Expert Systems With Applications 189 (2022) 116076

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Audio based depression detection using Convolutional Autoencoder

Sara Sardari[a,1], Bahareh Nakisa[b,*,2], Mohammed Naim Rastgoo[c,3], Peter Eklund[b,4]

[a] *Computer Science, Engineering and IT Department, Shiraz University, Shiraz, Iran*
[b] *School of Information Technology, Faculty of Science Engineering and Built Environment, Deakin University, Vic, Australia*
[c] *School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD, Australia*

ARTICLE INFO

Keywords:
Audio depression detection
Semi-supervised learning
Convolutional Autoencoder
Early depression detection

ABSTRACT

Depression is a serious and common psychological disorder that requires early diagnosis and treatment. In severe episodes the condition may result in suicidal thoughts. Recently, the need for building an effective audio-based Automatic Depression Detection (ADD) system has sparked the interest of the research community. To date, most of the reported approaches to recognize depression rely on hand-crafted feature extraction for audio data representation. They combine wide variety of audio-related features to improve the classification performance. However, combining many hand-crafted features including relevant and less-relevant can enlarge the feature space which can lead to high-dimensionality issues as not all the features would carry significant information regarding depression. Having high number of features can make the pattern recognition more difficult and increase the risk of overfitting. To overcome these limitations, an audio-based framework of depression detection which includes an adaptation of a deep learning (DL) technique is proposed to automatically extract the highly relevant and compact feature set. This proposed framework uses an end-to-end Convolutional Neural Network-based Autoencoder (CNN AE) technique to learn the highly relevant and discriminative features from raw sequential audio data, and hence to detect depressed people more accurately. In addition, to address the sample imbalance problem we use a cluster-based sampling technique which highly reduces the risk of bias towards the major class (non-depressed). To evaluate the performance and effectiveness of the proposed pipeline, we perform the experiments on Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) dataset and compare them with the hand-crafted feature extraction methods and other outstanding studies in this domain. The results show that proposed method outperforms other well-known audio-based ADD models with at least 7% improvement in F-measure for classifying depression.

1. Introduction

Depression is a serious medical condition which is globally common. According to World Health Organization (WHO) (2017) more than 300 million people around the world are suffering from depression. In another study, the WHO states that unipolar depression will be the second most significant cause of disability worldwide by 2030 (Mathers & Loncar, 2006). This mental disorder can cause feelings of sadness, sleep disorder, appetite fluctuations, difficulty participating in social activity, and in the most severe cases suicide (World Health Organization, 2017). Although there are effective treatments for depression, one of the major problems faced is that depression assessment is vulnerable to several factors. The severity of depression is usually assessed in a face-to-face clinical interview and questionnaires. The effectiveness of this evaluation is highly prone to the experience of the interviewer, quality of question protocol, and patient's willingness to honestly respond (Zhang et al., 2019). This raises the need for an effective, reliable, and accessible ADD system to assist interviewers. This tool can also be significant in telehealth, a platform designed for a long-distance contact between patient and health care provider (Palylyk-

* Corresponding author.
E-mail addresses: sara.sardari@shirazu.ac.ir (S. Sardari), Bahar.nakisa@deakin.edu.au (B. Nakisa), mohammadnaim.rastgoo@qut.edu.au (M.N. Rastgoo), peter.eklund@deakin.edu.au (P. Eklund).

[1] ORCID: 0000-0002-0042-5928.
[2] ORCID: 0000-0003-2211-2997.
[3] ORCID: 0000-0002-0657-0878.
[4] ORCID: 0000-0003-2313-8603.

https://doi.org/10.1016/j.eswa.2021.116076
Received 26 April 2021; Received in revised form 20 September 2021; Accepted 10 October 2021
Available online 20 October 2021

Colwell & Argáez, 2018). It allows the health care provider to monitor and remind them in certain critical situations, and generally deliver remote healthcare to patients (Palylyk-Colwell & Argáez, 2018). Providing an ADD system and embedding it in different telecommunication technologies and applications such as Skype and FaceTime, not only may help the healthcare provider in real-time depression detection, but it also can provide a self-awareness for the patient to identify and help prevent severe cases of depression.

There are many studies which have analyzed different bio-signals for automatically detecting a mental disease in individuals such as Parkinson and Alzheimer (Vásquez-Correa et al., 2018; Venugopalan et al., 2021). However, ADD is a new emerging research area which has not been investigated compared to other emotion and disease detection areas. Previous studies have shown that depression affects behavioral features in people differently. For example, patients with depression usually exhibit abnormal facial expressions and impaired verbal productivity compared to their normal non-depressed self. Interviewers may also determine visual symptoms, downward angling of the head, downward eye gaze, varied duration of smiles, and other gestures (Qureshi et al., 2019). More importantly, patient speech not only contains linguistic contents, but it also exhibits acoustic features that have a direct relationship with the severity of different kind of mental disorders (Vázquez-Romero & Gallardo-Antolín, 2020). Therefore, many studies have been conducted to evaluate the correlation between mental health and individual vocal characteristics. These studies include detection of Parkinson (Braga et al., 2019; Zlotnik et al., 2015), Alzheimer (Gosztolya et al., 2019; Lopez-de-Ipina et al., 2018) and depression (Cummins et al., 2015), which is highly correlated to this research's objective.

