs.

According to the RAVDESS confusion matrix, the highest results obtained for anger, calmness and disgust were 29%, 33% and 24% respectively. The lowest accuracy score was 15% and this belongs to neutrality. For measure F1, the highest results were 83%, 76%, 72% for anger, tranquility, and disgust, and the lowest result was 37% for sad. The highest accuracy rates were 94%, 85%, 80%, 79% and they were for anger, surprise, disgust, and fear, for the lowest rates are 28% and 53% which are for sad and happy.

4.1.2 *Training and validation graphs showing performance and losses*.

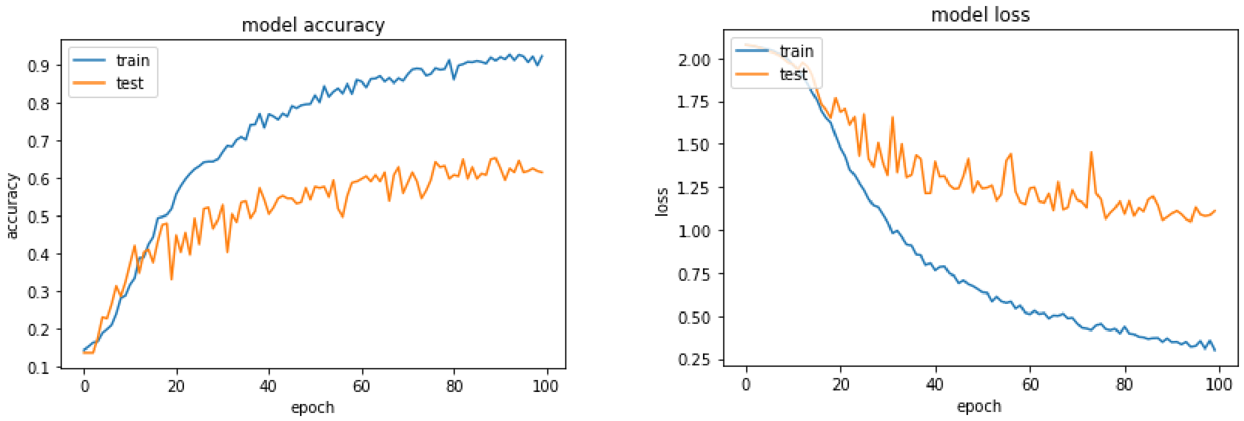


Figure 4. Shows system performance and illustrates training and validation accuracies and losses.

4.2 Comparison of results using the DSCNN model

Table 4. Comparison of the model with other authors

Data Set	Accuracy	Author
RAVDESS	61%	Our model
RADVESS	81%	(Mustaqeem & Kwon, 2020)
SAVEE	68.3%	(Wani et al., 2020)

5 DSCNN MODEL IMPLEMENTATION PROPOSAL

5.1 Identify patterns in the voice by means of the (DSCNN)

Traditional speech emotion recognition models have limited capabilities, as they can only identify a limited number of short words to classify emotions. The problem with these algorithms is that natural language has many facets such as accent, semantics, context, and foreign language words.

With the design of the deep convolutional neural network (DSCNN) improves the analysis and classification of emotions, because its algorithm is developed to process a large set of data, allows greater accuracy, and manages to learn and adapt to the different changes in the speaker's way of speaking and languages. Our model is not only better, but the database (RAVDESS) it uses has speech files and not only specific words, so it can be trainedfor real-time querying.

5.1.2 How would patterns be identified?

By using this model we could use it in real time, which would allow us to capture the frequency of the voice and create a spectrogram, which will then be analyzed and classified depending on the emotion that the speaker transmits, the therapist will have the results instantly, but to create a pattern would need that through the results a record of the patient's voice is created, before the therapist asks specific questions and during the therapeutic session, since in most of the procedures they ask about the close relationships of the individual, their likes, problems and most importantly the experiences lived, in this we can detect through the algorithm if the patient when answering these questions, is physically expressing this emotion or the hidden ones of the health professional. For example: if the patient was asked what he felt when his father left him and he answered "nothing", the psychologist would only write down in his notebook what he thinks of the patient, but, even if he sensed that the patient feels sad, he would not know exactly if it were that emotion or another that the patient feels at that moment, however, with the help of the algorithm we could detect and know precisely if the type of emotion that the patient reflects is the correct one or another.

