Received: 8 September 2022 | Revised: 3 December 2022 | Accepted: 9 January 2023

Cognitive Computation and Systems

DOI: 10.1049/ccs2.12076

ORIGINAL RESEARCH

# Speech emotion recognition with artificial intelligence for contact tracing in the COVID-19 pandemic

**Francesco Pucci**[1,2] | **Pasquale Fedele**[2] | **Giovanna Maria Dimitri**[1]

[1]DIISM, Universitá degli Studi di Siena, Siena, Italy

[2]Blu Pantheon, Siena, Italy

**Correspondence**
Giovanna Maria Dimitri
Email: giovanna.dimitri@unisi.it

**Abstract**

If understanding sentiments is already a difficult task in human-human communication, this becomes extremely challenging when a human-computer interaction happens, as for instance in chatbot conversations. In this work, a machine learning neural network-based Speech Emotion Recognition system is presented to perform emotion detection in a chatbot virtual assistant whose task was to perform contact tracing during the COVID-19 pandemic. The system was tested on a novel dataset of audio samples, provided by the company Blu Pantheon, which developed virtual agents capable of autonomously performing contacts tracing for individuals positive to COVID-19. The dataset provided was unlabelled for the emotions associated to the conversations. Therefore, the work was structured using a sort of transfer learning strategy. First, the model is trained using the labelled and publicly available Italian-language dataset EMOVO Corpus. The accuracy achieved in testing phase reached 92%. To the best of their knowledge, thiswork represents the first example in the context of chatbot speech emotion recognition for contact tracing, shedding lights towards the importance of the use of such techniques in virtual assistants and chatbot conversational contexts for psychological human status assessment. The code of this work was publicly released at: https://github.com/fp1acm8/SER.

**KEYWORDS**

affective computing, artificial intelligence, machine learning

## 1 | INTRODUCTION

Over the past 2 years, since humanity has been dramatically affected by COVID-19, our lives have drastically changed. COVID-19, acronym of Corona Virus Disease 19 [1], was first identified in Wuhan, China, in December 2019 and became a pandemic in 2020 [2].

Among the many challenges that the health system was facing during the pandemic, there was the crucial issue of contact tracing. By contact tracing we intend the search and contact management of a confirmed COVID-19 case. This was an essential public health action to pursue, to try to limit the ongoing epidemic. Indeed, identifying and managing the contacts of confirmed COVID-19 cases allowed to quickly identify and isolate any secondary cases, thus interrupting the infection transmission chain [3].

The operation of contact tracing required the employment of several human resources. In particular, this task affected dramatically the local Italian health authority (the Italian so called ASLs). Therefore an innovative approach to offer a solution to this issue consisted in the use of conversational interfaces based on artificial intelligence (AI) and Natural Language Processing (NLP).

The importance of emotion recognition in contact tracing dialogues is manifold. For instance it could help companies in detecting potential stress and psychological conditions of people involved in the conversations with the chatbots, as well as detecting potential liers. Furthermore, using the information obtained, the companies developing chatbots could decide to modify the flow of dialogues. Last, but not least, when worrying psychological emotions were to be detected by the automatic system, the company could decide to proceed

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Cognitive Computation and Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Shenzhen University.

further and make the human talk to a psychologist to help the person in need of further psychological help.

The startup company Blu Pantheon [4], with whom we collaborated, works in the development of virtual assistants. In particular, in our specific case, the goal of our work consisted in the implementation of a neural network-based system, capable of accurately identifying emotions starting from speeches. The speech emotion recognition model we implemented was tested on a novel dataset provided by Blu Pantheon. In particular, such dataset was collected thanks to a virtual contact tracing chatbot, developed by Blu Pantheon, which allowed to perform contact tracing during the first wave of the pandemic in Italy.

In our paper, we presented several novelties. First of all, to the best of our knowledge, we introduced, for the first time in the literature, the research question related to understanding emotions from contact tracing conversation in a chatbot context, with particular reference to the COVID-19 contact tracing task. Secondly, we used a completely novel dataset, collected by the startup Blu Pantheon. Thirdly, we used a transfer learning approach by using a dataset built for a similar but different context, characterised by the same language (i.e. Italian).

The paper is organised as follows: In Section 2, we report a comprehensive literature review for speech emotion recognition (SER) in AI.

In particular, we focussed on the role of SER for cognitive and psychological assessment of the health status of users. In Section 3, we report a comprehensive description of the datasets we used and of all of the pre-processing steps we performed. In Section 4, we describe the computational workflow and the methodologies used in our experiments. In Section 5, we described the experimental settings. In Section 6, we described the conclusions deriving from our work, together with a thorough description of the possible future developments.

## 2 | LITERATURE REVIEW AND BACKGROUND

In the latest years, machine learning (ML) and artificial intelligence (AI) has been successfully applied to many different fields [5–9]. Moreover, throughout the last 2 decades, research focussed on automatic emotions recognition using ML have been developed [10]. In the context of contact tracing, however, only a few works can be found, which relates emotion recognition to contact tracing conversations.

In Ref. [11], the authors analysed user reviews collected from the Irish Health Service Executive's (HSE) Contact Tracker app. The app was developed with the aim of identifying large-scale and automated analysis of reviews. A total of 1287 reviews from the Google/Apple playstores was collected to classify aspects of the app on which the users mostly focussed.

To the best of our knowledge, no works are present in the literature for what concerns the Italian language, and for non-app tracing related contexts, showing the complete novelty of the framework and analysis proposed in our work.

In the following subsections, we will present an overview of the main background concepts that are necessary to better understand the overall context in which our paper is focussed.

In Section 2.1, we will summarise concepts related to the field of Emotion AI. In Section 2.2, we will report a comprehensive summary of the available databases and corpora for emotion recognition. In Section 2.3, we will report a comprehensive overview of Speech Emotion Recognition Systems and the related ML approaches. In Section 2.4, we will present an overview of ML and related SER applications.

### 2.1 | Emotion AI

In a world where technology advances at exponential speed, Human-Computer Interaction (HCI) has become one of the most studied fields of research. The goal of HCI is not only to create a communicative interface that is as natural as possible between human and machine but it is also to create new communication paradigms that can improve human life. Emotions are, in fact, one of the predominant aspects of human interactions and, consequently, they have also become an important aspect of the development of HCI-based applications.

Emotions can be technologically captured and assessed in a variety of ways, such as facial expressions, physiological signals or speech. With the intention of creating more natural and intuitive communication between humans and computers, emotions conveyed through signals should be correctly detected and appropriately processed. Throughout the last 2 decades of research focussed on automatic emotions recognition, several machine learning techniques have been developed and constantly improved. This field of research is widely known as Emotion AI. Emotion AI can have several applications:

1. **Video gaming.** Using computer vision, the game console detects emotions via facial expressions during the game session and adapts to it [12]. This is often performed during the testing phase of a video game.
2. **Healthcare.** Several applications can be found in the healthcare sector. For example, voice analysis software can help doctors with the diagnosis of diseases such as depression and dementia. Another application is the use of 'nurse bot' not only to remind older patients on long-term medical programmes to take their medication but also talk with them every day to monitor their overall well-being [13].
3. **Education.** Learning software prototypes have been developed to adapt to kids' emotions. When the child shows frustration because a task is too difficult or too simple, the programme adapts the task so it becomes less or more challenging. Another learning system helps autistic children recognise other people's emotions [14].
4. **Employee safety.** Based on Gartner client inquiries[1], demand for employee safety solutions is on the rise. Emotion AI can help, for instance, to analyse the stress

and anxiety levels of employees who have very demanding jobs such as first responders.

5. **Automotive sector.** Automotive vendors can use computer vision technology to monitor the driver's emotional state. An extreme emotional state or drowsiness could trigger an alert for the driver. Another possible application is in the future of autonomous vehicles, where many sensors such as cameras and microphones could help to monitor what is happening and understand how users view the driving experience [15].
6. **Fraud detection.** Insurance companies use voice analysis to detect whether a customer is telling the truth when submitting a claim. According to independent surveys, up to 30% of users have admitted to lying to their car insurance company in order to gain coverage.
7. **Recruiting.** The software could be used during job interviews to understand the credibility of a candidate.
8. **Call centre intelligent routing.** An angry customer can be detected from the beginning and can be routed to a well-trained agent who can also monitor in real-time how the conversation is going and adjust [16].
9. **Connected home.** A VPA-enabled speaker (e.g. Google Home, Alexa, etc.) can recognise the mood of the person interacting with it and respond accordingly. Recent advances in Amazon's quest for humour detection are worth mentioning [17].
10. **Public service.** Partnerships between emotion AI technology vendors and surveillance camera providers have emerged. Cameras in public places in the United Arab Emirates can detect people's facial expressions and, hence, understand the general mood of the population. This project was initiated by the country's Ministry of Happiness.
11. **Retail.** Retailers have started looking into installing computer vision emotion AI technology in stores to capture demographic information and visitors' mood and reactions.