The emergence of machine learning algorithms has been widely deployed to provide automatic mental disease detection tools as an adjunct to medical expert decision-making. In this vein, since 2011, the "Audio-Visual Emotion Challenges" (AVEC) competition has been held with the aim of evaluating different machine learning methods for automatic audio-visual health and emotion detection. Each competition is designed for a specific mental disorder, examples being bipolar (Ringeval et al., 2018) and depression (Valstar et al., 2016; Ringeval et al., 2017; Valstar et al., 2013, 2014). Some of the tasks may include specific sub-challenges and researchers are invited to propose novel solutions. One of the most popular sub-challenges is Depression Classification Sub-Challenge (DCC) (Valstar et al., 2016). This challenge involves proposing a classifier model distinguishing depressed and non-depressed speakers using three main modalities: audio, visual, and textual (a transcript of interview).

More specifically, several datasets for different spoken languages such as German, Turkish and English were introduced specifically for depression problem (Gratch et al., 2014; Ozkanca et al., 2018; Valstar et al., 2014). The German database as a part of the AVEC 2014 challenge (Valstar et al., 2014) consists of conversations with 84 patients in German. According to (Demiroglu et al., 2020) some of the patients have multiple recordings. They were given a Beck Depression Inventory-II (BDI-II) (Beck et al., 1996). There is another dataset for depression detection in Turkish which was collected at a hospital in Istanbul with 70 subjects (Ozkanca et al., 2018). The patients in this study were also given a BDI-II score for depression assessment. It is worth to mention that both datasets were collected for the aim of depression-level assessment and not the depression detection. Another public dataset to mention is the Black Dog dataset (McIntyre et al., 2009) which includes recordings of audio/video from 40 speakers with a major depressive disorder, rated using the Quick Inventory of Depressive Symptomatology- Self Report (QIDS-SR) scale (Cohn et al., 2018).

Specifically, the aim of our research is to propose an audio-based ADD system based on the dataset provided by this sub-challenge, called DAIC-WOZ (Gratch et al., 2014), which comprises interviews between participants (patients) and a virtual questioner. As we are aiming to develop a generalized depression detection model, we use this dataset that is spoken in English and contains more sample data (higher number of participants compared to other dataset).

There are several studies based on both visual and vocal features for developing a depression detection system. For example, Yang et al. (2017) proposed a hybrid depression classification and estimation method using the fusion of audio, video and textual information and the experiments are carried out on DIAC-WOZ dataset. The study utilized a multivariate regression model fusing the depression degree estimations extracted from each modality. They used the text modality for developing depression/non-depression classification system using Paragraph Vector (PV), Support Vector Machine (SVM), and Random Forest where the whole system resulted in 0.667F-measure for depressed class in the development set. In other research, Al Hanai et al. (2018) proposed a depression detection model based on the sequences of audio and textual transcriptions of the speakers using the same dataset. The performance of a multi-modal Long Short-Term Memory (LSTM) model that combined the features extracted from both text and audio are evaluated. This model yielded the performance of 0.77F-measure score. Yang et al. (2016) proposed a gender specific Decision Tree classifier based on visual and vocal features. Their work was the winner of the AVEC-2016 challenge compared to the baseline model provided by the challenge and achieved an F1 score of 0.57 for depressed class in testing set. Despite the fact that fusion of different modalities has been shown to be an efficient and effective way to build a reliable depression detection system, this approach affects the acceptability of the system in real-world situations. Therefore, developing a depression detection system based on a single modality has received much attention in recent years. Because of the aforementioned reasons, we take the approach of building a depression detection system based on the single modality of audio.

To develop a depression detection system, we need to follow three phases: preprocessing, feature extraction, and classification. One of the major tasks in the preprocessing phase using DAIC-WOZ dataset is applying a technique to address the unbalanced data problem in depression detection, since depression occurs less frequently among the whole population. There are different studies that applied preprocessing techniques like random sampling to overcome the problem of unbalanced data problem (Ma et al., 2016; Vázquez-Romero & Gallardo-Antolín, 2020). However, both of the undersampling and oversampling as random sampling strategies have limitations. In random undersampling, some of the useful information from the majority class might be eliminated. In random oversampling the likelihood for overfitting in the model construction process increases (Lin et al., 2017). The aim of a clustering-based resampling (undersampling) is to group similar data samples from the majority class into different clusters. Therefore, the cluster centroid (or center) can be used to represent the data in the whole group. Thereby, we are reducing the size of the majority class without losing any useful information from this class (Lin et al., 2017). Hence, we are using the same strategy in the preprocessing phase of audio.

