PTSD in the wild: a video database for studying post-traumatic stress disorder recognition in unconstrained environments

Moctar Abdoul Latif Sawadogo, Furkan Pala, Gurkirat Singh, Imen Selmi, Pauline Puteaux, Alice Othmani

PTSD in the Wild: A Video Database for Studying Post-Traumatic Stress Disorder Recognition in Unconstrained Environments

Moctar Abdoul Latif Sawadogo, Furkan Pala, Gurkirat Singh, Imen Selmi, Pauline Puteaux, and Alice Othmani

***Abstract*—POST-traumatic stress disorder (PTSD) is a chronic and debilitating mental condition that is developed in response to catastrophic life events, such as military combat, sexual assault, and natural disasters. PTSD is characterized by flashbacks of past traumatic events, intrusive thoughts, nightmares, hypervigilance, and sleep disturbance, all of which affect a person's life and lead to considerable social, occupational, and interpersonal dysfunction. The diagnosis of PTSD is done by medical professionals using self-assessment questionnaire of PTSD symptoms as defined in the Diagnostic and Statistical Manual of Mental Disorders (DSM). In this paper, and for the first time, we collected, annotated, and prepared for public distribution a new video database for automatic PTSD diagnosis, called PTSD in the wild dataset. The database exhibits "natural" and big variability in acquisition conditions with different pose, facial expression, lighting, focus, resolution, age, gender, race, occlusions and background. In addition to describing the details of the dataset collection, we provide a benchmark for evaluating computer vision and machine learning based approaches on PTSD in the wild dataset. In addition, we propose and we evaluate a deep learning based approach for PTSD detection in respect to the given benchmark. The proposed approach shows very promising results. Interested researcher can download a copy of PTSD-in-the wild dataset from: http://www.lissi.fr/PTSD-Dataset/**

***Index Terms*—Affective computing in the wild, PTSD diagnosis, Video analysis, Mental disorder diagnosis, Affect recognition**

I. Introduction

POST-traumatic stress disorder (PTSD) is a chronic mental condition resulting after exposure to a traumatic event. These events generally involve direct threat, or represent a real risk of death or serious injury like natural disasters, sexual assault, military combat experience or even during a major stressful life experience like divorce or unemployment. The symptoms of PTSD are divided into three symptom clusters: reexperiencing, avoidance, and hyperarousal. In addition, trauma survivors often experience guilt, dissociation, alterations in personality, difficulty with affect regulation, and marked impairment in ability for intimacy and attachment. Disorders comorbid with PTSD include depression, substance abuse, other anxiety disorders, and a range of physical complaints. Over the past several decades, considerable progress has been made in the development and empirical evaluation of assessment instruments for measuring trauma exposure and PTSD as well as related syndromes, such as acute stress disorder. The measures that have been developed, including self-report questionnaires, structured interviews [1], and psychophysiological procedures [2], have been extensively validated and many have been widely adopted internationally.

Traditionally like most other debilitating psychological disorders, it was diagnosed by medical professionals in a clinical setting. Diagnosing the ailment would involve a comprehensive psychological evaluation comprising questionnaires and multiple in-person interview sessions. Occasionally blood tests are an additional requirement for filtering out the specific psychological affection. Collecting information through a self-report has limitations. People are often biased when they report on their own experiences. And these self-reports are subject to several biases and limitations, such as the introspective ability: the subjects may not be able to assess themselves accurately or the rating scales bias : Rating something yes or no can be too restrictive, but numerical scales also can be inexact and subject to individual inclination to give an extreme or middle response to all questions.

Early and automatic detection of PTSD can help reduce its effect. Hence, the interest in early detection and the audio-visual mechanisms involved in post-traumatic stress detection. To the best of our knowledge, no public video dataset exists for PTSD recognition and prediction. To meet this need, We present in this paper, a new publicaly available dataset for academics to facilitate the development of new approaches and tools for the automatic diagnosis of post traumatic stress disorder. This dataset as well as the proposed deep learning based approach for PTSD recognition will serve as a valuable starting point for scholars interested in affective computing of mental disorders and audiovisual data to investigate and to study PTSD.

The rest of this paper is organized as follows. In Section II, we detail current state-of-the-art methods related to PTSD diagnosis from clinical videos and existing video datasets. Section III describes the new video dataset we constructed,

M.A.L. Sawadogo is with LISSI laboratory of University Paris-Est Créteil and University Paris City in France. E-mail: moctar-abdoul-latif.sawadogo@u-pec.fr

F. Pala is with LISSI laboratory of University Paris-Est Créteil in France and Istanbul Technical University (ITU) in Turkey. E-mail: pala18@itu.edu.tr

G. Singh is with LISSI laboratory of University Paris-Est Créteil in France and the National Institute of Technology Warangal in India. E-mail: gs832025@student.nitw.ac.in

I. Selmi is with LISSI laboratory of University Paris-Est Créteil in France and The National Engineering School of Sfax (ENIS) in Tunisia. E-mail: imen.selmi@enis.tn

P. Puteaux is with the French National Centre for Scientific Research (CNRS) in France. E-mail: pauline.puteaux@cnrs.fr

A. Othmani is with LISSI laboratory of University Paris-Est Créteil. E-mail: alice.othmani@u-pec.fr

This research work is realized in the Laboratoire Images, Signaux et Systèmes Intelligents (LiSSi)- EA-395, Université Paris-Est Créteil (UPEC), Corresponding author: Dr. Alice Othmani, E-mail: alice.othmani@u-pec.fr

named PTSD-in-the-wild dataset. In Section IV, we give explanations on the used baseline models (audio, visual and text). Section V presents experimental results and finally, this paper is concluded in Section VI.

II. Related works and databases

In this section, we present related works and databases for PTSD and stress diagnosis. In Section II-A, we first detail current state-of-the-art PTSD diagnosis from clinical videos methods. In Section II-B, we describe existing video datasets, which were all acquired in lab-controlled conditions.

A. PTSD diagnosis from clinical videos

In recent years, Artificial Intelligence (AI)-based systems have become immensely popular for psychological diagnosis due to their scalability, flexibility and convenience. A system for the mental screening of military personnel – where audio interviews were fed into the *Sensibility technology ST Emotion* framework for emotion recognition – has been proposed in [7]. Soldiers who spent more time on a particular mission had higher stress levels and lower happiness compared to those who served for a shorter time.

In [8], audio-visual features comprising of face, voice, speech content and movement were extracted from medical interview recordings and fed into a neural network for PTSD and MDD (Major Depressive Disorder) prediction. It was found that textual features contributed the most to an accurate prediction. In [9], only the audio data was used. To compensate for the small sized PTSD dataset that was available, they decided to use transfer learning. A deep belief network was initially trained on audio data to learn general voice features after which fine-tuning was performed on the PTSD dataset.