## 2.2 | Emotional speech databases or corpora

Over the last few years, the development of SER systems has led to the creation of labelled databases for emotion recognition.

Most of them are freely available and can be retrieved online. The databases used for the development of the SER system differ in language, number of different speakers, emotions represented and finally the type. The three types highlighted in Ref. [18] are simulated, natural and induced. We will now briefly describe all of the three types.

### 2.2.1 | Simulated emotional speech databases

A simulated/acted emotional speech database is a type of database with audio samples collected from actors who are aware of the recording process [18].

The recording process consists of a list of sentences and a certain set of emotions. Actors are asked to simulate a sentence for each emotional state. The advantage of this type of database collection is that researchers get full control of the quality of the recordings and therefore it is also easier to construct.

However, the disadvantage is that the natural component is neglected, and the resulting model may not work in real-time emotion recognition application.

### 2.2.2 | Natural emotional speech databases

A natural emotional speech database is a type of database with audio samples recorded in a natural environment. In general, it consists of recordings taken from call centres conversations, talk shows or movies [18]. In this case, speakers are not aware of the recording process. The advantage of this type of database is its reliability, since emotions are natural and the resulting SER model may obtain much better accuracy in real-time emotion recognition applications.

The disadvantage is the complexity in constructing and analysing it because emotions in everyday life may be less expressive and therefore recognisable compared to acted emotions. Another drawback is that it often contains unbalanced emotional categories.

### 2.2.3 | Induced emotional speech databases

An induced/elicited emotional speech database is a type of database with audio samples recorded under simulated natural conditions.

This means that speakers are put into situations to induce a specific emotion. That is why these types of databases are considered more natural than simulated databases but they are not entirely natural [18].

### 2.2.4 | Examples of existing emotional speech databases

Other emotional speech databases can be derived from multimodal emotional databases which include textual and visual data in addition to audio samples. Some of the best known in the English language are Crowd-sourced Emotional Multimodal Actors Dataset (Crema-D) [19], Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess) [20], Surrey Audio-Visual Expressed Emotion (Savee) [21] and Toronto emotional speech set (Tess) [22]. Another notable example is the IEMOCAP database [23]. IEMOCAP stands for 'Interactive Emotional Dyadic Motion Capture Database.' It consists of a corpus collected by the University of Southern California (USC) in which 10 actors were recorded while keeping sensors on faces, heads and hands. It represents an extremely important database for sentiment analysis. Also, in this case, the recorded speech are in English.

## 2.3 | Speech emotion recognition and machine learning

ML methodologies have, nowadays, reached the state of the art performances in several different fields: from bioinformatics to computer vision, from anomaly detection to computer forensic [5, 24–29].

The capability of machine learning to automatically perform difficult prediction and classification tasks has allowed to move steps forwards in solving tasks which were unsolvable a few years ago. In this context, new advances in research stands for emotion detection and recognition.

Emotions can be technologically collected in different ways, and multiple approaches can be used for their recognition. One example is the multimodal approach which aims to combine textual, audio and visual data [30].

However, in the present work, only audio data was used, as the novel dataset we analysed was audio only. In this case, we are no longer referring to Emotion Recognition in general but to the field of SER. This can be defined as extraction of the speaker emotional state from the speech signal [31]. Input data for SER systems consist of audio signals that are analogue representations of a sound.

Several approaches have been proposed in the literature, trying to exploit different features sound. The approaches used range from classic feature extraction to the implementation of different types of classifiers (i.e., Gaussian mixture model, Hidden Markov model, Support Vector Machine, Artificial Neural Network, etc.) [32] as well as deep learning models that act directly on sound representations such as the spectrogram and the time series (i.e., waveform). These latter allow automatic feature extractions, however requiring the availability of big datasets for model training [33]. For instance in Ref. [34], the authors use a Hidden Markov Model (HMM) approach to predict emotions. They obtained an accuracy of 80% in recognising seven different emotional states (disgust, fear, anger, joy, surprise, sadness and neutral) using the best combination of low level sound features (pitch and energy).

In Ref. [35], instead, the authors use a Support Vector Machine to classify five different emotional states (disgust, boredom, sadness, neutral, and happiness). They obtained 66.02% classification accuracy only by using energy and pitch features and 70.7% by using exclusively Linear Prediction coefficients and Mel cepstrum coefficients (LPCMCC) features extracted from the audio files and obtaining an accuracy of 82.5% using both of them.

Moreover in Ref. [36], the authors use a Multilayer Perceptron to recognise four emotional states (happy, angry, sad and neutral) with an overall accuracy of 81%. The network used is composed of an input layer, one hidden layer and the output of the four classes. Features extraction considered both temporal and spectral features for the classification task. The latest development of ML, that is, deep learning has also reached the state of the art for what concerns emotion recognition and classification. In particular, Convolutional Neural Networks (CNNs) have often reached state-of-the-art performances. For instance in Ref. [37], the authors used CNN-based architectures to perform SER on unlabelled samples. Performances were assessed on four public databases. A further example is, for instance, [38] where a real-time SER system based on dilated convolutions was proposed (DCNN). Residual blocks were proposed in order to learn the long-term contextual dependencies in the input features, and the features were later concatenated to perform the final emotion tasks. The architecture was tested on the IEMOCAP and EMO-DB benchmark datasets, obtaining a high recognition accuracy of 73% and 90% for each of the benchmarks. Among the most recent approaches, we can find the Wav2Vec2 approach, which recently reached the state-of-the-art performances for what concerns speech emotion recognition [39]. In particular, in Ref. [39], the author proposed to mask the speech into the latent space and use a contrastive loss so that the speech input is learnt. The method has now reached state-of-the-art performances in several different fields. Another state-of-the-art model in SER, which is worth mentioning, is in Ref. [40]. The proposed Hidden-Unit BERT (HuBERT) models aim at approaching the self-supervised speech representation learning field. The idea is to use an offline clustering step, in order to provide aligned target labels, with the final result of having a BERT-like prediction loss. The computational pipeline implemented includes a CNN and transformer encoding part together with a K-Means approach, resulting in a final model that is able to improve performances of BERT.

Deep learning (DL) methods (mainly based on CNN) have the great advantage of not having to specify the features of the sound to be used in advance. In this way, the extraction step, which could include human biases as well as a time consuming approach, is removed from the pipeline. However the amount of labelled data and, in general, the dimensionality of the dataset, which can be used in this case, is relevant, and therefore DL methods cannot be used in small unlabelled dataset as the case of the novel dataset we analysed and presented in our work.

## 2.4 | Speech emotion recognition (SER) and COVID-19

In this section, we will introduce literature related to the case of Speech Emotion Recognition for COVID-19-related research. COVID-19 has in fact significantly affected our lives and our emotions. The virus highly affected our way of interacting and significantly affected emotional and facial recognition in human interaction. Several researches have in fact investigated the relationship existing between facial emotion recognition, COVID-19 and emotions recognition [41]. In this context, several studies have in fact focussed on the possibility of identifying emotions from masked faces [41–43].

For what concerns speech emotion recognition and text, however, not the same can be said. A few studies, in fact, can be found, which relate emotion recognition and COVID-19-related texts. For instance, in Ref. [44], the authors analysed emotions from the data obtained from the TraceTogether app

and conducted a cross-sectional survey at the large public hospital in Singapore after the COVID-19 lockdown.

Moreover in Ref. [45], the authors used Twitter-based analysis for understanding people's feelings on social distancing from Twitter's data.

The data streams were analysed through the use of a Deep Learning approach (Deep Belief Networks) with pseudo-labelling. Moreover in Ref. [46], the authors analysed over 500.000 tweets related to COVID-19 from UK cities (collected in the last 2 years from February 2020 to November 2021). Using different types of deep learning approaches (based on emotion recognition and topic modelling), it was possible to observe the difference in sentiments related to vaccination and epidemiological situation. Moreover in Ref. [47], the authors collected more than 2 millions tweets in the period from February-June. In this way, a multi-class classifier was trained and used for understanding the Twitter COVID-19-related sentiments, achieving a classification accuracy of 80.33%. In all of the research cases described so far, the text features were extracted from the Tweets by using several different deep learning models and allowing to therefore use them as input to different types of classifiers.

For what concerns speech emotion recognition and COVID-19 related emotions, to the best of our knowledge, no works can be found in the literature in which the COVID-19 sentiment analysis was performed by using speech audio or text derived from audio. In this sense, our work represents an absolute novelty in this context and in the context of human-computer interaction sentiment analysis dialogues systems.

## 3 | MATERIALS

### 3.1 | Blu Pantheon contract tracing system and dataset

Blu Pantheon [4] is a startup active in the process innovation market through the implementation of innovative solutions for telemedicine and computer vision, with a strong focus on the development of the aforementioned conversational interfaces based on AI and NLP. It is precisely through the use of such technologies that it proposes new application scenarios that use AI in sectors such as healthcare and smart cities.