Another challenging step in the pipeline of audio-based ADD is feature extraction. In fact, the performance of the model significantly depends on the quality of extracted features (Nakisa, Rastgoo, Rakotonirainy, Maire, & Chandran, 2020). Therefore, it is important to find the feature set which consists of the most relevant features. Two major approaches for feature extraction are hand-crafted feature extraction and using DL for automatic feature extraction. Hand-crafted features can be extracted either by using special algorithms or utilizing the knowledge of human experts. Most of the studies related to ADD (Pampouchidou et al., 2016; Valstar et al., 2016; Yang et al., 2016) rely on conventional hand-crafted feature extraction methods like COVAREP (*Cooperative Voice Analysis Repository for Speech Technologies*). There are two major challenges against hand-crafted technique that hinders it from being a preferable feature extraction method. First, taking advantage of human expert knowledge for feature extraction can be highly time consuming and needs to have access to the human experts. Second, if we use a general toolkit like COVAREP for feature extraction from audio, we might obtain a high dimensional feature set including relevant and irrelevant/less-relevant features to our specific problem of depression detection. However, in order to mitigate the curse of dimensionality, avoid overfitting, and have better classification result, we should reduce the dimensionality to a feature space that corresponds to the intrinsic representation of data (Van Der Maaten et al., 2009).

With the recent successful results of DL techniques in automatic feature extraction and classification, applied to different challenging problems (Nakisa et al., 2018a, 2018b, 2020; Rastgoo et al., 2018, 2019; Zhang et al., 2017; Mou et al., 2021; Banan et al., 2020; Shamshirband et al., 2019; Fan et al., 2020), it can be inferred that DL techniques are more reliable and effective in extracting highly relevant features compared to other hand-crafted feature extraction methods. Due to these successful performances in different applications, there are some research that developed ADD systems using DL methods. One of the studies in the speech-based depression detection is Depaudionet proposed by Ma et al. (2016). That system is focused on two important challenges of data representation and sample unbalanced issue for the audio files in the DAIC-WOZ dataset. In their study, the audio files were firstly converted to spectrograms (a time–frequency 2D representation) and a deep model based on Convolutional Neural Network (CNN) and LSTM models. To address the unbalanced issue in the dataset, a random sampling strategy is used. This method resulted in 0.52F-measure performance for depressed class detection in development set. Although they have achieved better performance compared to hand-crafted feature extraction methods, they converted the raw audio data into spectrogram data which is time consuming and limits its application to real-world applications.

In other work conducted by Vázquez-Romero and Gallardo-Antolín (2020), an ensemble of CNNs for depression classification is proposed. To address the unbalanced data challenge, a random sampling method is utilized. The authors achieved a notable performance (0.63F-measure) using the proposed method, utilizing the common practice of log-spectrogram representation of audio signals.

To the best of our knowledge, most of the studies (Ma et al., 2016; Vázquez-Romero & Gallardo-Antolín, 2020) focus on converting an audio signal to spectrograms features and then applying deep learning techniques. However, this preprocessing phase is not appropriate in real-time applications because of the computational effort. In addition, according to Chorowski et al. (2019) many of these deep neural networks comprise automatic supervised representation learning that rely on large human labeled datasets, which is not available in many cases. It has been stated (Gogoi & Begum, 2017) that some of the best results in classification often involve an unsupervised pre-training phase, such as unsupervised representation learning followed by a supervised learning phase such as training a classifier, an approach which can be interpreted as a semi-supervised classification technique.

Inspired by all the observations mentioned above, we aim to propose a framework based on DL technique to extract highly relevant and compact set of features from raw audio signals which can be applicable to real-world applications.

To extract highly relevant and compact set of features, an Autoencoder (AE) as an ideal candidate that can be adapted to the ADD problem. AE is a special type of deep neural network and unsupervised learning which aims to reconstruct the input signal in a manner to minimize reconstruction error. The reconstruction error is the difference between the original input vector and the reconstructed output vector (An & Cho, 2015). In other words, the training process of AE is in such a way that the input and output of the network become as close as possible to each other. The AE technique can achieve its aim using two processing phases of encoding and decoding. The encoder component maps the input vector (the features) into a hidden representation (the latent representation) of the input with lower dimensionality. Then a decoder performs the reconstruction transform on the hidden representation to achieve a reconstructed version of the input signal (Ribeiro et al., 2020). The hidden representation of the original input can be interpreted as a compressed feature vector which represents the most important and relevant relationships of the original input features (Ribeiro et al., 2020). Therefore, an AE can be understood as an unsupervised non-linear dimensionality reduction and feature extraction method before applying a classifier. For example, Sahu et al. (2018) investigated the performance of a speech-based Adversarial AE for emotion recognition. High dimensional feature sets from audio files are extracted, and then the extracted features are passed to Adversarial AE to extract hidden representation of each sample. SVM classifier is applied to classify sample into 4 emotional labels of neutral, sad, angry, and happy and they achieved an Unweighted Average Recall (UAR) of 56.38, which is close to the performance of classifier on the original feature space (57.88). In another study, a de-noising AE and SVM were proposed for hidden representation learning and classifying samples into neutral and emotional classes, this resulted in about 2 percent improvement in emotion recognition accuracy (Xia & Liu, 2013).