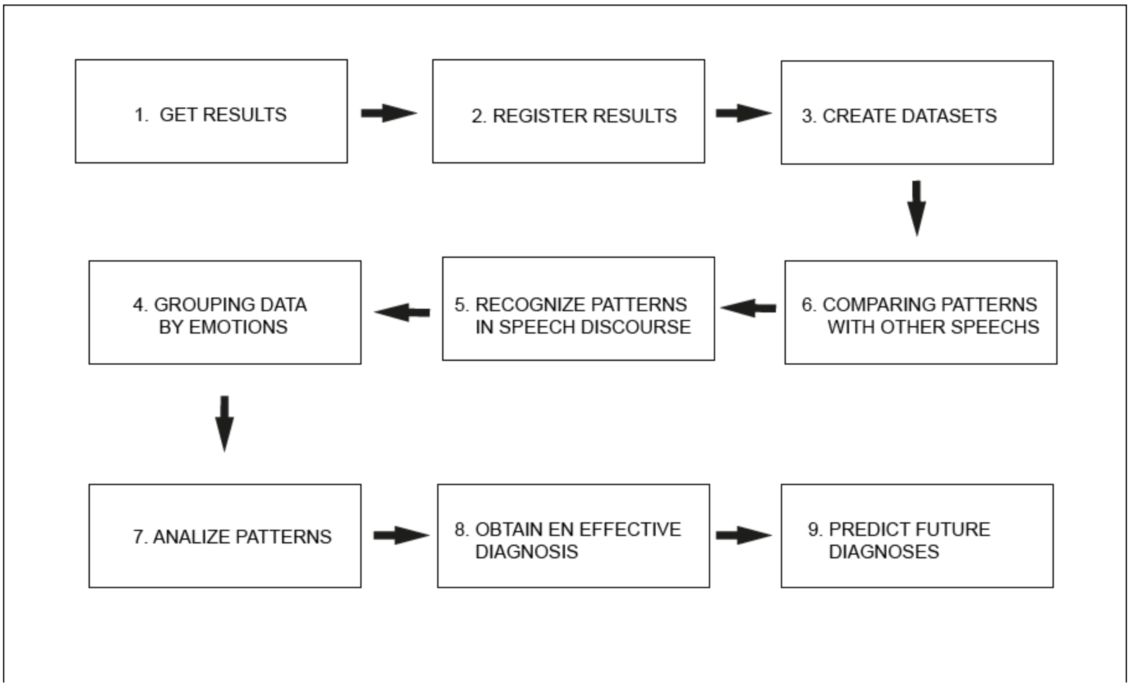

1. GET RESULTS
2. REGISTER RESULTS
3. CREATE DATASETS
4. GROUPING DATA BY EMOTIONS
5. RECOGNIZE PATTERNS IN SPEECH DISCOURSE
6. COMPARING PATTERNS WITH OTHER SPEECHS
7. ANALIZE PATTERNS
8. OBTAIN EN EFFECTIVE DIAGNOSIS
9. PREDICT FUTURE DIAGNOSES

Figure 5. Pattern identification process

5.1.3 Determine what psychological pathology the patient suffers from by identifying patterns.

The human being has 27 emotions according to a study conducted at the University of Berkeley. There are 6 basic or primary emotions such as happiness, sadness, anger, surprise, and disgust, from these the other emotions are derived, for this reason the complexity increases at the time of recognizing a specific emotion that corresponds to the physiology of the patient and allows to identify why he/she has a certain type of behavior.