The accuracy of audio-based PTSD classifiers could be increased by using multi-view learning, as shown in [10]. The classifier itself was trained on EEG (electroencephalography) and audio data, but used only audio signals for prediction during inference. A prediction on audio data is also performed in [11]. After preprocessing, prosodic, excitation and vocal tract features were extracted and fed to an architecture based on extreme gradient boosting with teamwork-based optimization for prediction. In [12], it was shown that elicitation of multimodal neurophysiological responses to audio-visual stimuli was a suitable alternative to using clinical interviews as input for PTSD diagnosis. Neurophysiological responses comprising of ECG (electrocardiography), EEG, GSR (galvanic skin response), head motion, video, speech, along with the emotions evoked were fed into an SVM classifier. A comprehensive study on post-traumatic neuropsychiatric sequelae is proposed in [5]. It provided a huge multimodal dataset on which PTSD models could be trained.

Owing to the overlap between PTSD and the symptoms of other mental disorders, it is important to go over the work done on the latter as well. In [13], audio-visual data was used along with the text. The model utilized a DCNN and DNN based architecture to get predictions from audio-video modalities, while paragraph vector, SVM and random forest were utilized to predict from the textual input. Fusion using a multivariate regression model fused the data from all three modalities for the final prediction.

B. Existing video datasets

Few video datasets for PTSD and stress diagnosis exist, as shown in Table I. There are only three datasets with audio data for PTSD recognition: eDAIC-WOZ [4], FEMH (Far Eastern Memorial Hospital) [3] and Aurora [5]. Moreover, these datasets were all acquired in lab-controlled conditions.

a) eDAIC-WOZ [4]: It is an extended version of DAIC-WOZ dataset for detection of PTSD from an interview without any human intervention. This is an audio visual dataset that was also used for the AVEC 2019 Challenge. The dataset contains clinical interviews which were collected as part of a larger effort to create a computer agent which identifies verbal and non verbal signs of mental illness. It contains 189 interviews ranging between 7-33 minutes. This dataset is made publicly available only for academics researchers under request and signature of an End User License Agreement (EULA).

b) FEMH dataset (Far Eastern Memorial Hospital) [3]: It consists of voice signals from 200 patients labeled as normal (50), neoplasm (40), phonotrauma (60) and vocal palsy (50). These voice signals contain 3-second sustained /a:/ sounds, with a sampling rate of 44.1KHz and a 16-bit resolution. The voice recording lengths are between 5 seconds to 40 seconds.

c) Aurora dataset [5]: RecOvery afteR traumA (AURORA) dataset conducts a large scale (n = 5,000 target sample) in-depth assessment of adverse post-traumatic neuropsychiatric sequelae (APNS) development. These APNS include traditionally categorized outcomes such as PTSD, depression, MTBI (minor traumatic brain injury), and regional or widespread pain. The 5,000 patients were screened, recruited and receive initial baseline evaluation, including blood collection and psychophysical, survey and neurocognitive evaluation. They were closely monitored over the next 8 weeks using a wrist wearable, a smart phone app and daily flash surveys weekly, web-based neurocognitive tests, periodic mixed-mode surveys, serial saliva collection, deep phenotyping blood collection, fMRI (functional magnetic resonance imaging), psychophysical evaluation and then followed less intensively using similar procedures including deep phenotyping over the remainder of a 52-week follow-up period.

Two others video datasets exist for recognizing stress and positive/negatives emotions. We notice that these two datasets could be interesting for transfer knowledge from stress and negative emotions domain to stress post-traumatic domain. These two datasets are : Ulm-TSST and MuSe-CaR datasets.

d) Ulm-TSST [6]: This dataset contains about 10 hours long audio visual, as well as biological recordings such as ECG, EDA (electrodermal activity), respiration and heart rate. It consists of a richly annotated dataset of self reported and external dimensional ratings of emotion and mental well being. It has videos of 105 participants out of which 69.5% are female which are aged between 18 to 39 years. This dataset is available online for researchers who fulfil the requirements of the EULA.

Ref.	Dataset	Modalities	Public
[3]	FEMH (Far Eastern Memorial Hospital) Dataset	Audio	NO
[4]	eDIAC-WOZ	Audio + visual features	only for academic research under request
[5]	Aurora	Audio + text	NO
[6]	Ulm-TSST (Muse challenge)	Video + ECG + EDA + Respiration + Heart Rate	only for academic research under request

TABLE I: Existing audio and video datasets for stress and PTSD diagnosis.

e) MuSe-CaR [14]: It is a large multimodal dataset which contains about 303 videos, *i.e.* over 40 hours of user-generated video material. This dataset mainly focuses on how positive and negative sentiments as well as emotional arousal are linked to the person. This dataset is extensively annotated and is available online for researchers upon request.

III. PTSD IN-THE-WILD DATASET COLLECTION

In this section, we describe a new video dataset called PTSD in-the-wild dataset for studying PTSD in unconstrained environments. In Section III-A, we explain the procedure of videos collection from the web. In Section III-B, the videos annotation process is detailed and the categories of trauma and their distribution are presented.

A. Videos collection from the web

The procedure of collecting this dataset was realized semi-automatically. It was done first by downloading more than 1,000 videos from YouTube and then, by choosing the videos of interviews to predict whether a person has a PTSD or not.

In our search, we used keywords such as: "PTSD", "Understanding PTSD", "my story with PTSD", "celebrities Answers the Web's Most Searched Questions", "interviews with celebrities", *etc*. For data with PTSD, we selected a group of people who had experienced a traumatic event like an accident or a terrorist attack. We collected this data from news, entertainment and medical channels on YouTube like Make the connection[1], Veterans Health Administration[2] and GQ[3]. For data without PTSD, we selected a group of celebrities and artists who have been interviewed.

All these videos are in English. Moreover, there are different age groups and genders in this data, as well as nationalities. The average video duration is about 2 minutes. In most of the videos, the interviewee speaks more than the interviewer. There are also videos without any interviewer: only an interviewee talks and answers questions.

In Fig. 1, as examples, we present 4 pictures extracted from videos of our dataset. Pictures depicted in Fig. 1.a and Fig. 1.b are taken from videos of people with PTSD, while those in Fig. 1.c and Fig. 1.d show people without PTSD. One can see from Fig.2b that the two genders are represented both in the With PTSD and Without PTSD classes. In Fig. 1, the interviewees with PTSD seem to be troubled, whereas it is not the case of the interviewees that do not have PTSD. Note also that, within the 'With PTSD' class examples, the PTSD disorder originates from two different types of trauma: war veteran (Fig. 1.a) and plane crash (Fig. 1.b).

[1]https://www.youtube.com/c/VeteransMTC
[2]https: //www.youtube.com/c/VeteransHealthAdmin
[3]GQ: https://www.youtube.com/c/GQ

B. Annotation

The videos were annotated by human annotators. Four people participated to the annotation of the dataset. After extraction, each video was watched by an annotator and different labels were given. Apart from the class label, other categories of information were collected. These are the class, the gender and the type of developed trauma. In the paragraphs below, we explain how the dataset was annotated and how the categories were determined.