Utilizing a CNN instead of a simple Dense Neural Network in AE, adds a powerful feature learning ability to the system because of its unique ability of local feature extraction. In other words, CNN can have a local perception of sequential data like signal or text to extract local features using kernels, which simple Artificial Neural Networks are not able to perceive (Wen & Zhang, 2018). Another advantage of CNNs over fully-connected deep architectures is that it can scale well to high dimensional data unlike fully-connected architectures, since there is no dependence to the input dimensionality in the number of free parameters describing their shared weights (Masci et al., 2011). This interesting characteristic of CNN motivates many researchers to utilize CNN AE as a local feature extractor from high dimensional data consisting of spatial information like signal and image. In a recent study, Wen and Zhang (2018) proposed a deep CNN AE model for feature learning from an Electroencephalography (EEG) signal for epilepsy prediction. The experimental results in two EEG datasets showed that the proposed feature learning method performed better compared to other dimensionality reduction methods.

In this study, motivated by the good performance of the CNN AEs in different research contexts, we present a novel end-to-end framework based on CNN AE in audio-based depression detection. Using an end-to-end learning approach, the network is trained from the raw audio signals without any *a priori* feature extraction. Following this phase, a classifier is used to distinguish depressed speakers from non-depressed speakers. The classifiers explored in this study are four typical classifiers of SVM, Multi-Layer Perceptron (MLP), Gradient Boosting (GB), and Random Forest (RF). We compare the performance of CNN AE against hand-crafted features extracted from different studies as well as a simple Dense AE model.

In summary, the contributions of this study are as follows:

- Specifically, we propose an end-to-end framework to automatically extract highly relevant and compact set of features using CNN AE to accurately build an audio-based depression detection model. To the best of our knowledge, there is no other research conducted on developing an automated feature extraction using Convolutional AE on the audio modality for depression detection. One of the important characteristics of the proposed audio-based ADD compared to the previous DL studies is that it is applied on raw audio data without converting them to spectrograms features.
- To overcome the problem of unbalanced data using DAIC-WOZ dataset, in this study, we apply a cluster-based resampling technique for the preprocessing step which can cover the problems ahead of random sampling. We believe that this preprocessing technique is better than other applied preprocessing technique because it reduces the size of the majority class without losing any useful information.
- We conduct various experiments to compare the performance of the proposed CNN AE method with hand-crafted feature extraction technique. To address the problem of high-dimensionality reduction. different dimensionality reduction techniques were explored for

hand-crafted features and their performance are explored and compared with the proposed model.
- The performance of the proposed method is analyzed and compared to previous outstanding studies using both hand-crafted and DL techniques.

It is worth mentioning that the implementation of the preprocessing phase and the proposed method is publicly available at https://github.com/SaraS92/CAE_ADD.

The remainder of the paper is organized as follows: Section 2 presents the proposed speech-based CNN AE approach for depression detection. Section 3 contains a description of the experimental framework and its results. Finally, in Section 4 a conclusion and some suggestion for future works are presented.

2. Proposed method

This section presents the proposed end-to-end speech-based depression detection approach utilizing an CNN AE as feature learning technique. We firstly provide a description of the depression dataset used in evaluation of this research. Next, we discuss the preprocessing steps carried out on the dataset. Moreover, we provide a description for the proposed CNN AE architecture to find the latent representation of the audio samples.

As depicted in Fig. 1, the overall process of our framework consists of two phases. First, the proposed CNN AE model is trained using training samples of speech to construct an automatic feature learner. Prior to this, a series of preprocessing steps including silence removal and addressing the class imbalance problem are conducted on the speech signals. The trained CNN AE will be utilized to modify any raw speech signal to a low-dimensional feature vector with minimal loss in its discriminative power. Therefore, in the phase of depression classification after preprocessing the data, each of the training and testing samples are being fed into the encoder of the trained CNN AE to capture their latent representation (feature vector). These latent attributes are then used for classifying samples into depressed/ non-depressed classes using a SVM classifier.

2.1. Dataset description

The Distress Analysis Interview Corpus (DAIC) database (Gratch et al., 2014) involves a series of clinical interviews in order to diagnose depression, anxiety, post-traumatic stress, and other mental disorders. The DAIC-WOZ (Distress Analysis Interview Corpus, Wizard of Oz) dataset is a subset of DAIC in which a virtual interviewer named Ellie asks questions designed to diagnose depression. This dataset contains audio, video, and the transcriptions of the interviews. The length of each audio file ranges between 7 and 33 min (averaging 16 min) with a 16KHz sampling frequency. All the recordings were provided by University of Southern California (USC)'s Institute of Creative Technologies as a part of AVEC 2016 challenge. Each participant has been given a PHQ-8 (the eight-item Patient Health Questionnaire depression) score based on their answers to questionnaires. This score determines depression severity, and Kroenke et al. (2009) proposed a threshold of 10 to classify patients in two groups of "depressed" and "non-depressed" cases (they consider patients with the PHQ-8 score above 10 as depressed).