5.1.4 Most frequent Psychological Pathologies in society.

Mental illnesses have been on the rise in recent years, but two of them top the list as the most common or frequent in the population and these are depression and anxiety, WHO estimates that it affects 3.8% of the population, including 5% of adults and 5.7% of adults over 60 years of age. Worldwide, approximately 280 million people suffer from depression and some 264 million from anxiety.

The number of people suffering from these mental illnesses was already shocking, now it has become a bigger problem, since the beginning of the COVID-19 pandemic caused hundreds of millions more people to develop depressive and anxiety disorders. A study was conducted, and it was determined that 76 million more people suffer from anxiety and 53 million from depression. For a change, medical treatments are not optimal and delay in making a diagnosis to determine what factors influence the patient and thus give them a specific, effective, and fast diagnosis.

5.1.5 What type of emotions do depression and anxiety have?

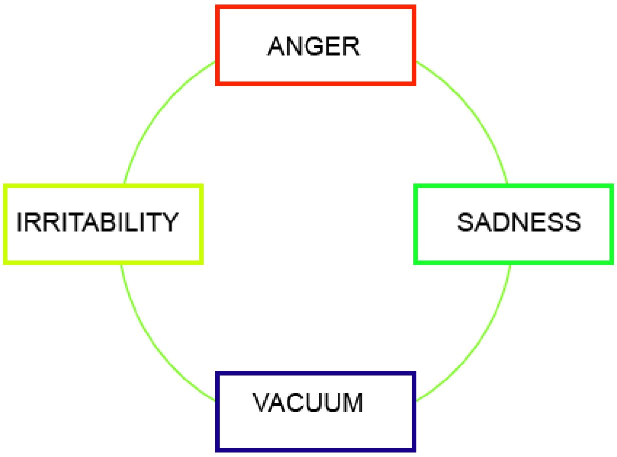


Figure 6. Depression Emotions

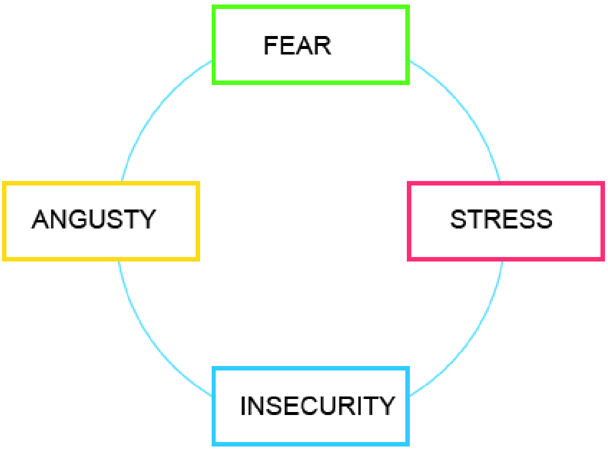


Figure 7. Anxiety Emotions

.

5.1.6 How would the patient be diagnosed by emotion?

The intelligent affective system would have a big data, which already has the records of the patterns of emotions related to each pathology, it is important to note that these data have already been trained with the neural network (DSCNN) managing to identify and classify with high precision each emotion that differentiates one pathology from the other.

The system runs in real time, i.e. the input data obtained through a microphone with sensors will be analyzing each part of the patient's speech, and then perform a recognition of emotions and group the data according to the classification of these, it proceeds to compare with the data that are in the system which determine what type of pathology corresponds to these emotions, if it is depression or anxiety.

Finally, the psychotherapist would have a more accurate report ofthe moods of his patient and what kind of factors cause him to have a certain type of behavior,and thus give an effective diagnosis and treatment to the patient and short duration.

6 SURVEY RESULTS

We conducted a survey of our university classmates to determine whether they present symptoms related to these mental disorders and what type of external factors could be influencing them. We chose to establish an age range between 21 and 30. The number of respondents are 100 people between men, women and people who do not identify with thesegenders.