1) Class label: Videos were classified into 'With PSTD' and 'Without PTSD'. Videos labelled as 'With PTSD' are about people who developed PTSD after a traumatic event and they are talking about their trauma during an interview. Videos labelled as Without PTSD are interview videos about random topics which are not PTSD-related. The videos are labelled by taking into account the speech of the interviewee and the description of the video. If the interviewee mentions that he was diagnosed with PTSD or it is mentionned in the video's description that the interviewee has PTSD, then the video is labelled as 'With PTSD'. On the other hand, if neither the speech nor the description is about PTSD or a trauma, the video is labelled as 'Without PTSD'. The dataset contains 634 videos equally balanced: there are 317 videos of subjects with PTSD and 317 videos of healthy control subjects with no PTSD symptoms.

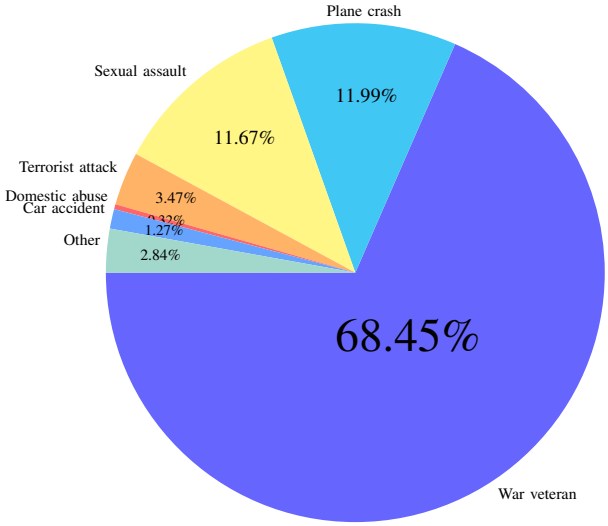
2) Gender: The gender of each interviewee was determined from the video. The considered gender types were Male (M) and Female (F). Some videos have more than one interviewee. Therefore, the number of interviewees exceeds the number of videos. The dataset globally contains 467 male interviewees and 207 female interviewees. There are different gender distributions across the classes With PTSD and Without PTSD. Fig. 2b summarizes the distribution of the gender within the two labels.

3) Type of trauma: The types of trauma were also considered in the annotation of the dataset. A trauma is a response to a terrible event. The type of the trauma concerns the event that caused the trauma. It was determined from the interviewee's speech or from the video's description. Many types were found. They are grouped into the following categories: War veteran, Plane crash, Sexual assault, Terrorist attack, Domestic abuse, Car accident and Other. Fig. 2a shows the count of the different types of trauma in the dataset.

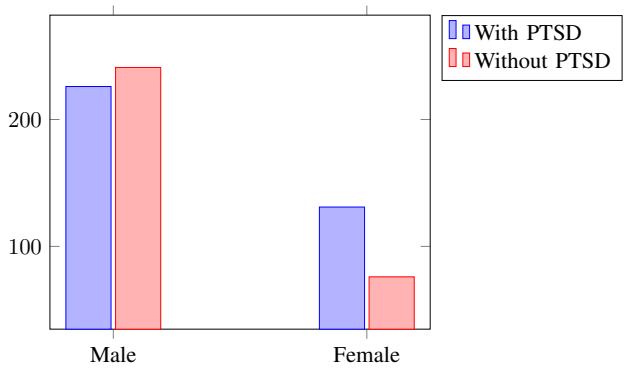
- **War veteran:** This category contains videos of people who have been in the army and have been deployed in hostile territories or war zones. They subsequently developed PTSD at some point during their service or after ending their service. This is the most represented category.

(a) (b) (c) (d)

Fig. 1: Four videos frames from PTSD-in-the-wild dataset of two subjects with PTSD ((a) and (b)) and two subjects without PTSD ((c) and (d)). (a) A woman who had PTSD after war. She was on a mission and suddenly a bomb exploded. She found out that she has PTSD in October 2004. (b) Man who suffered from PTSD after a plane crash. He was the captain of the plane. After the plane crash, his body was damaged a lot and he always felt guilty. (c) Dove Cameron answers the Web's Most Searched Questions with WIRED auto-complete interview. Dove Olivia Cameron is an American singer and actress. She played a dual role as the eponymous characters in the Disney Channel comedy series Liv and Maddie. (d) Ken Jeong answers the Web's Most Searched Questions with WIRED auto-complete interview. Kendrick Kang-Joh Jeong is an American stand-up comedian, actor, producer, writer and licensed physician.



(a) Distribution of the types of trauma.



(b) Distribution of the gender in the dataset. Note that, in some videos With PTSD, there are more than one interviewee. This is why the number of interviewees exceeds the number of videos.

Fig. 2: Distribution of the type of trauma and the gender across the PTSD-in-the-wild dataset. (a) Distribution of the types of trauma. (b) Distribution of the gender in the dataset.

- **Sexual assault:** This category contains videos of people who have endured sexual assaults and harassment. Some of the videos in this category are Military Sexual Trauma (MST), they are people in the military who have suffered sexual assaults from other soldiers.
- **Terrorist attack:** This category gathers the videos of people who have been direct victims of a terrorist attack or have been witnesses of a terrorist attack. Terrorist attacks are events where violence is used to create fear in a population. Most of the time, it involves killing people. In the collection of the dataset, we considered many such events among which school shootings and September 11 attacks.
- **Plane crash:** This videos regroup people who have survived a plane crash or a plane related incident.
- **Domestic abuse:** This category contains videos of people who have developed a trauma from domestic abuse by their partner.
- **Car accident:** This category contains videos of people who have been victims or witnessed a terrible car accident which led them to develop PTSD.
- **Other:** This category regroups videos of people whose type of trauma could not be classified in any of the aforementioned categories. This contains categories like physical abuse, gang violence and videos of people who have suffered several and repeated traumatic events. Furthermore, people whose type of trauma could not be determined are classified in this category.

IV. Baseline

In this section, we present our baseline for PTSD diagnosis in unconstrained environments, using our proposed PTSD in-

the-wild dataset. In Section IV-A, we explain how we built sets for training, validation and test. Then, we describe the models we proposed, based on different modalities: audio (Section IV-B), visual (Section IV-C) and text (Section IV-D).

A. Training, validation and testing sets

Our dataset consists of 317 videos of subjects with PTSD and 317 videos of healthy control subjects with no PTSD symptoms. Two benchmarks are proposed for the evaluation of new approaches for PTSD diagnosis and recognition. Then, for splitting the dataset into training, validation and testing sets, we used the two strategies or partitions:

a) Train/Validation/Test split: We randomly split the dataset into training, validation and testing sets with the percentages of 80%, 10% and 10%, respectively. Distribution of the two classes of With and Without PTSD in the three sets of training, validation and test can be seen in the Table II: among the 317 samples for each class, we used 253 videos for training, 32 for validation and 32 for testing.

Split	PTSD samples	NO PTSD samples
Training	253	253
Validation	32	32
Testing	32	32

TABLE II: Distribution of the classes in the splits.

b) N-fold cross validation: To further evaluate the effectiveness of our baseline models and any new approach for video PTSD diagnosis, we used n-fold cross validation strategy with $n = 3$ for the splitting of training and testing sets. By doing so, we ensure that our obtained results can be generalized to an independent subset of data. We also prevent overfitting and selection bias.