The dataset is publicly available at a website provided by USC's Institute of Creative Technologies (http://dcapswoz.ict.usc.edu). The archive contains 189 interview folders of sessions 300 to 492, where each of them belongs to one individual. Some sessions have been excluded from the dataset due to technical problems with data capture. The training set, development set, and testing set are composed of 107, 47, and 35 files, respectively. Transforming the PHQ-8 scores of samples into binary labels of "non-depressed" and "depressed", converts the original data into an unbalanced binary classification problem.

2.2. Preprocessing

One of the most important steps before analyzing the audio files for depression is to apply some preprocessing techniques on the audio. The audio files contain long silences and the voices of both participant and interviewer. Each participant's interview is preprocessed to remove silence, as well as the virtual interviewer's voice using a python library called *PyAudioAnalysis*. *PyAudioAnalysis* (Giannakopoulos, 2015) provides easy-to-use and high-level Python wrappers for several audio analysis tasks and has been used in several other audio-based studies (Wroge et al., 2018; Chernykh & Prikhodko, 2017). We used this

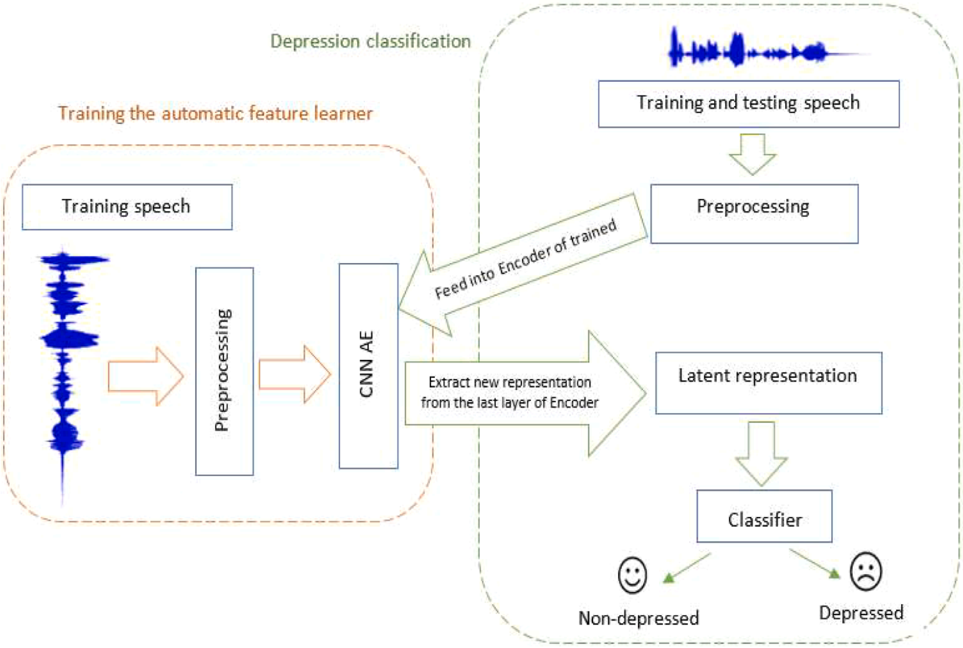
Depression classification
Training the automatic feature learner
Training speech
Preprocessing
CNN AE
Feed into Encoder of trained
Training and testing speech
Preprocessing
Extract new representation from the last layer of Encoder
Latent representation
Classifier
Non-depressed
Depressed

Fig. 1. General overview of the proposed semi-supervised Automatic Depression Detection (ADD) system. On the left, the proposed Convolutional Neural Network-based Auto-Encoder (CNN AE) model is trained using training samples of speech to construct an automatic feature learner. A series of preprocessing steps including silence removal and addressing the class imbalance problem are conducted on the speech signals. On the right, these latent attributes are then used for classifying samples into depressed/ non-depressed classes using a Support Vector Machine (SVM) classifier.

wrapper just for the diarization (deleting the voice of the second speaker) and silence removal in audio data. *Pyannote.audio* (Bredin et al., 2020) is another new toolkit for speaker diarization. Since, *PyAudioAnalysis* is very popular compared to the other toolkits for performing these tasks and has demonstrated good results in different studies (Wroge et al., 2018; Chernykh & Prikhodko, 2017), we preferred to use this wrapper. Both silence removal and speaker diarization tasks are successfully accomplished for approximately 84 percent of the audio files. 29 troubled audio files are omitted from the dataset because of the low volume levels or the proximity of speaker and interviewers' voices. The result is 160 audio files, including only the patients' voices, with a mean length of 587.249 s and standard deviation of 230.775 s.