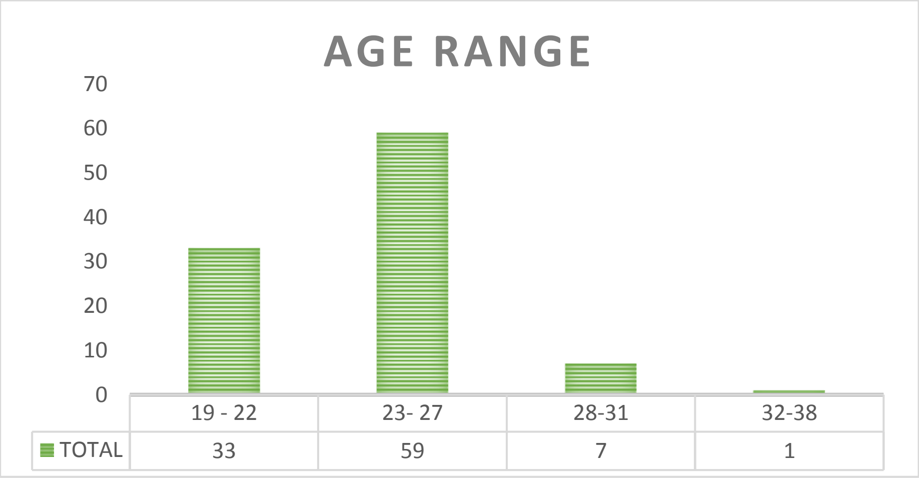


Figure 8. Age of respondents

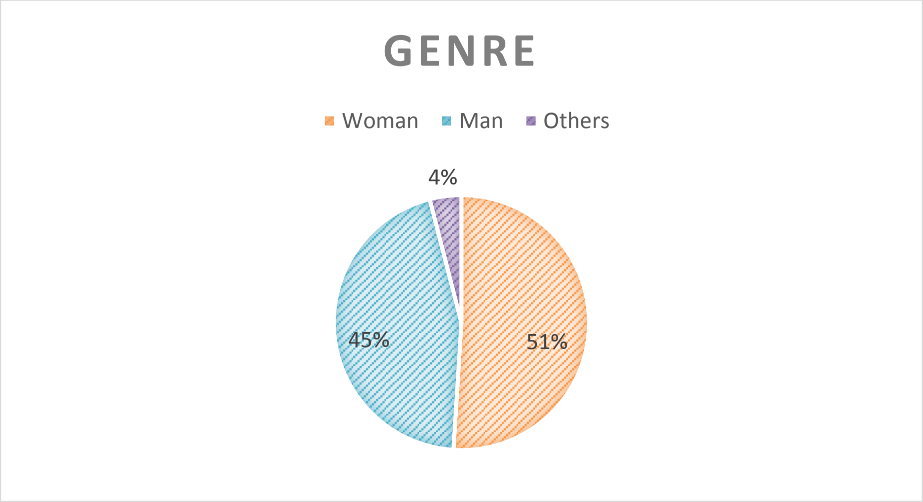


Figure 9. Gender of respondents

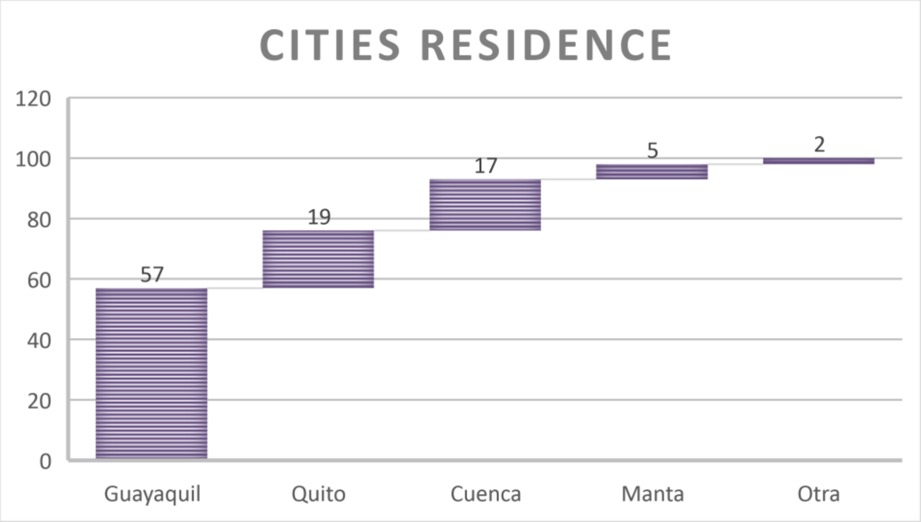
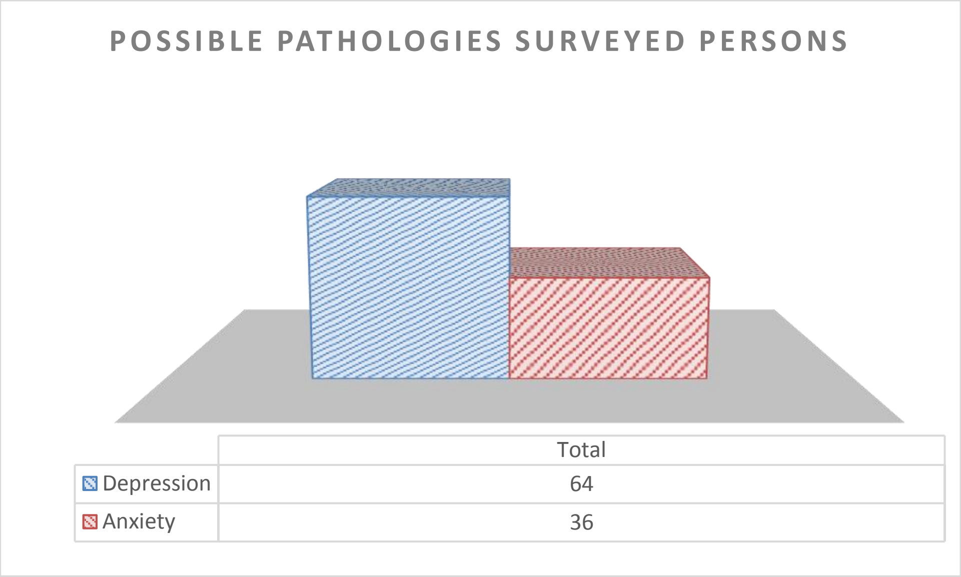


Figure 10. City of Residence

6.1 Factors that may influence the suffering of depression and anxiety

The survey was conducted with questions based on possible symptoms that may occur in a person with depression or anxiety, which allowed us to know that most of our respondents could be suffering from these pathologies, however, the pathology of anxiety has not been the one that has the largest number of people who could suffer from it, rather it is depression and is the one that worldwide is the most common and dangerous because many people incur suicide.



	Total
▨ Depression	64
▨ Anxiety	36

Figure 11. Percentage of respondents who may have these mental disorders

Among the external factors that could influence the people surveyed were economic problems, lack of employment, love breakup, death of a family member or close person, and the Covid-19 that caused the confinement and led to an increase in depression and anxiety. Below are two graphs obtained from the survey where we can see these types of factors that can affect the individual.