B. Audio baseline model

Audio can be a powerful asset to identify emotions or affects from the speech [15]. More recently, several approaches have been proposed for mental disorders diagnosis based on audio data [16], [17], [18], [19]. Audio signal has been also used in trauma prediction [9].

Therefore, we propose a baseline model for binary audio classification into With PTSD and Without PTSD. The model used for audio classification is based on the VGGish model [20]. Our actual implementation is based on the Keras implementation[4] of the original VGGish model. The VGGish model is an adaption of VGG16 for audio files. Our starting point is the original VGGish model, pre-trained on the Audio Set dataset [21].

The model architecture consists of four blocks of convolution layers. Each block is followed by a max-pooling layer with a kernel size of 2×2 and a stride of 2×2. The first block consists of a single convolutional layer which outputs 64 feature maps. The second block has a single convolutional layer with an output of 128 feature maps. The third block comprises two convolutional layers with 256 output feature maps each. The third block is also composed of two convolutional layers that both output 512 feature maps. All convolutional layers have a convolutional kernel of size 3×3 and a stride of 1×1.

On top of this, for the sake of our work, we replaced the original last three fully connected layers with a global pooling layer and two fully connected layers of 32 units each. The number of units of these layers were found empirically by searching through a set of values. These layers are followed by a dropout layer and a fully connected layer of 2 units with a softmax activation representing the two classes. Fig. 3 shows a representation of the modified VGGish model's architecture.

We pass to the model the log-mel-spectograms of each audio of size 2976×64. These inputs were computed using the original VGGish implementation[5] in a window size of 30 seconds.

C. Visual baseline model

The second proposed baseline model aims to predict PTSD based on the video frames. Video frames contain visual, spatial and temporal information to retrieve in order to extract PTSD patterns to discriminate between healthy and PTSD subjects. The proposed approach is based on PTSD symptoms recognition from patients' responses to questions asked by an interviewer.

In our proposed approach, the first step of the visual-based classification approach consists in pre-processing the data. Video Frames (images) are pre-processed in order to separate the patient's face (interviewee) from that of the interviewer. Indeed, from each video in the dataset, only 60 frames containing face of the interviewee are extracted manually and are considered for training the visual baseline model.

Once interviewee (patient or control subject) frames are selected, the face is extracted using the MTCNN model [22]. Interviewee frames and face selection is an important step in the proposed approach and it impact considerably the performances of the proposed approach for visual PTSD classification.

For our visual baseline model, we used ResNet50v2 [23], originally trained on the ImageNet dataset [24] for extracting meaningful features. Then, the extracted features from the ResNet50v2 are fed to a sequence model consisting of recurrent layers of LSTM [6]. Our hybrid architecture is popularly known as a CNN-RNN architecture.

ResNet50v2 is a modified version of ResNet50 that performs better due to a modification in the propagation formulation of the connections between blocks. Our actual implementation is based on the Keras implementation[7]. The fully connected layer on the top of the model is removed for feature extraction. This technique of training a model from its pre-trained weights is called transfer learning, that focuses on storing knowledge gain, while solving one problem and applying it to a different but related problem. Note that transfer

[4]https://github.com/DTaoo/VGGish

[5]https://github.com/tensorflow/models/tree/master/research/audioset/vggish

[6]https://keras.io/api/layers/recurrent_layers/lstm/

[7]ResNet50v2: https://keras.io/api/applications/resnet/#resnet50v2-function

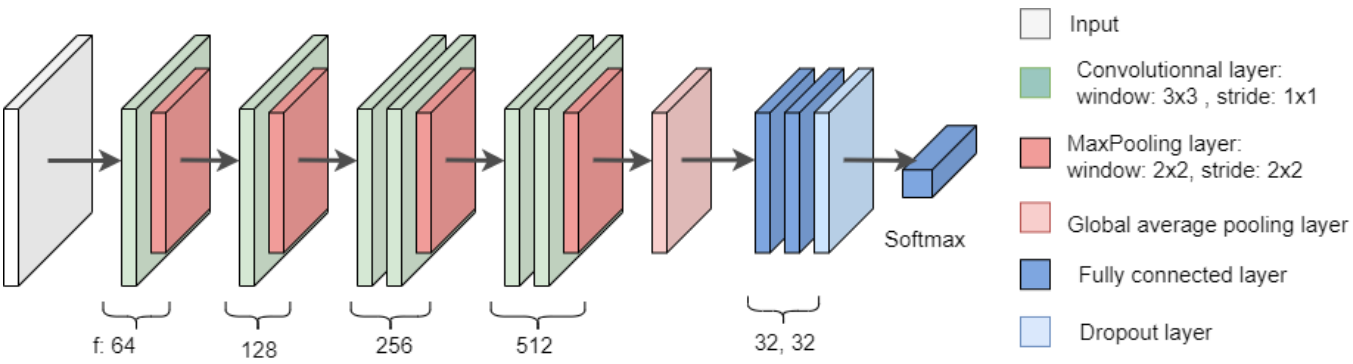


Fig. 3: VGGish-based audio baseline model's architecture. The size of the feature maps (f) of each convolutional and fully-connected layer are shown below each block of operations.

learning is a very useful technique when the size of dataset is small, as in our case.

The resulting vector is passed through a fully connected layer for further classification as With PTSD or Without PTSD related videos. Fig. 4 shows a representation of our proposed visual baseline model's architecture based on Resnet50v2 and LSTM.

D. Text baseline model

Natural language processing (NLP) offers a valuable approach in recognizing the underlying thought in the text by understanding the meanings of individual words as well as the context which has been investigated and harnessed in several fields like sentiment analysis and emotion recognition [25], [26], [27]. Thus, we consider it useful to utilize the NLP methods and tools in classifying the individuals by whether they have PTSD or not based on their transcriptions. Since we are working with videos, we need to : (1) extract the audio signal and (2) to convert it to transcribed text before (3) being fed to an automatic speech recognition (ASR) model. Our text classification network uses BERT [28] language model as the backbone to get the contextualized word embeddings which are then fed into a fully connected classification layer. We detail the ASR and text classification pipelines in the following sections.

1) Automatic Speech Recognition: Classical ASR approaches rely on a large amount of transcript annotations of the speech audio to train, which is not feasible considering the human effort required.

We used the wav2vec 2.0 [29] architecture as our speech-to-text converter which uses a self-supervised approach to train on a limited amount of labelled data. As shown in Fig. 5.A, the model is composed of three blocks : **i) Feature Encoder** is a multi-layer convolutional neural network that learns the latent audio representations for T time steps from the raw audio. **ii) Context network** is a Transformer based architecture that takes the output of the feature encoder and generates T representations carrying the information from the whole audio sequence. As in masked language modelling [28], a ratio of the learned speech representations coming out of the feature encoder are randomly masked before feeding into the context network. To discretize the learned speech representations, **iii) a quantization module** is used by taking the output of the feature encoder with no masking as input and choosing a speech unit for each representation. The model jointly learns the quantized speech units and contextualized latent representations which boosts the performance significantly.