The product of the company is the result of the activity of the R&D team led by Dr. Pasquale Fedele, a computer engineer and serial entrepreneur who received the award Knight Order of Merit of the Italian Republic in 2017 for having created with his other company LiquidWeb, a medical device (i.e. Brain-control AAC) that gives people affected by amyotrophic lateral sclerosis (ALS) the possibility to control assistive technologies through their thoughts [48].

We will now here briefly describe the tool designed by the company to help the ASLs in the contact tracing process. The proposed solution involves the use of a conversational AI system, capable of discriminating the information useful for contact tracing thanks to the syntactic/semantic analysis of the interaction flow. In the case of the identification of the list of names and related contacts communicated by the positive patient to COVID-19, the conversational voice-bot processes the speech through a Speech-to-Text module. Subsequently, the agent analyses the text, extrapolating the words that can be mapped back to names, surnames and telephone numbers. Once identified, it proceeds to store them in a specific database.

The virtual learning agent process takes place both through the supervision of specialised technicians and through self-learning through the application of machine learning algorithms. This module interacts with the ASL telephone platform, as well as with the customer's database. A summary of the architecture is shown in Figure 1. More precisely, a call consists of a set of questions, following a path derived on the base of the answers given by the user.

The questions can be divided into seven types:

- *Person identification.* The identity of the user is identified. For privacy reasons, the conversations relating to this section are not reported in the dataset provided to us.
- *Clinical interview.* In this step the questions are asked that concern the patient's symptoms (symptomatic or asymptomatic). Based on this, there are different procedures to follow according to the Italian law.
- *Tracing.* The patient is asked if he/she lives with other people and if he/she can provide their names, surnames, dates of birth and telephone numbers.
- *Cohabitant interview.* The name, surname, date of birth and telephone number of each cohabitant are stored.
- *Job investigation.* A person is asked if she/he is working, the last time he went to work, the name and telephone number of the company. Also this information for privacy reasons has not been provided in our novel dataset.
- *Social interview.* The user is asked if he/she had other contacts outside his/her cohabitants and work. If so, all contact information is required (name, surname, date of birth and telephone number).

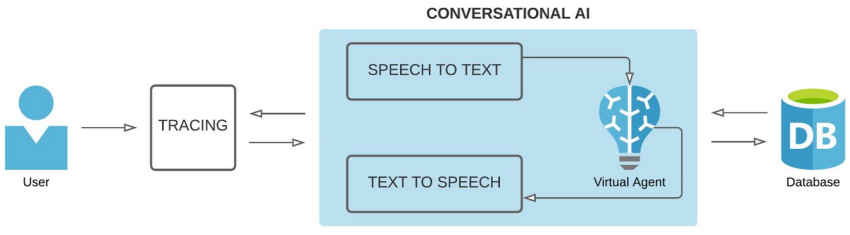


**FIGURE 1** Blu Pantheon contact tracing service architecture.

- *Quarantine guidelines and Greetings.* At the end, a series of recommendations to follow are listed, and the user is asked if he wants to listen to them again before the phone call ends.

It is worth noticing that during the call, a section can be skipped depending on the answers given by the user. Furthermore, if an answer is not understood, the user is asked to repeat up to a maximum of three times. Subsequently, the call is dropped.

The service has been operating on an experimental basis since November 2020 in support of some ASLs in the Veneto Italian region, finding a context with various problems to be solved such as epidemiological investigation, neophyte personnel to be trained or unpredictability of the COVID-19 pandemic and consequent frequent regulatory adjustments. During such validation period, a single human operator, supervising a virtual agent, has been able to carry out on average 9.6 cases per hour for an estimated management cost of 28.26 €/h, while to carry out the same workload in the traditional way (i.e., proceeding with a human operator to make contact tracing calls and manually fill the database) requires seven human operators for an estimated cost of 91.96 €/h.

Moreover, this service allows ASLs to manage phone calls peaks more easily and allows healthcare personnel to save time that is useful to carry out tasks that require greater competence. Furthermore, the modularity of the service allows its use on multiple channels, for example, through text messages or voice interaction. This favours social inclusion towards people with disabilities such as deaf-mute, visual impairment or motor difficulties in the upper limbs.

The service has been operating on an experimental basis since November 2020 in support of some ASLs in the Veneto region, finding a context with various problems to be solved such as

- Epidemiological investigation designed for small numbers and transferred to very large numbers in a short time
- Neophyte personnel to be trained and standardised
- IT tools not suitable for large-scale investigations
- Databases and applications not in communication with each other
- Unpredictability of the COVID-19 pandemic and consequent frequent regulatory adjustments

The novel dataset provided by Blu Pantheon for our experiments consists of 3005 audio samples in WAV[1,2] format, with a sampling frequency of 48 kHz. The sampling frequency of 48 kHz was chosen in accordance to the sampling frequency of the EMOVO dataset. In this way, uniforming the one of EMOVO and the one of our novel dataset, we could proceed training the dataset on EMOVO and then testing in the novel Blue Pantheon dataset.

The Blu Pantheon dataset was obtained from the recordings of the conversations between users and the virtual agent during the contact tracing calls. Each audio file corresponds to a single user response to a specific question of the virtual agent. For example, suppose the agent asks if the user had close contacts outside the family environment and the user replies 'yes,' then the audio file will consist of the recording of the user pronouncing the word 'yes.' The dataset provided is totally novel in which no descriptive label of the emotional state of the users is reported.

### 3.2 | EMOVO dataset

EMOVO is the first publicly available emotional corpus for the Italian language [49].

It is a database built using the voices of six actors (three males and three females) who played 14 sentences simulating six emotional states (disgust, fear, anger, joy, surprise and sadness) plus the neutral state [49]. These emotions are the well-known Big Six found in most of the literature related to emotional speech [50]. EMOVO is a perfectly balanced dataset made up of 588 audio samples.

In Figure 2 we show the duration of the conversations per sentiment. As we can see, they all appear to be quite balanced and of the same length. The sentences were designed with the emotionally neutral semantic content so as not to generate bias in the recognition of emotions by both machines and humans. EMOVO has all the phonemes of the Italian language, and all its sentences are characterised by a fair balance between voiced and unvoiced consonants.

The performances of the actors were recorded in the laboratories of the Fondazione Ugo Bordoni in Rome with professional equipment by using a sampling frequency of 48 kHz. The recordings were saved in *the WAV* format.

## 4 | METHODS

The workflow of our implemented methodology can be summarised into four main steps: data collection, data transformation, modelling and testing. The workflow is depicted in Figure 3. All of the code of our project is available at https://github.com/fp1acm8/SER.

In the following subsections we will describe each of the steps performed.

### 4.1 | STEP 1: Data collection

In this first phase we performed a thorough search to identify a possible Italian labelled dataset to be used in our study for training. As the novel dataset provided was unlabelled, so we decided to look for labelled dataset on which we could train a machine learning model to be later tested in the novel context.

[1]Gartner is a global research and advisory firm providing information, advice, and tools for leaders in IT, finance, HR, customer service and support, communications, legal and compliance, marketing, sales, and supply chain functions.
[2]Waveform Audio File Format (WAVE or WAV due to its filename extension).

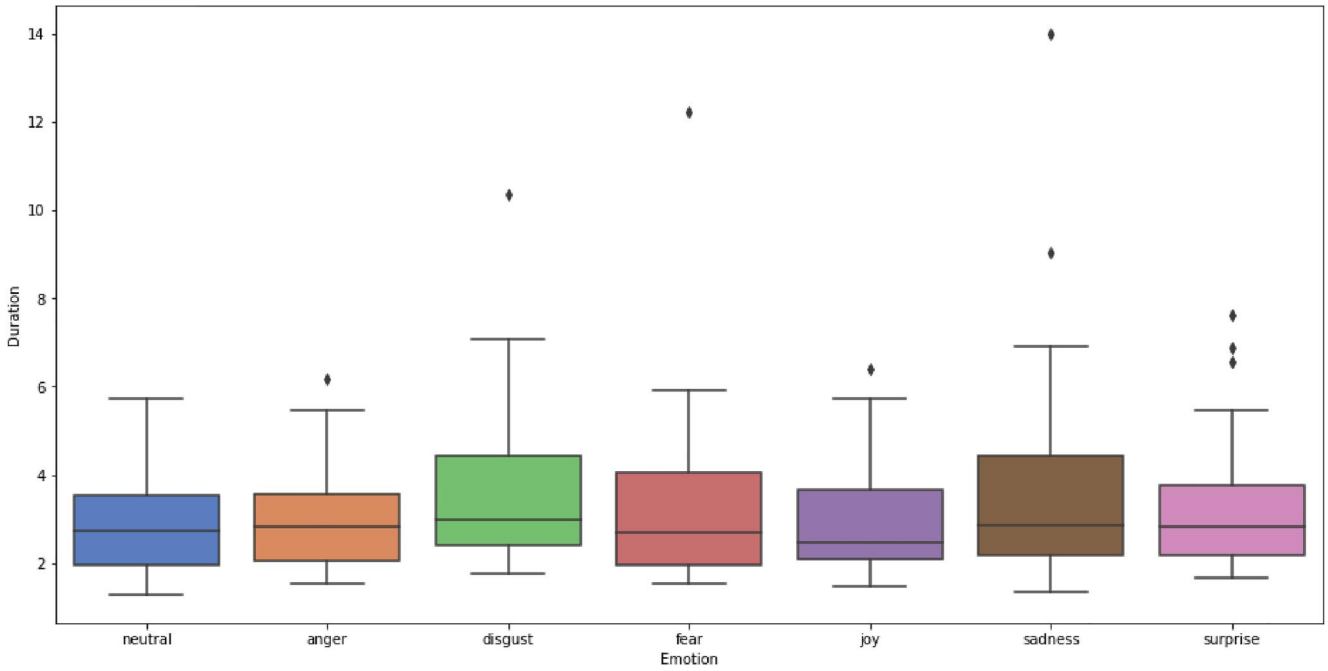


**FIGURE 2** Boxplot showing the duration of the conversations per emotion. In this way, we can see how the dataset is well balanced among the examples of different emotions.