To perform the end-to-end learning, we need the actual raw audio signal without any *a priori* feature extraction. Therefore, we create our first dataset called dataset *A*., by performing a sampling on signals of each audio file using a python library called *Librosa* (McFee et al., 2015). DAIC-WOZ audio files were recorded at a 16 kHz sampling rate. The sampling rate has a direct impact on the dimensionality of the input data and consequently the computational cost. Therefore, to build a low-dimensional dataset with acceptable quality, signals are down-sampled to 1 kHz.

After silence removal, speaker diarization, and down-sampling, we applied a sliding window to segment the audio signals to create a set of successive fixed-size windows with a fixed degree of overlap. This segmentation is applied in such a way that each down-sampled audio file is cropped into 150 sequential instances, each instance of 4 s length and with the stride (or step) of 1 ms shifting forward for each segment. It should be noted that a certain amount overlap is used to maximize the use of the audio information. This method can also be considered as a data augmentation method, it increases the number of samples extracted from one audio signal by reusing certain parts of a signal in new instance (Vázquez-Romero & Gallardo-Antolín, 2020). In fact, observing that we only have 107 training audio files with different and relatively long durations, we should use a technique to ovoid these challenges. These characteristics (dynamic long durations and small number of samples) can hinder any deep learning method, including our model from training. Cropping the audio files into fixed and same sized samples, which have been utilized in a related study (Vázquez-Romero & Gallardo-Antolín, 2020), can be a good method to overcome the small number of samples with different length. Therefore, we used the same strategy for achieving both data augmentation and having same short length samples goals. Due to these advantages of cropping, we didn't use any other type of data augmentation methods for audio signals including time stretching, pitch shifting, and insertion of noise (Nanni et al., 2020) on audio signals. The result is we will have 150 instances for each speaker, each with the length of 4,000 signal samples. In Fig. 2 the general process of audio signal segmentation with a window size of 4 and overlapping stride size of 0.5 s is illustrated.

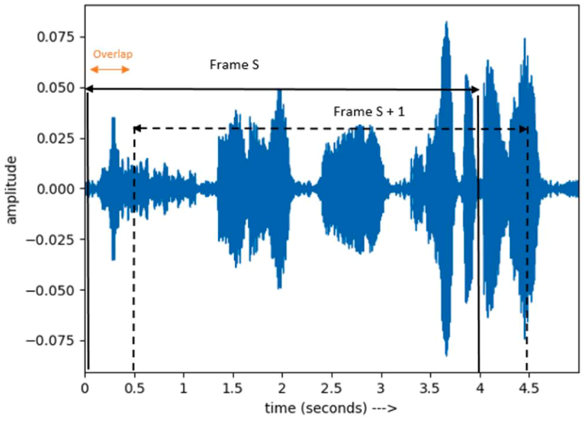

Fig. 2. Input audio signal segmentation into multiple frames with a fixed degree of overlap. A window size of 4 and overlapping stride size of 0.5 s is illustrated.

The distribution of samples in the original dataset and the dataset *A* are depicted in the Table 1. One of the most significant problems with this dataset is that it contains uneven sample distributions and as such is a severely unbalanced dataset. An unbalanced dataset of this sort may cause overfitting and result in a classifier with a bias towards the class with larger number of samples (in this case the non-depressed class). Specifically, this becomes an important issue in this study since our aim is to detect the minority class (depressed) accurately. The unbalanced dataset is to be expected, depression being a condition that thankfully occurs in only a small percentage of the underlying population.

To avoid overfitting and decreasing the classification performance, some studies suggested random undersampling methods before training the models (Ma et al., 2016; Vázquez-Romero & Gallardo-Antolín, 2020). However, a random undersampling technique might cause a loss of useful information, which may diminish the classification performance. To solve the unbalanced problem in this dataset, we applied a particular sampling technique called *ClusterCentroids under-sampling* (Lemaître et al., 2017) which is a cluster-based resampling technique. The ClusterCentroids method under-samples the majority class by replacing the samples from each cluster with the centroid of that cluster using the k-means algorithm. This results in an equal proportion of positive and negative samples (depressed and non-depressed samples). Using this way we reduce the number of samples in the majority class with avoiding the information loss in the whole data.

2.3. CNN AE as a representation learning method

This subsection presents an architecture of the CNN AE to detect depression using raw audio signals. The proposed CNN AE model is shown in Fig. 3 and is similar to a simple AE model. It encompasses two major components of encoder and decoder. The encoder comprises four blocks consisting of convolutional, max-pooling (down-sampling) and dropout layers. It should be noted that the convolutional layers have a kernel size of 3, Rectified Linear Unit (ReLU) activation function and filters of 64, 128, 256, and 512, respectively. Each convolution layer is followed by a 1-dimensional max-pooling layer to down-sample the input representation with pooling sizes of 2. In addition, a dropout layer with the dropout rate of 0.25 follows each pooling layer to avoid overfitting. Finally, a flatten layer and a Dense layer are introduced to generate a one-dimensional vector of 700 features as a latent representation.