In the pre-training phase, the aforementioned blocks are trained in an end-to-end manned on the unlabeled data. Fine-tuning is done by minimizing the Connectionist Temporal Classification (CTC) loss [30], [31] on the labelled data. Nonetheless, we neither pre-trained nor the fine-tuned the ASR model on our dataset, *i.e.* we used the ASR model just for inference, since we do not have the labels for the ASR task. We just executed inference on the audio files to get the transcribed text using a publicly available pre-trained and fine-tuned wav2vec 2.0 model. Also, we have long videos in our dataset with an average duration of 2 minutes, which makes it not applicable to directly run inference on the whole audio file because of the insufficient GPU memory. Thus, we needed to split the audio into small chunks. However, simply splitting the audio into a fixed duration of chunks may lead to poor results since we can cut in the middle of a spoken word. Although one possible solution is splitting based on the silence periods, it is not practical since we need to determine a threshold for loudness which is not an one-size-fits-all value. Hence, we run inference on partially overlapping chunks so that the model has the correct context focused in the center. The amount of overlapping is defined as stride. Then, we drop the logits on the sides, *i.e.* take the logit at the center. Chaining the remaining logits gives us better results than the naive chunking.

2) Text Classification: Language models (LMs) are powerful NLP tools used in many areas like machine translation, question answering, named entity recognition and text classification due to their ability to understand the underlying abstract meaning of the natural language. However, training such data-hungry models require a large and diverse dataset to grasp the characteristics of the natural language. Trying to train a LM from scratch on a small corpus is most probably going to be not enough for it to generalize well.

BERT [28] model addresses this issue by proposing an architecture pre-trained on a large unlabelled text which can be used on the downstream tasks by fine-tuning with a single layer on top of it. BERT uses a deep bidirectional

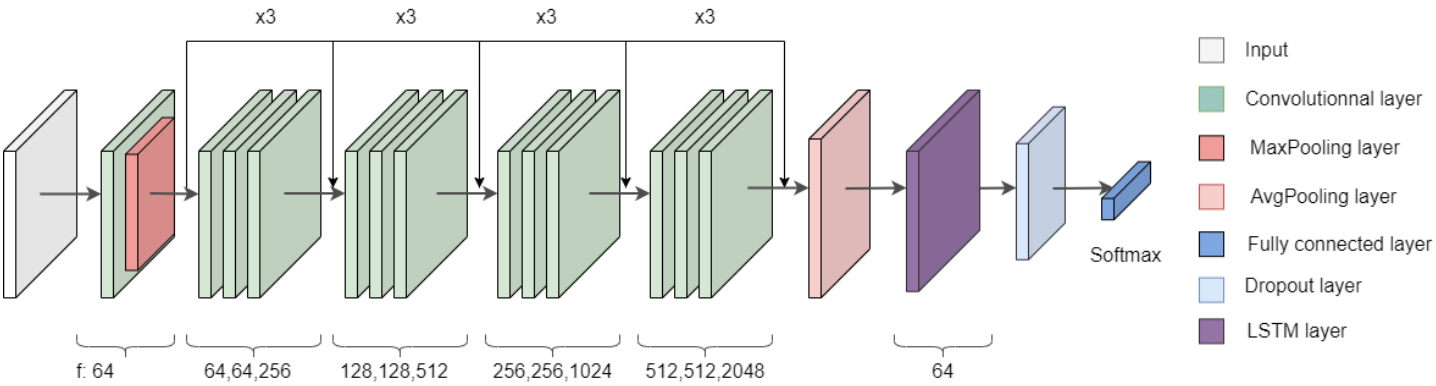


Fig. 4: Resnet50v2 and LSTM-based visual baseline model's architecture. The size of the feature maps (f) of each convolutional and fully-connected layer are shown below each block of operations.

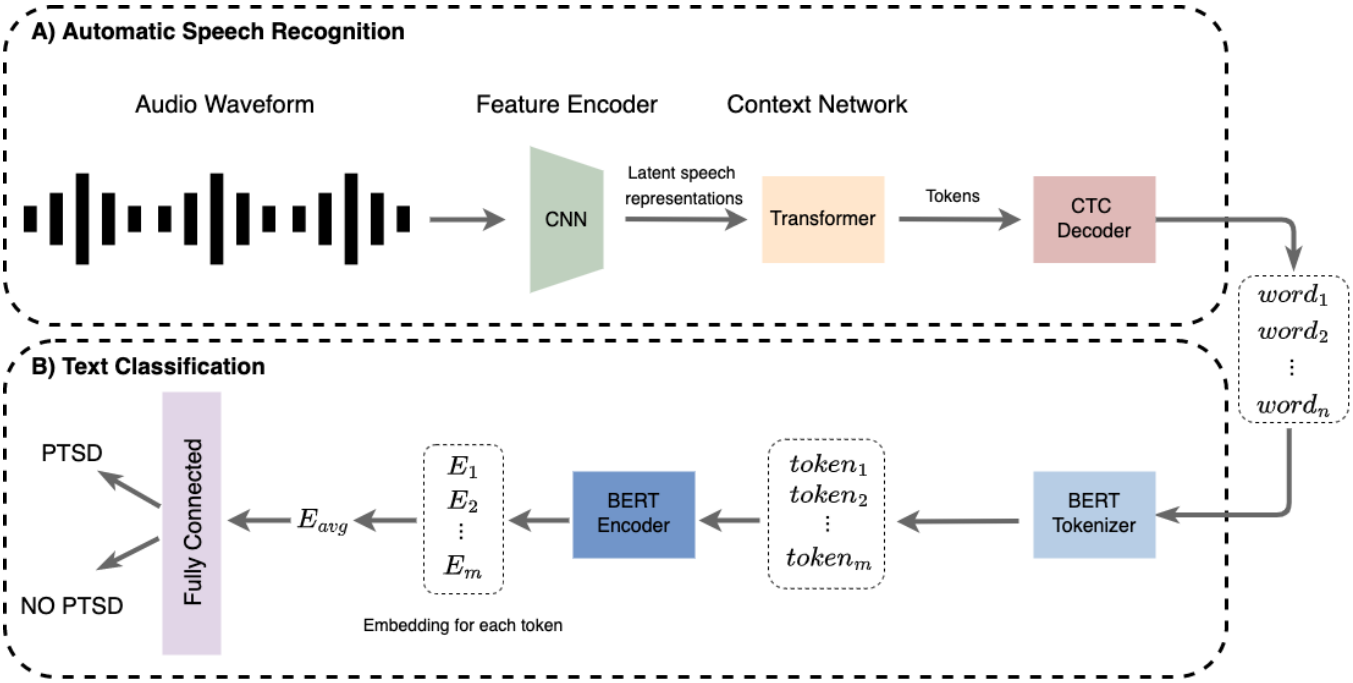


Fig. 5: The pipeline of our text baseline model. **A) Automatic Speech Recognition:** To get the transcript from the audio waveform for each sample in our dataset, we used the wav2vec 2.0 [29] architecture which is composed of a CNN feature extractor that feeds the latent speech representations into a Transformer to obtain the tokens which are then decoded to one of the 29 letters plus a space character using a CTC decoder. **B) Text Classification:** Given the plain text transcript of a sample, our aim is to classify it by PTSD or NO PTSD via the textual meaning. Thus, we first tokenize the text using the BERT tokenizer then we feed the tokens into the BERT encoder to get the embeddings for each token. We get an average text embedding for that sample by computing the mean of embeddings over the token dimension. Using one fully connected layer on top of BERT encoder handles the binary classification task.