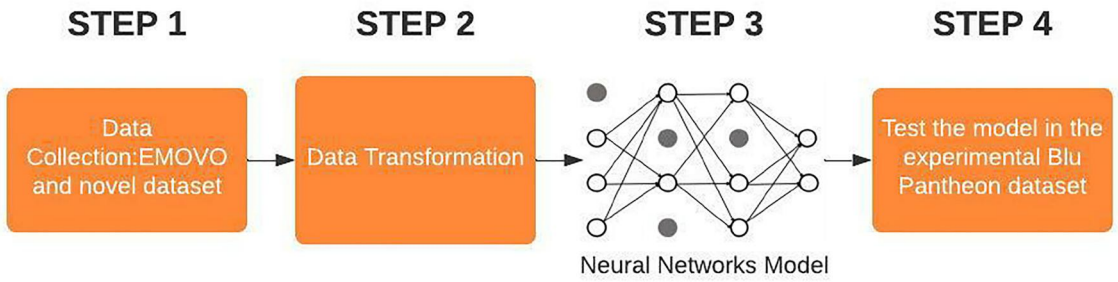


**FIGURE 3** Workflow of our experimental setting.

Language was not the only feature to consider when we looked for publicly available dataset. In fact, we also tried to look for dataset where there was a correspondence between the emotions that we wanted to identify in our model, considering the novel dataset and the contact tracing application.

For example, in the context in which we developed our project emotions such as 'happy' or 'joy' are not something we should expect in a conversation between COVID-19 positive patients and a virtual agent.

Moreover, we looked for dataset where there was a similar duration of the audio samples collected to have a correspondence between distributions in the training and test set. This is because the duration of an audio sample directly affects the amount of information extracted from the audio signal features. Therefore, having datasets with audio samples of extremely different duration could have affected the performance of our model.

Last, but not least, it is important to take into account the quantity and the quality of the collected data. It is crucial, in fact, to make sure that the audio samples collected are high quality data, not affected by excessive noise. For this reason, among the possible publicly available solutions, we decided to use the EMOVO dataset (described in Section 3) to train the SER model.

## 4.2 | STEP 2: Data transformation

We performed data transformation steps in order to increase the similarity between EMOVO and our experimental dataset.

Firstly, we decided to remove the emotion joy from the EMOVO dataset (as this was an emotion not pertinent to our experimental dataset). We further combined similar emotional states in a unique class (anger and disgust were merged in a unique class named disappointed).

Moreover, we performed data augmentation by using techniques such as noise addiction and change the pitch of the conversation.

Data were standardised using the $z - score$ normalisation and an extensive set of data cleaning steps were performed in the novel dataset. In particular, the following cleaning and data filtering steps were performed on our novel dataset. First, we deleted audio samples with a duration longer than 14 s that is, those audio with sampling errors made by the system developed by Blu Pantheon. Secondly, we deleted audio samples for which the question asked by the bot could not be retrieved. Such cleaning and transformation steps led us to obtain a dataset made of 2871 observations, which represented 96% of the samples originally provided to us by Blu Pantheon. These cleaning and transformation steps led us to obtain a dataset made of 2871 observations, which represented 96% of the samples originally provided to us by Blu Pantheon. Moreover, in our novel dataset, we performed extensive processing steps, merging the BOT and the user answer, as well as classifying the questions made by the BOT into seven classes: clinical interview, tracing, telephone number, name of the contact, social interview, guidelines and repeat.

## 4.3 | STEP 3: Modelling

In this central steps, we performed mainly two tasks: features extraction and machine learning modelling of the features. We will describe the two steps separately describing the methodologies used.

### 4.3.1 | Features extraction

First of all, we extracted the relevant features to be used as input to our classification model. In particular, we decided to use the python library librosa to extract the following features from the audio data:

- Chroma vector: a 12 element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music
- Root mean square (RMS) value for each frame from the audio samples
- Spectral features: Spectral flatness, spectral centroid, spectral bandwidth and spectral rolloff
- Zero-crossing rate
- The Mel-frequency cepstral coefficient

The final dimension of the EMOVO dataset was 1754 observations and 39 features. The choice of such features, rather than others which are commonly used in speech experiments (such as the Mel Frequency Spectrograms) was mainly driven by the dataset dimension. Having only a small amount of labelled data (the EMOVO dataset) did not allow to properly train an ML model with such features, and this is the reason why instead of spectrogram information we decided to use the set of features described above.

## 4.4 | Machine learning modelling

In our project, the chosen ML architecture is consisted in a Multilayer Perceptron (MLP), which we will describe more in details in Section 4.5 Moreover, we compared the performances of the MLP to two further extremely famous ML methods: the Support Vector Classification and The Extreme Gradient Boosting (XGBR) method. Support vector machines (SVMs) represent among the most powerful machine learning models, developed since 1995 and they are based on the VC theory (so called as proposed by Vapnik and Chervonenkis [51]). They can be used succesfully both for regression and for classification. Since in our case, we used them for classification. We used the acronym SVC: Support Vector Classification. Also XGBR represents one of the most popular algorithms for classification and regression nowadays (since the introduction in 2015) and is based on the gradient boosting algorithm [52]. For more details, please check the reference paper [52].

## 4.5 | Multilayer perceptron

Since their introduction in the 80s, neural networks models have proved to be extremely successful in performing a wide variety of different classification and regression tasks [24] and have been successfully applied to several different fields from biology to natural language processing, from object detection to scene classification [53–56]. In our project, we implemented a classifier based on a multilayer perceptron (MLP) neural network. The architecture implemented is made of 39 units in the input layer with Relu activation function, a dense and a dropout hidden layer and a final dense output layer. We report the implemented neural network scheme in Figure 4. We used Adam optimiser and categorical cross-entropy loss function. The categorical cross-entropy loss function is defined as follows:

$$Loss = -\sum_{i=1}^{n} y_i \cdot \log \widehat{y_i} \tag{1}$$

where $n$ is the output size (i.e., the number of label classes), $y_i$ is the target value and $\widehat{y_i}$ is the $i$-th scalar value in the model output. In particular $y_i$ and $\widehat{y_i}$ are probabilities associated respectively to the true and predicted $i$-th class. We used 10-folds-cross validation, batch size of 32, 10 epochs and early stopping. To implement the MLP architecture, we used the Keras python library.

## 4.6 | STEP 4: Test the model in the experimental dataset

In step 4, we tested the trained model on EMOVO in our new unlabelled experimental dataset. An emotion was predicted for each new sample.