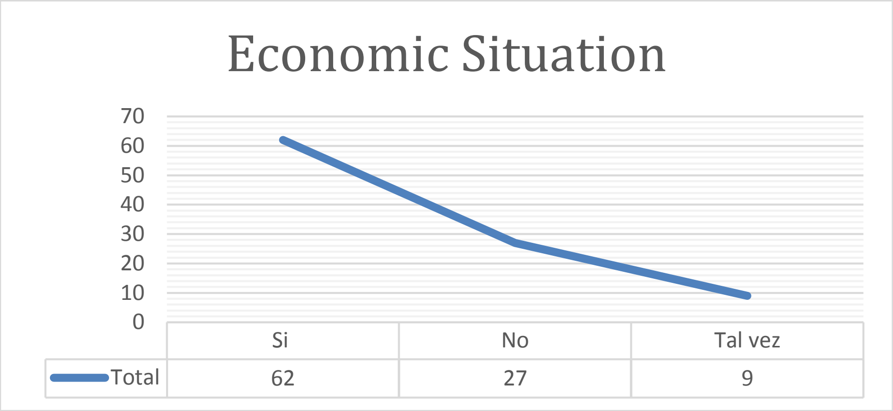


Figure 12. Factor 1 economic situation

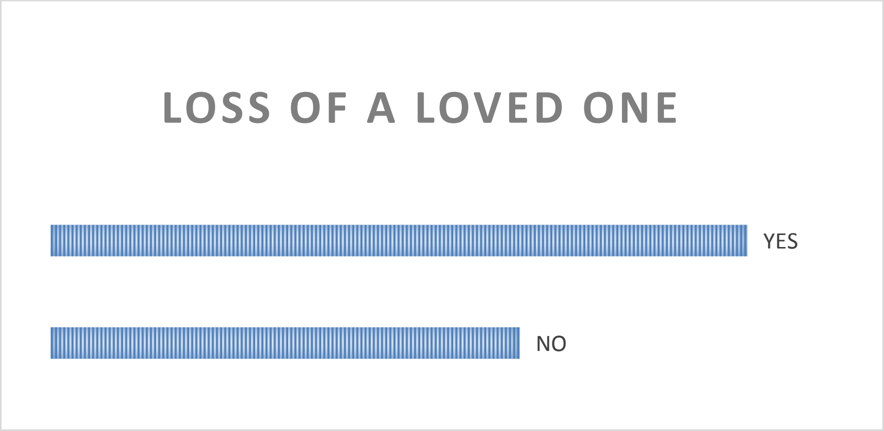


Figure 13. Factor 2 Loss of a loved one

7. DISCUSSION

In this study, the proposed model architecture (DSCNN) with the use of deep stride strategy is considered an important advance in (SER), as it manages to reduce computational complexities as well as massive data processing due to this improves the accuracy of emotion recognition. Researchers have used many methods using manual features and high-level features for (SER) but none have been effective, including the use of (CNN) does not provide optimal results. For this study our model (DSCNN) uses a clustering scheme to recognize the speaker's voice signals. Our model used input spectrograms which represent the frequency of the voice in a two-dimensional image. Using deep frequency features the model obtained an accuracy of 61%. The training was with 100 epochs. This model is simple; however, an efficient and effective system (SER) is obtained.

8. CONCLUSION

Innovations and studies in psychology are increasing more and more, due to the highdemand of wanting to understand the patient's behavior and emotions, in certain types of events,experiences and traumas caused. In addition to knowing what types of external factors can influence and provoke the suffering of depression or anxiety.

The (SER) is the basis of the study for the recognition of emotions through speech, this technology alone presents difficulties in accuracy and processing large amounts of data. Due to these challenges, we tested a deep CNN neural network model which is called (DSCNN), this model allows extracting high-level discriminative features and achieves higher accuracy than other neural networks and decreases computational complexity.

The model was trained with the RAVDESS database, considering 8 emotions and a total of 1440 audio-only files. Clean spectrograms were used for model input, which were previously processed to eliminate background noise. The results of the training of 100 epochs had a percentage of 61% in the prediction with this we can show the effectiveness of the model, the more the model is trained with more epochs the better its prediction will be.