The decoding phase comprises the reverse transform of the input data. In this phase the latent representation is transformed into a reconstructed vector. A Dense layer with activation function of ReLU

Table 1
The distribution of depressed and non-depressed patients in the original dataset and the dataset A.

		Depressed	Non-depressed	Total
Training set	Original set	30	77	107
	After silence removal and speaker diarization	29	67	97
	After speech segmentation	4,350	10,050	14,400
Testing set	Original set	14	33	47
	After silence removal and speaker diarization	11	27	38
	After speech segmentation	1,650	4,050	5,700
Development set	Original set	12	23	35
	After silence removal and speaker diarization	9	17	26
	After speech segmentation	1,350	2,550	3,900

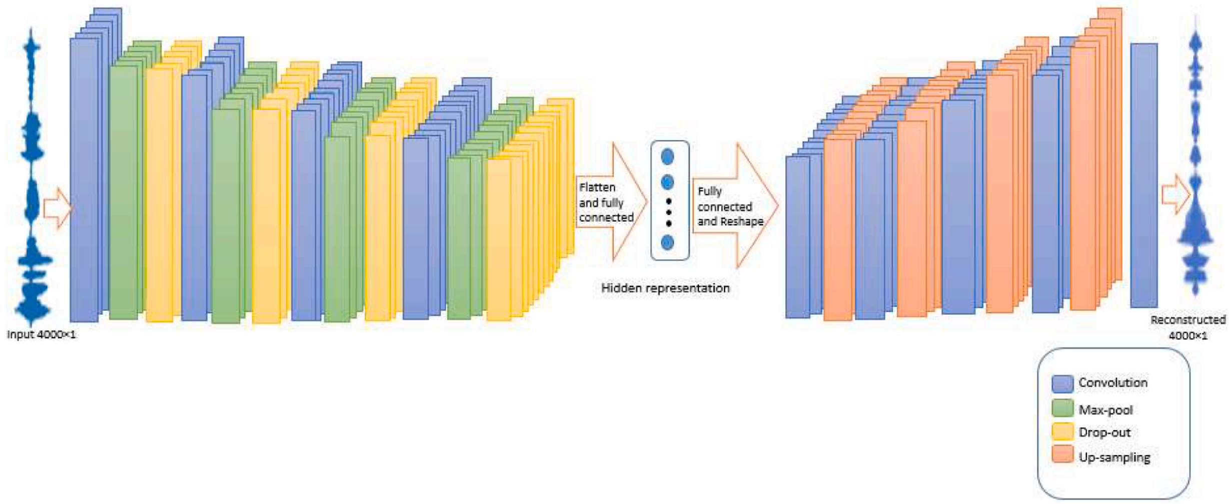

Fig. 3. Proposed CNN AE architecture for depression identification.

and a reshape layer are considered to reshape data into the form to apply convolutional layers. Four blocks of convolutional, and 1 dimensional up-sampling layers are considered. The convolution layers have a kernel size of 3, ReLU activation function and filters of 512, 256, 128, and 64, respectively. The 1 dimensional up-sampling layers have the size of 2. Finally, a convolution layer with kernel size of 3 and filter of 1 is used to reconstruct the input. A brief description of the proposed model layers and their output shape is shown in Table 2.

The encoder component of the CNN AE will be assumed as an unsupervised feature learning system, by capturing the latent representation (feature vector) of each sample from its last layer. After performing the unsupervised feature learning procedure for training and testing samples to obtain the feature sets, we use a simple classifier algorithm such as SVM, MLP, GB, and RF to quantify the discriminative ability of the new feature set.

Table 2
-Proposed model layer details.

Layers	Hyperparameter settings	Output shape
Input layer	–	(4000,1)
Conv. 1D	Kernel size = 3, Activation = ReLU	(4000,64)
Max-pool 1D	Pool size = 2	(2000,64)
Drop-out	Rate = 0.25	(2000,64)
Conv. 1D	Kernel size = 3, Activation = ReLU	(2000,128)
Max-pool 1D	Pool size = 2	(1000,128)
Drop-out	Rate = 0.25	(1000,128)
Conv. 1D	Kernel size = 3, Activation = ReLU	(1000,256)
Max-pool 1D	Pool size = 2	(500,256)
Drop-out	Rate = 0.25	(500,256)
Conv. 1D	Kernel size = 3, Activation = ReLU	(500,512)
Max-pool 1D	Pool size = 2	(250,512)
Drop-out	Rate = 0.25	(250,512)
Flatten	–	(128000)
Dense	Activation = ReLU	(700)
Dense	Activation = ReLU	(128000)
Reshape	–	(250,512)
Conv. 1D	Kernel size = 3, Activation = ReLU	(250,512)
Up-sampling 1D	Size = 2	(500,512)
Conv. 1D	Kernel size = 3, Activation = ReLU	(500,256)
Up-sampling 1D	Size = 2	(1000,256)
Conv. 1D	Kernel size = 3, Activation = ReLU	(1000,128)
Up-sampling 1D	Size = 2	(2000,128)
Conv. 1D	Kernel size = 3, Activation = ReLU	(2000,64)
Up-sampling 1D	Size = 2	(4000,64)
Conv. 1D	Kernel size = 3, Activation = Sigmoid	(4000,1)