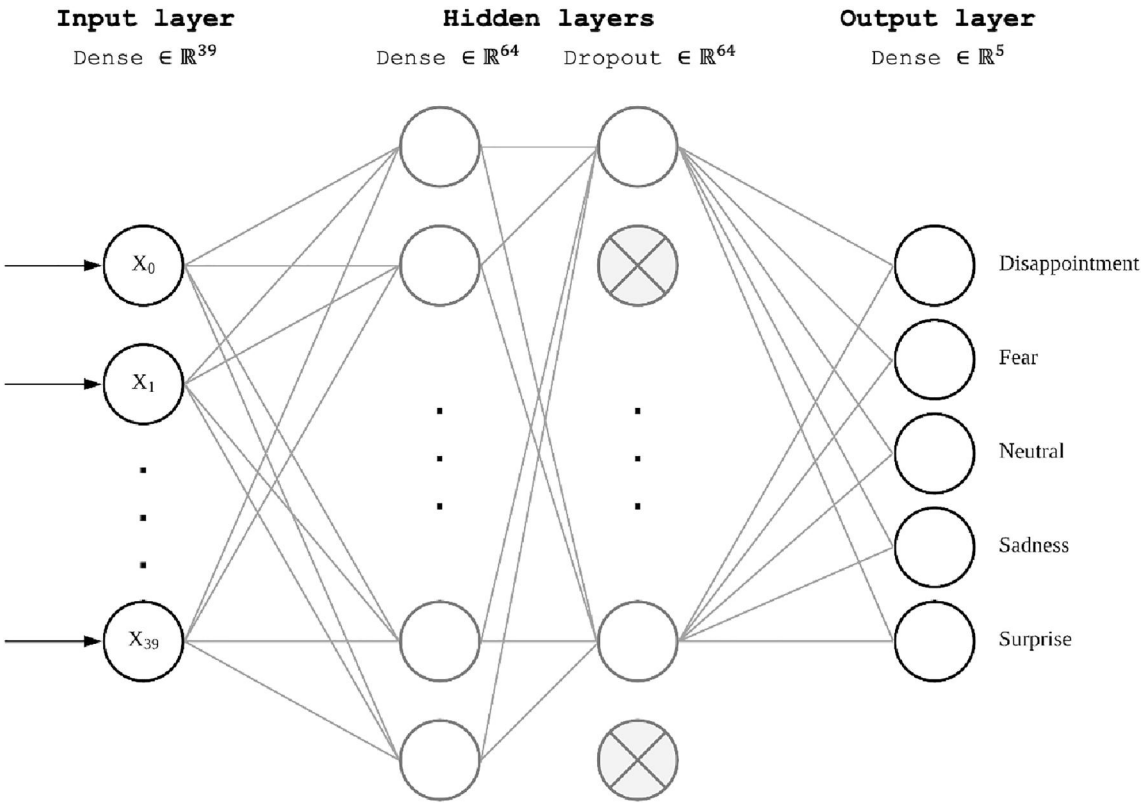


**FIGURE 4** Representation of the multilayer perceptron architecture (MLP) with dropout represented by the nodes with a cross in the second hidden layer.

Moreover, we proceeded in the following way. Since we had also time and the user id information, we merged the various pieces of conversation of a single user, considering to obtain the overall stream of a conversation. Knowing that an entire conversation is made up of multiple audio samples, we actually predict the emotional state of a single user several times based on the length of the conversation (i.e. the number of audio samples per user). We therefore made the assumption that it is unlikely that a user during a conversation of a few minutes will change his emotional state many times. So if the model predicts 3/4 different emotions for a single user, we could conclude that the prediction will be inconsistent. While this assumption is plausible, it does not sufficiently explain the results obtained. Therefore we explored the data further through data visualisation techniques (see Results and Experiments section). For instance, investigating if there were any patterns in the predictions depending on the sentence pronounced by the user or by the agent.

## 5 | EXPERIMENTS AND RESULTS

### 5.1 | Training on EMOVO dataset

We first trained our model on the EMOVO dataset. For evaluating performances in the EMOVO dataset, we used a 10-fold cross-validation approach, with 90% of this data was used to train the dataset, while the remaining 10% was used as a test.

**TABLE 1** In the table, we show the performances in terms of precision, recall, *F*-Score and accuracy for the test folder in the 10-fold cross validation.

| Metric | XGB | SVC | MLP |
|---|---|---|---|
| Precision | 0.81 ± 0.11 | 0.80 ± 0.06 | **0.93 ± 0.05** |
| Recall | 0.74 ± 0.13 | 0.82 ± 0.09 | **0.91 ± 0.09** |
| FScore | 0.76 ± 0.08 | 0.84 ± 0.03 | **0.92 ± 0.03** |
| Accuracy | 0.76 ± 0.13 | 0.80 ± 0.09 | **0.92 ± 0.05** |

*Note*: We report the mean and standard deviation over the 10 test folds. Bold values represent the best performing cases.

The training results obtained were promising, in terms of accuracy obtaining a mean accuracy of 0.92 (0.02 standard deviation). In Table 1 we report the test set performances (mean and standard deviation in the 10 fold). We report them for the MLP, the SVC and the XGB classifier implemented as baseline experiments (and which we described in the methods section). Performances were evaluated according to the following performance indicators: accuracy, precision, recall and F1 score (we report their definition in the Supplementary Material S1).

As we can see from Table 1, the MLP outperformed in all of the performance metrics than the other two machine learning models implemented. Therefore we chose the MLP as the optimal model and proceeded in testing it in our novel dataset from Blu Pantheon. In Figure 5 we also present the confusion matrix with the mean accuracies over the 10 folds obtained in the Emovo test set.

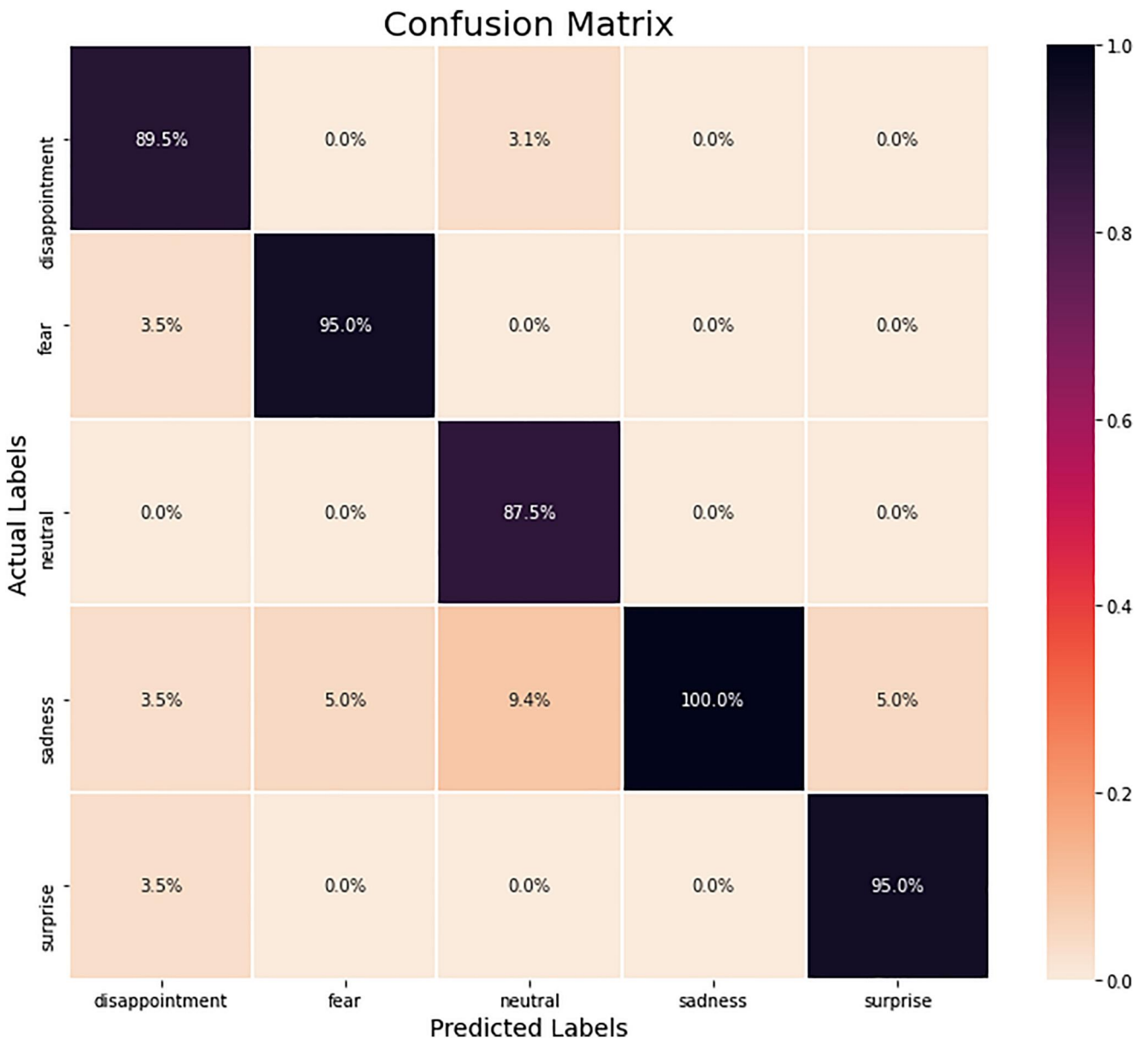


**FIGURE 5** Average accuracies of the test set in the 10-folds cross validation, divided per emotion.

## 5.2 | Testing on Blu Pantheon novel dataset

We therefore proceeded using the fitted model to test it in our novel dataset. This was possible due to the high similarity between the EMOVO and the original dataset, in terms of audio samples lengths and distributions.

Comparing the two datasets it can be noticed that the two datasets distributions are very similar up to the 75% percentile, in terms of sounds tracks. The minimum duration, the median and the 25% percentile, differs only by a few hundredths of a second. A difference of 1.10 s between EMOVO and the experimental dataset can be found in the mean value of the duration of the respective audio samples. This difference is due to the fact that in the experimental dataset, there are audio samples with a duration of up to 57 s. However, inspecting these samples, it was noticed that the long durations are often due to sampling errors made by the system developed by Blu Pantheon.