An intelligent affective system based on the (SER) and with the architecture (DSCNN) will be able to identify emotional patterns and effectively predict what type of pathology or mental disorder a person is suffering from. The architecture of the model must continue to be improved to achieve a higher percentage and perhaps in the future 100% in the prediction of emotions by means of voice signals converted into spectrograms.

9. REFERENCES

Alharthi, H. (2020). Predicting the level of generalized anxiety disorder of the coronavirus pandemic among college age students using artificial intelligence technology. *Proceedings - 2020 19th Distributed Computing and Applications for Business Engineering and Science, DCABES 2020*, 218-221. https://doi.org/10.1109/DCABES50732.2020.00064. https://doi.org/10.1109/DCABES50732.2020.00064

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). What Is AI? In *SpringerBriefs in Ethics*. https://doi.org/10.1007/978-3-030-51110-4_2.

Benhamou, S. (2020). Artificial intelligence and the future of work. *Revue d'Economie Industrielle*, *169*(1), 57-88. https://doi.org/10.4000/rei.8727.

Bzdok, P. D., Ph, D., & Meyer-lindenberg, P. A. (n.d.). *Machine learning for precision psychiatry*. 1-16.

Chang, K. H., Chan, M. K., & Canny, J. (2011). AnalyzeThis: Unobtrusive mental health monitoring by voice. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/1979742.1979859

Frogner, J. I., Noori, F. M., Halvorsen, P., Hicks, S. A., Garcia-Ceja, E., Torresen, J., & Riegler, M. A. (2019). One-dimensional convolutional neural networks on motor activity measurements in detection of depression. *HealthMedia 2019 - Proceedings of the 4th International Workshop on Multimedia for Personal Health and Health Care, Co-Located with MM 2019*. https://doi.org/10.1145/3347444.3356238. https://doi.org/10.1145/3347444.3356238

Galiatsatos, D., Konstantopoulou, G., Anastassopoulos, G., Nerantzaki, M., Assimakopoulos, K., &

Lymberopoulos, D. (2015). *Classification of the most Significant Psychological Symptoms in Mental Patients with Depression using Bayesian Network.* https://doi.org/10.1145/2797143.2797159

Gjoreski, M., Gjoreski, H., Luštrek, M., & Gams, M. (2016). Continuous stress detection using a wrist device - in laboratory and real life. *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* https://doi.org/10.1145/2968219.2968306

Ishiguro, A. (1999). Experimental Studies of Endoscopic Mucosal Resection for Colorectal Tumor. *Gastroenterological Endoscopy, 41*(7), 1293-1300. https://doi.org/10.11280/gee1973b.41.7_1293.

Jain, V., & Agarwal, P. (2017). Symptomatic diagnosis and prognosis of psychiatric disorders through personal gadgets. *Conference on Human Factors in Computing Systems - Proceedings, Part F127655.* https://doi.org/10.1145/3027063.3048417

Morshed, M. Bin, Saha, K., Li, R., D'Mello, S. K., De Choudhury, M., Abowd, G. D., & Plötz, T. (2019). Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3*(3). https://doi.org/10.1145/3351233

Mustaqeem, & Kwon, S. (2020). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors (Switzerland), 20*(1). https://doi.org/10.3390/s20010183. https://doi.org/10.3390/s20010183

Nobles, A. L., Glenn, J. J., Kowsari, K., Teachman, B. A., & Barnes, L. E. (2018). *Identification of Imminent Suicide Risk Among Young Adults using Text Messages.* https://doi.org/10.1145/3173574.3173987

Rabiul Alam, M. G., Cho, E. J., Huh, E. N., & Hong, C. S. (2014). Cloud based mental state monitoring system for suicide risk reconnaissance using wearable biosensors. *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, ICUIMC 2014.* https://doi.org/10.1145/2557977.2558020. https://doi.org/10.1145/2557977.2558020

Ramos, L. (2014). COGNITIVE PSYCHOLOGY AND ARTIFICIAL INTELLIGENCE: MYTHS AND TRUTHS Cognitive psychology and artificial intelligence: myths and truths. *Unife, 22*(1), 21-27. file:///C:/Users/Pc/Downloads/270-Article text-1134-1-10-20180205.pdf.