Transformer architecture to jointly train on two objectives: masked language modelling (MLM) and next sentence prediction. It utilizes a WordPiece tokenizer which allows BERT to recognize relevant words since they may share the same tokens. BERT model outputs a contextualized word embedding vector for each input token, *i.e.* the representation for each token is built based on its surrounding tokens, thus the same words with the different meanings can be distinguishable.

We used the BERT language model for text classification in our text baseline (Fig. 5.B). We tokenize the transcribed speech of each sample obtained from the ASR using the BERT tokenizer that outputs a sequence of tokens. Each sequence starts with `[CLS]` and ends with `[SEP]` tokens. We use `[PAD]` tokens to pad the sequence to a constant length of 512. We truncate the longer sequences. Then, we use the base BERT encoder to get the contextualized embeddings for each token. We get the text embedding by averaging the embeddings over the token dimension. We feed the text embedding into a fully-connected layer that acts as a binary classifier.

Hyperparameters	Set of considered values	Optimal
Optimizer	Adam, Adamax, Adagrad, Adadelt, SGD	**Adam**
Learning rate	1e-05 to 1e-02	**3e-05**
Units	32, 40, 48, 56, $\cdots$, 248, 256	**32**
Dropout rate	0.1, 0.2, 0.3, 0.4, 0.5	**0.4**
Kernel constraint	0.1, 0.2, 0.3, $\cdots$, 0.9, 1.0, 2.0, 3.0	**3.0**
Batch size	8, 16, 24, 32, $\cdots$, 120, 128	**128**
Epochs	1, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50, $\cdots$, 100	**100**

(a) Audio baseline model.

Hyperparameters	Set of considered values	Optimal
Optimizer	Adam, Adamax, Adagrad, Adadelta, SGD, RMSprop, Nadam, Ftrl	**SGD**
Learning rate	3e-04, 1e-03, 1e-02, 1e-01	**3e-04**
LSTM units	2, 4, 8, 16, 32, 64	**64**
Dropout rate	0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5	**0.5**
Momentum	0.0, 0.1, 0.2, 0.9	**0.0**

(b) Visual baseline model.

TABLE III: Hyperparameters search and optimization using KerasTuner for the audio and visual baseline models.

V. Results on the baseline model

In this section, we present the results of the experiments performed using the baseline models on the PTSD in-the-wild dataset. In Section V-A, we describe the metrics used to evaluate the performances of our deep learning based baseline models. In Section V-B, we give some details on the implementation of our models, with a specific interest to the used parameters. Experimental results obtained with our audio baseline model, visual baseline model and text baseline model are described in Section V-C, Section V-D and Section V-E respectively.

A. Evaluation metrics

We are interested in evaluating a binary classification task. There are two classes: With PTSD and Without PTSD. As we have to test the presence of a PTSD, we define the TP (true positives), FN (false negatives), TN (true negatives) and FP (false positives) values as such:

- TP is the number of samples labelled With PTSD and predicted as With PTSD,
- FN is the number of samples labelled With PTSD and predicted as Without PTSD,
- TN is the number of samples labelled Without PTSD and predicted as Without PTSD,
- FP is the number of samples labelled Without PTSD and predicted as With PTSD.

From the confusion matrix, we compute the different popular classification metrics : the accuracy, the precision, the recall and the F1-score. The different metrics are described in the following.

Accuracy is the fraction of correctly classified samples and computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision indicates the strength of the model to classify a negative sample as non-positive, which can be computed as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall indicates the ability of the model to find all the positive samples and can be computed as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score is the harmonic mean of precision and recall, thus can be computed as:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

or also:

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

B. Implementation details

Finding the best configuration of the hyperparameters of deep neural networks – both related to the model and to the training – is a computationally expensive task. In fact, the validation error function is very costly to evaluate for each hyperparameter value in order to find the best value by solving an optimization problem. This function is often not convex: it is then impractical to find its global minimum.

a) Parameters setting of the audio baseline model: To find the best hyperparameters, we used a bayesian optimizer built-in KerasTuner [32]. KerasTuner is a tool for hyperparameters search and optimization. Bayesian optimization uses principles based on Bayes theorem to efficiently and effectively conduct a search of global optimization problem. We describe the list of hyperparameters and the sets of considered values in Table III.a. The optimal hyperparameters are presented in the last column. Then, we selected the Adam optimizer [33] with a learning rate of 3e-05. The model has a dropout rate of 0.4 to prevent overfitting after layers which have 32 units. The kernel constraint is equal to 3.0. We consider a batch size of 128 samples and the number of epochs is set to 100.

We also had to prepare the audio files for the training. To train the model, we extracted audio files from the videos. The classification into With PTSD and Without PTSD through audio can be sensitive to the interference of the interviewer's

speech and to background sound. For this reason, we dropped some audios of the two categories of type of trauma: Plane crash and Terrorist attack videos because the speech of the interviewer interfered a lot with the interviewee's speech and the background sound makes some audios not clearly audible. Moreover, the audios were split into chunks of 30 seconds. Thus, we dropped the chunks that do not contain the speech of the interviewee and those in which the interviewer speaks the most.

In addition, we expand the labeled audio training sets by using data augmentation techniques. Three types of audio augmentation techniques are then performed on the audio frames to perturb the raw audio signals and generate new ones:

- **Noise injection**: a random noise is added to the audio signal. The noise is multiplied by an α factor;
- **Pitch change**: The pitch of the audio is changed by a factor 0.5, 2 and 3 in semi-tones;
- **Time shifting**: The samples are moved forward or backward by a given number of seconds.

b) Parameters setting of the visual baseline model: We also used KerasTuner to find the best hyperparameters, but using the built-in Random Search algorithm as a tuner. The Random Search algorithm bases its optimization strategy on a stochastic process. The list of hyperparameters and their sets of considered values are presented in Table III.b. The optimal hyperparameters can be found in the last column. Then, we selected the SGD (Stochastic Gradient Descent) optimizer with a learning rate of 3e-04 and a momentum of 0. The model has a dropout rate of 0.5 to prevent overfitting after LSTM layers and its 64 units.

c) Parameters setting of the text baseline model: For each video in the dataset, we extracted the audio and then run inference using the base wav2vec 2.0 model that is pre-trained and fine-tuned on the 960 hours of Librispeech speech audio (wav2vec2-base-960h). The chunk length is 60 seconds and the stride is 10 seconds. We have written each transcript to the disk. For the text classification, we used the pre-trained base uncased BERT model for English (bert-base-uncased) to create the contextualized representations. Just before the fully-connected layer, we added a dropout layer [34] with a dropout rate of 0.1 to prevent the overfitting. We fine-tuned the BERT on the aforementioned text classification task for 5 epochs using the AdamW [35] optimizer with the learning rate of 5e-05, betas of (0.9, 0.999) and weight decay of 0.01. For single train/validation/test split methodology, we monitored the validation F1-score during the training and chose the model with the highest validation F1-score for testing. In N-Fold cross-validation, we trained for 3 folds as well and evaluated the model using the left-out testing set as usual.