For instance, in the 57.64 s recording, the user answers the question asked by the virtual agent in the first few seconds of the recording while the system continues to record background noises thinking it is the user who is answering.

Following these considerations, we decided to discard all those defective records assuming the same maximum duration as EMOVO (i.e. 14 s). In this way, the difference between the mean values drops to 0.72 s, explained by the higher standard deviation in the case of the experimental dataset.

In conclusion, we can say that from the point of view of the duration, the two datasets are similar enough to justify the use of the model trained on EMOVO to be taeted later on the experimental dataset by using a transfer learning approach.

## 5.3 | Experimental sentiment analysis of the novel Blu Pantheon dataset

The first experimental analysis we performed was per user (Subsection 5.3.1). Subsequently, we analysed the predictions more in depth on the basis of the question asked by the virtual agent (Subsection 5.3.2).

Eventually, we made some further qualitative evaluation, verifying whether by listening to some audio samples it was possible to distinguish the different predicted emotions (Subsection 5.3.3).

### 5.3.1 | Analysis by user

In this section we analysed prediction grouping by users.

Each user corresponds to a certain conversation, and each conversation is made up of multiple audio samples. So it can happen that for a single user, several predictions of his emotional state have been made (remember that the model predicts an emotion for each audio sample provided).

We therefore decided to associate each user with the *mode* among all the emotional states predicted for him/her. For example, suppose the model predicts five times 'neutral,' three times 'fear' and two times 'disappointment' for a generic user $x$.

Then the emotional state associated with the user $x$ is 'neutral.' In this way, we have only one prediction per user for a total of 353.

A summary representation can be found in Figure 6. The prediction made is the mode of all the predicted emotions for a single user, so the emotion associated with a certain user is not necessarily the same throughout the conversation but it can change.

Only in 14% of the cases observed (i.e. 48 cases out of 353), the emotion remained unchanged. A summary representation of these cases, for which the model predicted only one emotion during the whole conversation, can be seen in Figure 7. From these two graphs, we see how 'disappointment' and 'sadness' are the emotions most likely to be predicted for the entire duration of the conversation. This concept of uniqueness of prediction is important because, as we mentioned before, it is unlikely that a user will experience three or four different emotions during a 10 min conversation. On the other hand, if only one emotion is detected during the entire conversation, then it will be likely that the prediction made for that user is reliable.

At the same time, this does not mean that the model does not work properly. There may be nuances in the user's voice depending on the question asked by the virtual agent that cause one emotion to be predicted rather than another. It is therefore important to analyse the emotions also on the basis of the questions received by the users.

### 5.3.2 | Analysis by sentence

Sentiment analysis by sentence was carried out by analysing the predicted emotions according to the questions asked by the virtual agent which we have divided into seven classes defined in Section 3.

Six of these classes (i.e. Clinical interview, Tracing, Telephone number, Name of the contact, Social interview and

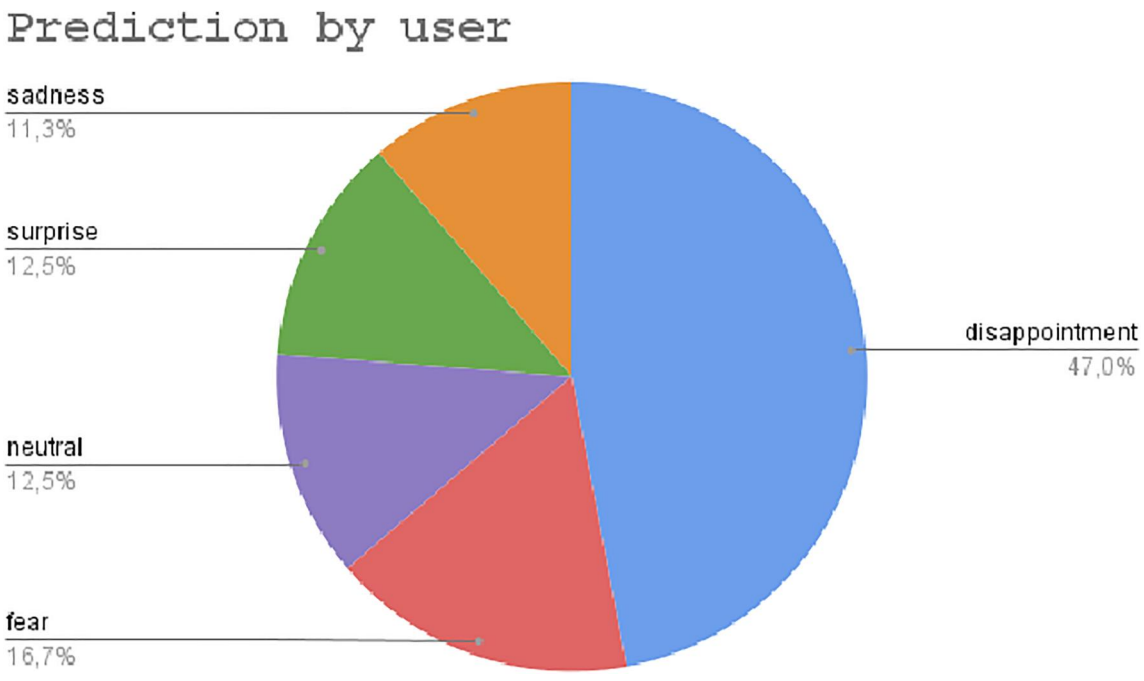


**FIGURE 6** Pie chart of emotions by the user obtained through the mode of all predictions made for that user.

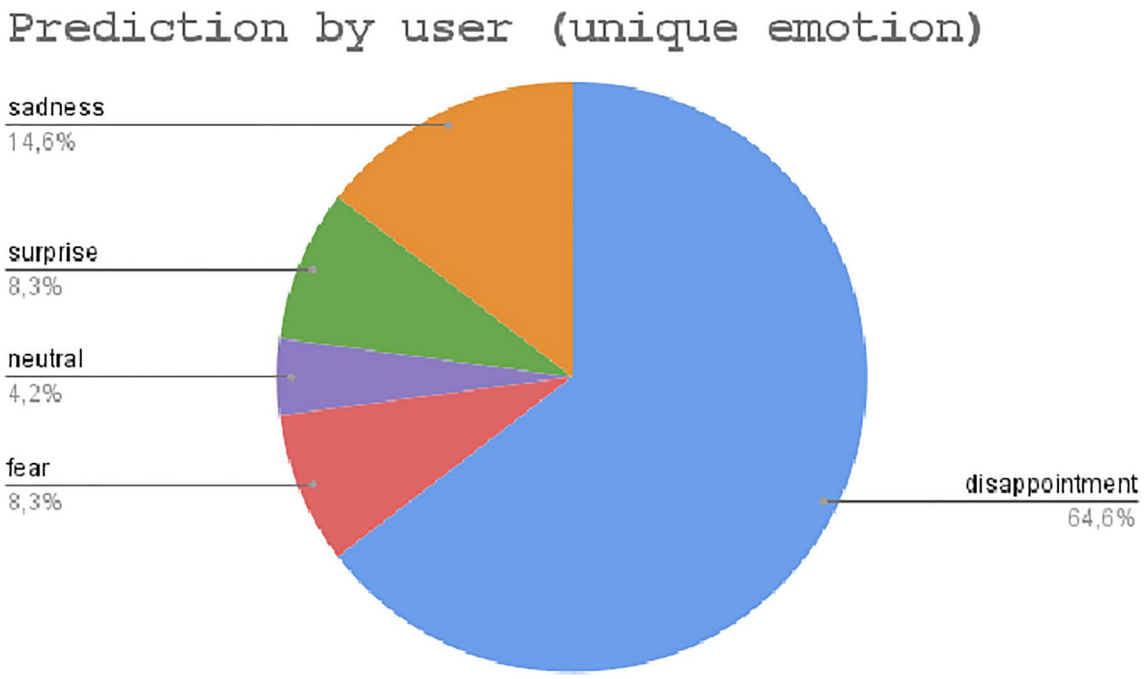


**FIGURE 7** Pie chart of emotions per user for which a single emotion was predicted throughout the conversation.

Guidelines) represent a specific section of the conversation between the patient and virtual agent. With this classification, it is possible to keep track at a macro level of the evolution of emotions.

The six classes have in fact a precise temporal order within the conversation (e.g. first a clinical interview is carried out with the patient, then the close contacts are traced up to the social interview and the remainder of the instructions to follow during the quarantine).

The 7th class, on the other hand, is 'Repeat.' Sometimes, in fact, it happens that the bot is unable to correctly isolate the answer given by the user and therefore rephrases the question. We therefore decided to enter this category to see if repeating the question significantly alters the patient's emotional state, highlighting a clear disservice to be solved by the company.