Rastogi, N., Keshtkar, F., & Miah, M. S. (2018). A multi-modal human robot interaction framework based on cognitive behavioral therapy model. *Proceedings of the Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era - 20th ACM International Conference on Multimodal Interaction, ICMI 2018.* https://doi.org/10.1145/3279963.3279968

Rutledge, R. B., Chekroud, A. M., & Huys, Q. J. (2019). Machine learning and big data in psychiatry: toward clinical applications. *Current Opinion in Neurobiology, 55*, 152-159. https://doi.org/10.1016/j.conb.2019.02.006.

Salekin, A., Eberle, J. W., Glenn, J. J., Teachman, B. A., & Stankovic, J. A. (2018). A Weakly Supervised Learning Framework for Detecting Social Anxiety and Depression. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(2). https://doi.org/10.1145/3214284

Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019). Passive mobile sensing and psychological traits for large scale mood prediction. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3329189.3329213.

Tardón, C. G. (2008). *Social Behavior Simulator. Generation and application of. 38*(38), 61-73.

Van Den Broek, E. L., Van Der Sluis, F., & Dijkstra, T. (2013). Cross-validation of bimodal health-related stress assessment. *Personal and Ubiquitous Computing, 17*(2), 215-227. https://doi.org/10.1007/s00779-011-0468-z

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., & Ismail, N. (2020). Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks. *Proceedings - 2020 6th International Conference on Wireless and Telematics, ICWT 2020.* https://doi.org/10.1109/ICWT50448.2020.9243622. https://doi.org/10.1109/ICWT50448.2020.9243622

Wilbourne, P., Dexter, G., & Shoup, D. (2018). Research driven: Sibly and the transformation of mental health and wellness. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3240925.3240932.

10. ANNEX

Annex 1. Implementation of a supervised learning algorithm for voice study.

Table 5. Proposed implementation of a monitoring system by the authors of the article. (Alharthi, 2020).

What type of system?	What kind of algorithm?	How would it help?	What emotion would be studied?	What report would it show?	How would the data be obtained?
Intelligent affective computing monitoring system	Machine learning classification algorithm of the supervised learning type	Assist the psychotherapist in assessing the effectiveness of psychotherapy.	Types of emotions expressed through the voices of both the patient and the psychotherapist	This assessment would be made possible by means of a report containing graphical displays	From the follow-up system at the end of the psychotherapy session and allowing to evaluate the dyadic interaction in terms of vocal synchronization.

Appendix 2. Gartner chart

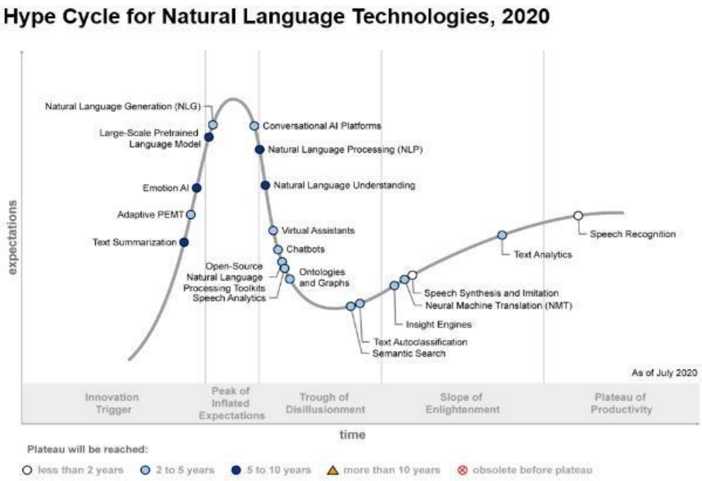


Figure 14. Speech Recognition Trend (Gartner 20020)