C. Results obtained using the audio baseline model

In order to evaluate how the model performs with the PTSD-in-the-wild dataset, we trained the model using the training, validation and testing sets for 100 epochs.

The classification results on the unseen testing set are presented in Table IV (first row). The accuracy is equal to 0.93. Moreover, a 0.93 F1-score value indicates that the trade-off between precision (equal to 0.94) and recall (equal to 0.93) is interesting. Thus, the audio baseline model achieves a good performances in the classification task, and it is very accurate in distinguishing between With PTSD and Without PTSD samples.

In addition, we performed a 3-fold cross-validation to validate the results. The results are very close to the previous results. Table V summarizes the results of the cross-validation. The model has an average accuracy of 0.88, with a standard deviation of 0.05. It performed better on the second and the third folds, with results similar to those obtained using the testing set (as presented in Table IV). The overall results show that the audio baseline model is able to learn and to generalize from a new dataset.

D. Results obtained using the visual baseline model

The classification results using the visual baseline model on the testing set are presented in Table IV (second row). With a value of 0.82, we can see that the accuracy is slightly worse than the value obtained using the audio baseline model. It is still interesting, as well as the value of the F1-score: they both remain high, which shows good classification performances.

Moreover, the 3-fold cross validation results are provided in Table V. The visual baseline model has an average accuracy of 0.84 and performed better on the first fold. Nevertheless, whatever the considered fold, the results are similar to those obtained using the testing set (as presented in Table IV). These results prove the generalization ability of the visual baseline model.

E. Results obtained using the text baseline model

In Table IV (third row), we report the classification results obtained using the text baseline model on the testing set. Being equal to 0.98, the accuracy is very close to the maximum value. Moreover, the F1-score is equal to 0.98. It indicates that the trade-off between precision (equal to 1.00) and recall (equal to 0.97) is excellent.

To further see the generalizability of the text baseline model, we also reported the 3-fold cross validation results in Table V. The text baseline model has an average accuracy of 0.98. The performances are almost the same on every fold and comparable to those using the testing set (as presented in Table IV): the classification is nearly perfect. Therefore, the text baseline model outperforms the two other baseline models using audio or visual information to distinguish between With PTSD and Without PTSD samples. The text is the best modality among the three evaluated modalities in PTSD recognition and detection.

VI. Discussion and conclusion

The analysis of human mental health and disorders is a very complex, challenging and sensitive problem. The lack of understanding of mental illness and the prejudice and the negative beliefs towards it have caused universally the stigma of mental illness. Consequently, a more attention is required

Baseline model	Accuracy	Precision	Recall	F1-score
Audio	0.93	0.94	0.93	0.93
Visual	0.82	0.82	0.82	0.82
Text	0.98	1.00	0.97	0.98

TABLE IV: Classification results on the testing set as a function of the used baseline model.

		Accuracy	Precision	Recall	F1-score
Audio baseline model	Fold 1	0.80	0.80	0.80	0.80
	Fold 2	0.93	0.93	0.93	0.93
	Fold 3	0.92	0.92	0.92	0.92
	Average	**0.88**	**0.88**	**0.88**	**0.88**
Visual baseline model	Fold 1	0.88	0.88	0.88	0.88
	Fold 2	0.82	0.82	0.82	0.82
	Fold 3	0.82	0.82	0.82	0.82
	Average	**0.84**	**0.84**	**0.84**	**0.84**
Text baseline model	Fold 1	0.99	0.98	0.99	0.98
	Fold 2	0.97	0.96	0.99	0.97
	Fold 3	0.97	0.97	0.97	0.97
	Average	**0.98**	**0.97**	**0.98**	**0.98**

TABLE V: 3-fold cross-validation classification results as a function of the used baseline model.

from the scientific community to join forces to improve mental health and to propose innovative solutions for better and more precise diagnosis, prognosis and assistance of the patients. The development of innovative clinical applications for mental disorders diagnosis is plagued by ethical and privacy issues. However, artificial intelligence (AI) research for psychiatry and mental health faces several strict limitations : (1) the lack and the small size of the available datasets. (2) the limited available data and modalities and (3) the absence of audiovisual data that could be relevant for the diagnosis. To overcome these problems, AI can be used as a tool to guarantee the privacy and the protection of sensitive data without being a threat [36].

Artificial intelligence and affective computing based tools and systems can play a key role in the diagnosis of several mental disorders, the follow-up of mental state of patients and the development of personalized medicine in psychiatry. One of the main limitations is the lack of datasets and cohorts. In this paper, we proposed and we made publicly available for academic research a new dataset called PTSD-in-the-wild dataset for trauma and post-traumatic stress disorder recognition and analysis from video data. The dataset is collected in unconstrained environments and conditions. It means that the video are acquired in real-world scenarios without any lab-controlled conditions. To be able to cope with such challenging conditions, the proposed approaches should be robust to indoor, outdoor, different color backgrounds, occlusions, background clutter and face misalignment. The proposed dataset can be used for binary learning and classification of PTSD. However, it can not be used for PTSD assessment as no self-assessment questionnaire like the Post-traumatic Stress Disorder Checklist (PCL) test is performed. The PCL-5 is a 20-item, widely used DSM-correspondent self-report that assesses the 20 DSM-5 symptoms of PTSD.

In this paper, we propose a baseline models for PTSD recognition using the PTSD-in-the-wild dataset. Two benchmarks are given for the evaluation and the comparison of proposed approaches on the PTSD-in-the-wild dataset. The visual, audio and textual baseline models show very promising and high performing results. However, our baseline visual model is very dependant on the manual frame selection of the interviewee. The performance of the audio baseline model strongly dependant on the drop of the two categories of plane crash and Terror attack videos because of the strong intervention of the interviewer during the interview and the bad quality of the video and the high ratio signal to noise. The three modalities of facial images, audio signals and speech are very accurate to recognize and to discriminate PTSD patterns during interviews sessions of healthy control subjects and PTSD subjects who experienced a traumatic event. The high performances of the baseline models can be explained by the big difference between the content of the two classes: (1) the first class of interviews of individuals who experienced a traumatic event and who have PTSD symptoms and talking about the traumatic event and showing severe emotional distress or physical reactions when reminding it and (2) the second class of interviews of random topics with actors and public figures, showing a positive affects, thoughts and attitudes. The question arises: Is it possible to discriminate between a PTSD person and a person with a negative affect (e.g. a sad person) or a person with anxiety disorder or also a depressed person ?