We present the results in Figure 8. In the stacked bar chart, each column is a question class. Each bar is also divided into five distinct parts and the five different colours that represent the percentage of predicted emotions for each class. The columns are arranged in chronological order with the class 'Repeat' left for last. The analysis shows how the emotional state of disappointment prevails over the others; however, this should not be interpreted exclusively as dissatisfaction with the service, but more generally as a more decisive tone of voice.

In fact, it is typical to respond to a virtual agent articulating the words well in order to be sure that they have been understood.

This can be found above in the 'Telephone number' and 'Name of the contact' columns which are the key questions of the service and in which the degree of disappointment rises while sadness decreases. Similar to the considerations made for 'disappointment,' 'sadness' identifies a lower tone of voice that could also indicate a certain level of boredom and little involvement in the conversation.

Moreover, it is interesting to see the behaviour of 'surprise' and 'fear' that rise in percentage terms, especially towards the end of the conversation.

In particular, when the bot asks the user if there have been other contacts outside the family environment and if he wants to listen to the instructions to follow during the quarantine again.

While in the first case, it is immediate to understand the reason for the surprise or fear (i.e. either because the user does not have an immediate answer or he is afraid of putting other people in unpleasant situations), in the case of the quarantine indications, the result can be explained by the fact that the bot after a long series of indications in which the patient could get distracted, he is asked if he wants to listen to the indications again, finding him sometimes displaced. Finally, as regards the 'Repeat' class, it is noted that the results are in line with the average values achieved for the other classes.

This means that there is insufficient evidence to show that repeating a response adversely alters the patient's emotional state.

### 5.3.3 | Empirical analysis

By empirical analysis, we mean an analysis carried out by directly listening to some of the audio samples and indicating whether the emotion predicted by the model is actually detectable by the human ear. Specifically, we sampled 15 random audio samples for each emotional state for a total of 75 plays.

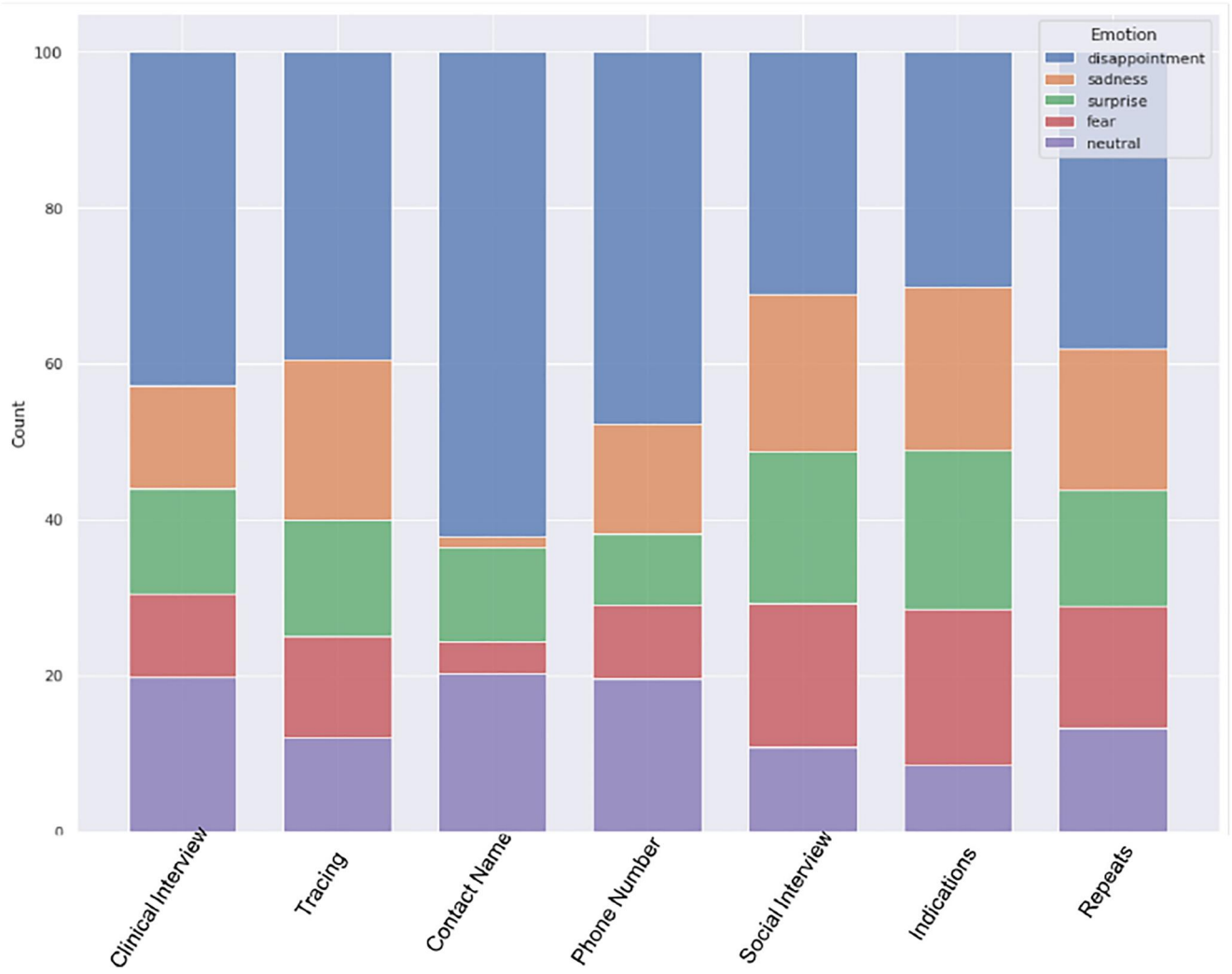


**FIGURE 8** Stacked bar chart coloured by emotion with questions on the *x*-axis and the percentage count on the *y*-axis.

Listening to the audio samples, due to the short and predefined answers, it was not possible to clearly distinguish between emotions.

However, one could hear tonal differences between the various classes as hypothesised previously (Subsection 5.3.2). For example, 'disappointment' generally appeared to have a more decisive tone than sadness.

Therefore it can be concluded from the analysis made that the model shows some promising results about its correct functioning but some structural problems such as the guided and brief answers that the user is forced to give to the bot do not allow to fully evaluate the performance of the model. In this regard, it is recommended to create sections of the call where the user can express himself more freely.

This can help identify the patient's emotional state more accurately.

Furthermore, the analysis could be improved by sending users a questionnaire aimed at tracking their emotional state. If we cross-reference the results of the questionnaire and the predictions made by the model, we could achieve better results and more specific insights.

## 6 | CONCLUSIONS

In this work, we present a neural network-based model to predict emotions in a chatbot developed for contact tracing during COVID-19. The importance of the work presented in this manuscript is manifold.

In particular, considering the emotions recognised by the system, the folowing are taken.

- Provide further assistance to patients with 'at risk' emotional states and in need of psychological support. For example, if fear or sadness is detected, it can be decided to entrust the case to a human operator who is more specialised in the treatment of specific patients.
- Adapt the language and the type of questions asked by the virtual agent based on the patient's emotional state. At present, the flow of the conversation and the questions asked by the virtual agent follow a tree pattern dependent on the patient's responses. This pattern can be redesigned based on the emotions identified during the conversation.

- More specific questions can be created to assess the patient's emotional state.
- Highlight possible disservices during the call and remedy those. A quality service would have an immediate positive economic impact for the supplier company and its customers and also positive effects on the collective trust in Human-Computer Interaction (HCI) applications.

In conclusion, the sentiment analysis carried out with the methodology proposed by this thesis would greatly help improve the quality of the service which means not only an immediate positive economic impact for the supplier company Blu Pantheon and its customers but also positive effects on the collective trust in Human-Computer Interaction (HCI) applications.

HCI solutions oriented towards sentiment analysis assumed increasing importance during the COVID-19 pandemic crisis, where authorities mandated both public and private organisations to embrace new practices for working remotely and maintaining social distancing.

The model implemented showed very promising results. An in-depth analysis of patients' emotional states during the call can, in fact, help the company significantly improve the service and provide an extremely important decision support system tool.

Future work may include the collection of new audio samples, and therefore, possible implementation of deep learning and multi-modal approaches for the identification of emotions in voice-bot conversation.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Code available at: https://github.com/fp1acm8/SER Data available upon requests to the authors.

## ORCID

*Giovanna Maria Dimitri* https://orcid.org/0000-0002-2728-4272

## REFERENCES

1. Covid-19 OED Online (2021, September). Oxford University Press. https://www.oed.com/viewdictionaryentry/Entry/88575495 (2021). Accessed September 2021
2. Page, J., Hinshaw, D., McKay, B.: Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market–The man with the first confirmed infection of the new coronavirus told the WHO team that his parents had shopped there. Wall St. J. (2021). https://www.wsj.com/articles/in-hunt-for-covid-19-origin-patient-zero-points-to-second-wuhan-market-11614335404
3. Ministero della Salute FAQ - Covid-19 domande e risposte. https://www.salute.gov.it (2021)
4. blupantheon.com (2022)
5. Bianchini, M., et al.: Deep neural networks for structured data. In: Computational Intelligence for Pattern Recognition, pp. 29–51. Springer (2018)
6. Bishop, C.M., Nasrabadi, N.M.: Pattern Recognition and Machine Learning, vol. 4. Springer (2006)
7. Flach, P.: Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge university press (2012)
8. Dimitri, G.M., et al.: Unsupervised stratification in neuroimaging through deep latent embeddings. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1568–1571. IEEE (2020)
9. Miconi, F., Dimitri, G.M.: A Machine Learning Approach to Analyse and Predict the Electric Cars Scenario: The Italian Case. Plos one (2023)
10. Zhang, J., et al.: Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. Inf. Fusion 59, 103–126 (2020). https://doi.org/10.1016/j.inffus.2020.01.011
11. Rekanar, K., et al.: Sentiment analysis of user feedback on the HSE's Covid-19 contact tracing app. Ir. J. Med. Sci. 1971(1-10), 103–112 (2022). https://doi.org/10.1007/s11845-021-02529-y
12. Ouellet, S.: Real-time emotion recognition for gaming using deep convolutional network features. arXiv preprint arXiv:14083750 (2014)
13. Oh, K.J., et al.: A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In: 2017 18th IEEE International Conference on Mobile Data Management (MDM), pp. 371–375. IEEE (2017)
14. Kalantarian, H., et al.: Labeling images with facial emotion and the potential for pediatric healthcare. Artif. Intell. Med. 98, 77–86 (2019). https://doi.org/10.1016/j.artmed.2019.06.004
15. Braun, M., Weber, F., Alt, F.: Affective automotive user interfaces–Reviewing the state of emotion regulation in the car. arXiv preprint arXiv:200313731 (2020)
16. Petrushin, V.: Emotion in speech: recognition and application to call centers. In: Proceedings of Artificial Neural Networks in Engineering, vol. 710, pp. 22 (1999)
17. Ziser, Y., Kravi, E., Carmel, D.: Humor detection in product question answering systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 519–528 (2020)
18. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif. Intell. Rev. 43(2), 155–177 (2015). https://doi.org/10.1007/s10462-012-9368-5
19. Cao, H., et al.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. 5(4), 377–390 (2014). https://doi.org/10.1109/taffc.2014.2336244
20. Livingstone, S.R., Russo, F.A.: The Ryerson audio-visual database of emotional speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One 13(5), e0196391 (2018). https://doi.org/10.1371/journal.pone.0196391
21. Jackson, P., Haq, S.: Surrey Audio-Visual Expressed Emotion (Savee) Database. University of SurreyGuildford, UK (2014)
22. Dupuis, K., Pichora-Fuller, M.K.: Toronto Emotional Speech Set (Tess)-younger Talker_happy (2010)
23. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Comput. Humanit. 42(4), 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6
24. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
25. Larranaga, P., et al.: Machine learning in bioinformatics. Briefings Bioinf. 7(1), 86–112 (2006). https://doi.org/10.1093/bib/bbk007
26. Spiga, O., et al.: Machine learning application for patient stratification and phenotype/genotype investigation in a rare disease. Briefings Bioinf. 22(5), bbaa434 (2021). https://doi.org/10.1093/bib/bbaa434
27. Sebe, N., et al.: Machine Learning in Computer Vision, vol. 29. Springer Science & Business Media (2005)
28. Vea, C., et al.: Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease. Faseb. J. 33(11), 12696–12703 (2019). https://doi.org/10.1096/fj.201901529r
29. Dimitri, G.M., et al.: Modeling brain–heart crosstalk information in patients with traumatic brain injury. Neurocritical Care 36(3), 738–750 (2022). https://doi.org/10.1007/s12028-021-01353-7
30. Soleymani, M., et al.: A survey of multimodal sentiment analysis. Image Vis Comput. 65, 3–14 (2017). https://doi.org/10.1016/j.imavis.2017.08.003

31. Selvaraj, M., Bhuvana, R., Karthik, S.P.: Human speech emotion recognition. Int. J. Eng. Technol. 8, 311–323 (2016)
32. Shaikh Nilofer, R., et al.: Automatic emotion recognition from speech signals: a Review. Int. J. Sci. Eng. Res. 6(4) (2015)
33. Lieskovská, E., et al.: A review on speech emotion recognition using deep learning and attention mechanism. Electronics 10(10), 1163 (2021). https://doi.org/10.3390/electronics10101163
34. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Commun. 41(4), 603–623 (2003). https://doi.org/10.1016/s0167-6393(03)00099-2
35. Shen, P., Changjun, Z., Chen, X.: Automatic speech emotion recognition using support vector machine. In: Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, vol. 2, pp. 621–625. IEEE (2011)
36. Shaw, A., Vardhan, R.K., Saxena, S.: Emotion recognition and classification in speech using artificial neural networks. Int. J. Comput. Appl. 145(8), 5–9 (2016). https://doi.org/10.5120/ijca2016910710
37. Huang, Z., et al.: Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 801–804 (2014)
38. Kwon, S., et al.: MLT-DNet: speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Syst. Appl. 167, 114177 (2021). https://doi.org/10.1016/j.eswa.2020.114177
39. Baevski, A., et al.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 33, 12449–12460 (2020)
40. Hsu, W.N., et al.: Hubert: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio, Speech, and Lang. Proces. 29, 3451–3460 (2021). https://doi.org/10.1109/taslp.2021.3122291
41. Castellano, G., De Carolis, B., Macchiarulo, N.: Automatic facial emotion recognition at the COVID-19 pandemic time. Multimed. Tool. Appl., 1–19 (2022). https://doi.org/10.1007/s11042-022-14050-0
42. Yang, B., Jianming, W., Hattori, G.: Face mask aware robust facial expression recognition during the COVID-19 pandemic. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 240–244. IEEE (2021)
43. Scarpina, F.: Detection and recognition of fearful facial expressions during the coronavirus disease (COVID-19) pandemic in an Italian sample: an online experiment. Front. Psychol. 11, 2252 (2020). https://doi.org/10.3389/fpsyg.2020.02252
44. Huang, Z., et al.: Public perception of the use of digital contact-tracing tools after the COVID-19 lockdown: sentiment analysis and opinion mining. JMIR Format. Res. 6(3), e33314 (2022). https://doi.org/10.2196/33314
45. Srikanth, J., et al.: Sentiment analysis on COVID-19 Twitter data streams using deep Belief neural networks. Comput. Intell. Neurosci. 2022, 2022–11 (2022). https://doi.org/10.1155/2022/8898100
46. Alhuzali, H., et al.: Emotions and topics expressed on Twitter during the COVID-19 pandemic in the United Kingdom: comparative geolocation and text mining analysis. J. Med. Internet Res. 24(10), e40323 (2022). https://doi.org/10.2196/40323
47. Choudrie, J., et al.: Applying and understanding an advanced, novel deep learning approach: a Covid 19, text based, emotions analysis study. Inf. Syst. Front 23(6), 1431–1465 (2021). https://doi.org/10.1007/s10796-021-10152-6
48. www.braincontrol.eu (2022)
49. Costantini, G., et al.: EMOVO corpus: an Italian emotional speech database. In: International Conference on Language Resources and Evaluation (LREC 2014), pp. 3501–3504. European Language Resources Association (ELRA) (2014)
50. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Commun. 40(1-2), 5–32 (2003). https://doi.org/10.1016/s0167-6393(02)00071-7
51. Chapelle, O., et al.: Choosing multiple parameters for support vector machines. Mach. Learn. 46(1), 131–159 (2002). https://doi.org/10.1023/a:1012450327387
52. Chen, T., et al.: Xgboost: extreme gradient boosting. R Package Version 04-2 1(4), 1–4 (2015)
53. Dimitri, G.M., et al.: Multimodal and multicontrast image fusion via deep generative models. Inf. Fusion 88, 146–160 (2022). https://doi.org/10.1016/j.inffus.2022.07.017
54. Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. IEEE Transact. Neural Networks Learn. Syst. 32(2), 604–624 (2020). https://doi.org/10.1109/tnnls.2020.2979670
55. Goldberg, Y.: A primer on neural network models for natural language processing. J. Artif. Intell. Res. 57, 345–420 (2016). https://doi.org/10.1613/jair.4992
56. Nogueira, K., Penatti, O.A., Dos Santos, J.A.: Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn. 61, 539–556 (2017). https://doi.org/10.1016/j.patcog.2016.07.001


## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Pucci, F., Fedele, P., Dimitri, G.M.: Speech emotion recognition with artificial intelligence for contact tracing in the COVID-19 pandemic. Cogn. Comput. Syst. 1–15 (2023). https://doi.org/10.1049/ccs2.12076