In future work, we are planned: (I) to propose an automatic approach to perform the frame selection and interviewee face detection and interviewee audio signal extraction to make our approach fully automatic and (II) to propose a second version of the PTSD-in-the-wild dataset with other low mental mood and affects conditions and disorders.

REFERENCES

[1] E. de Beurs, K. Thomaes, H. Kronemeijer, and J. Dekker, "[the PTSD checklist for DSM-5 (PCL-5): comparing responsivity with the outcome questionnaire (OQ-45) and practical utility]," *Tijdschr Psychiatr*, vol. 62, no. 6, pp. 448–456, 2020.

[2] M. R. Bauer, A. M. Ruef, S. L. Pineles, S. J. Japuntich, M. L. Macklin, N. B. Lasko, and S. P. Orr, "Psychophysiological assessment of PTSD: a potential research domain criteria construct," *Psychol Assess*, vol. 25, pp. 1037–1043, July 2013.

[3] K. A. Islam, D. Perez, and J. Li, "A transfer learning approach for the 2018 femh voice data challenge," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5252–5257, IEEE, 2018.

[4] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, *et al.*, "The distress analysis interview corpus of human and computer interviews," tech. rep., UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, 2014.

[5] S. A. McLean, K. Ressler, K. C. Koenen, T. Neylan, L. Germine, T. Jovanovic, G. D. Clifford, D. Zeng, X. An, S. Linnstaedt, *et al.*, "The aurora study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure," *Molecular psychiatry*, vol. 25, no. 2, pp. 283–296, 2020.

[6] L. Stappen, E.-M. Meßner, E. Cambria, G. Zhao, and B. W. Schuller, "Muse 2021 challenge: Multimodal emotion, sentiment, physiological-emotion, and stress detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5706–5707, 2021.

[7] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *2011 Defense Science Research Conference and Expo (DSR)*, pp. 1–5, IEEE, 2011.

[8] K. Schultebraucks, V. Yadav, A. Y. Shalev, G. A. Bonanno, and I. R. Galatzer-Levy, "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," *Psychological Medicine*, vol. 52, no. 5, pp. 957–967, 2022.

[9] D. Banerjee, K. Islam, K. Xue, G. Mei, L. Xiao, G. Zhang, R. Xu, C. Lei, S. Ji, and J. Li, "A deep transfer learning approach for improved post-traumatic stress disorder diagnosis," *Knowledge and Information Systems*, vol. 60, no. 3, pp. 1693–1724, 2019.

[10] X. Zhuang, V. Rozgić, M. Crystal, and B. P. Marx, "Improving speech-based ptsd detection via multi-view learning," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 260–265, IEEE, 2014.

[11] S. Gupta, L. Goel, A. Singh, A. K. Agarwal, and R. K. Singh, "Toxgb: Teamwork optimization based xgboost model for early identification of post-traumatic stress disorder," *Cognitive Neurodynamics*, pp. 1–14, 2022.

[12] V. Rozgic, A. Vazquez-Reina, M. Crystal, A. Srivastava, V. Tan, and C. Berka, "Multi-modal prediction of ptsd and stress indicators," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3636–3640, IEEE, 2014.

[13] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pp. 45–51, 2017.

[14] L. Stappen, A. Baird, L. Schumann, and B. W. Schuller, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *CoRR*, vol. abs/2101.06053, 2021.

[15] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021.

[16] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103107, 2022.

[17] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. Othmani, "Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis," *Machine Learning with Applications*, vol. 2, p. 100005, 2020.

[18] M. Muzammel, H. Salam, and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106433, 2021.

[19] A. Othmani, D. Kadoch, K. Bentounes, E. Rejaibi, R. Alfred, and A. Hadid, "Towards robust deep neural networks for affect and depression recognition from speech," in *International Conference on Pattern Recognition*, pp. 5–19, Springer, 2021.

[20] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135, IEEE, 2017.

[21] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780, IEEE, 2017.

[22] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, pp. 630–645, Springer, 2016.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 2009.

[25] L. Abualigah, H. E. Alfar, M. Shehab, and A. M. A. Hussein, "Sentiment analysis in healthcare: a brief review," *Recent Advances in NLP: The Case of Arabic Language*, pp. 129–141, 2020.

[26] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6351–6358, 2013.

[27] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (New York, NY, USA), p. 369–376, Association for Computing Machinery, 2006.

[31] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," 2019.

[32] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, "Kerastuner." https://github.com/keras-team/keras-tuner, 2019.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017.

[36] R. Kusters, D. Misevic, H. Berry, A. Cully, Y. Le Cunff, L. Dandoy, N. Díaz-Rodríguez, M. Ficher, J. Grizou, A. Othmani, *et al.*, "Interdisciplinary research in artificial intelligence: Challenges and opportunities," *Frontiers in Big Data*, vol. 3, p. 577974, 2020.

Moctar Abdoul Latif Sawadogo is currently a Master student at University Paris City in France. He was a master intern at Laboratory of Images, Signals and Smart Systems of University Paris-Est Créteil (UPEC). His main interests are machine learning and deep learning.

Furkan Pala graduated from the Istanbul Technical University (ITU) Computer Engineering bachelor program in 2022. He is currently pursuing a double major degree in Mathematics Engineering at ITU. He published two conference papers about medical imaging accepted in MICCAI 2021 and 2022 during his internship at Brain And SIgnal Research & Analysis (BASIRA) laboratory, ITU. His work focuses on deep learning, computer vision, medical imaging and NLP.

Gurkirat Singh is currently pursuing Bachelor in Technology at National Institute of Technology, Warangal, India majoring in Computer Science. His main research interests include but are not limited to Robotics, Deep Learning and Computer Vision.

Imen Selmi is a computer science and applied Mathematics engineering student at The National Engineering School of Sfax (ENIS), University of Sfax. She worked on many projects in deep learning and computer vision and she was a master intern at Laboratory of Images, Signals and Smart Systems of University Paris-Est Créteil (UPEC).

Pauline Puteaux received her M.S. degree in Computer Science and Applied Mathematics with specialization in Cybersecurity from the University of Grenoble, France, in 2017 and her PhD degree in computer science from the Université de Montpellier, France, in 2020. She is currently working as a researcher for the CNRS (French National Centre for Scientific Research) with the Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL), France. Her work has focused on multimedia security, and in particular, image analysis and processing in the encrypted domain. Since 2016, she has published eight journal articles and thirteen conference papers. She is a reviewer for Signal Processing (Elsevier), the Journal of Visual Communication and Image Representation (Elsevier), the IEEE Transactions on Circuits & Systems for Video Technology, and the IEEE Transactions on Dependable and Secure Computing. Photo credit: © Xavier PIERRE / CNRS

Alice Othmani is an associate professor at Université Paris-Est Créteil since 2017. Her research works concern developing computer vision and Artificial Intelligence solutions for Healthcare, emotional intelligence and psychiatry. She has been working in several international institutions like Ecole Normale Supérieure de Paris, Collège de France and Agency for Science, Technology and Research (A*STAR) in Singapore.