3. Results and discussion

This section explains the tools used for developing the audio-based depression detection system and the experimental results compared to other methodologies. To construct different models and perform multiple experiments, the Google Collaboratory platform with the programming language Python 3.6 has been used. In addition, the deep models evaluated in this study has been programmed using Keras (Chollet, 2015) in Python. The training, development and testing sets provided by the challenge of DAIC-WOZ (Gratch et al., 2014) are utilized for developing and evaluating the proposed model. It is worth mentioning that a Minmax normalization is performed on all training, testing, and development samples before feeding them into any model. In addition, it should be noted that different hyperparameters of the model such as activation function, kernel size, filter size, batch size, dropout rate, pooling size, and number of layers should be selected through a process for optimizing the error rate. In this study, the optimum configuration of different hyperparameters have been selected through the classical trial-and-error procedure as suggested by Ortiz-Rodriguez et al. (2013).

Since we have adopted the DAIC-WOZ dataset which was designed and proposed by the AVEC 2016 (Valstar et al., 2016) depression detection challenge, we used the same experimental protocols designed in this challenge for model evaluation. These protocols comprise reporting three metrics of precision, recall, and F-measure. In fact, it is a common practice to use these metrics for model comparison when the datasets are unbalanced. Therefore, F-measure as the harmonic average of precision and recall is the main preferred metric considered by many researchers in this area, including our research. However, compared to other studies, in order to provide more transparency and clearer illustration of hit/misclassified rate we have used confusion matrix as another metric. As the aim of this study is to detect depressed people using only audio signals, utilizing a confusion matrix provides a clearer image of the hit/misclassified rate in the proposed depression detection system. Since, many other metrics (such as accuracy, sensitivity, and specificity) can be calculated based on the confusion matrix, we highly recommend that in future other studies in this field, the confusion matrices should be mentioned as an evaluation metric for better comparison.

In the following we initially provide experiments regarding the feature extraction ability of various models. First of all, we compare the ability of CNN AE and Dense AE as automatic feature extraction methods from raw audio to assess the possible potential of an end-to-end system for depression detection. Then the winner end-to-end method, which is the pipeline of CNN AE and SVM, is compared to a handcrafted-feature approach and lower dimensional representation of these features to evaluate the performance of CNN AE in finding a compact discriminative representation of audio. Then we follow up with a comparative analysis of different methods proposed for depression detection.

3.1. Comparison of Dense AE and CNNAE using raw audio data

In order to investigate the discriminative ability of the feature vectors extracted using proposed deep CNN AE, a deep fully connected (Dense) AE is also trained on the raw audio signals in the dataset *A*, depicted in Fig. 4. The encoder component consists of two fully connected layers. The first and the second layers have 2,000 and 1,000 hidden units and the activation functions are Tanh and ReLU, respectively. The bottleneck layer has 700 hidden units using the ReLU activation function, this means that the encoder transforms the original 4,000-dimensional feature space to a 700-dimensional latent space. The decoder component has 1,000 and 2,000 hidden units and Tanh activation functions. Finally, the decoder reconstructs the output layer with the size of 4,000 and activation function of ReLU.

To compare the performance of both AE architectures, they are trained using 300 epochs with a batch-size of 32 samples using an *Adadelta* optimizer and *Mean Squared Error (MSE)* as the loss function. Several studies (Lee et al., 2018; Nogas et al., 2020; Zhao et al., 2017) have been conducted on training CNN AEs for different applications. They have shown that *Adadelta* is performing very well as an optimizer for training CNN AEs. This optimizer is one of the well-known optimizers that has been shown promising in different application. The reason that we have selected this optimizer is that this optimizer is performing well is that determining a good learning rate by introducing a new dynamic learning rate using the first order information. Using this optimizer, we don't no need to tune the learning rate manually. This optimizer is robust to noisy gradient information. Model architecture choices, various data modalities and hyperparameter selection. We should note that this optimizer requires minimal computation over gradient descent and robustness for large gradient and different architecture choices. Based on the results of these studies and the trial-and-errors process we decided to use *Adadelta* with the learning rate of 0.001 as the main optimizer in our study. To avoid overfitting and diminish the number of epochs (decrease the total running time) the early stopping call-back is exploited. In this way, the training halts when the loss function for validation set stops improving, and the trained parameters which led to the lowest validation error are kept. In this study, an early stopping with minimum change of 0.00005 in the validation loss and the patience of 3 are considered.

We compared the reconstruction performance (training and validation loss) of the proposed CNN AE method and Dense AE using the raw audio and presented it in Fig. 5. This shows that in spite of the fact that the Dense AE is being trained using more epochs compared to the CNN AE, the overall training and validation loss in CNN AE is less than for Dense AE. This suggests that CNN AE outperforms simple Dense AE in learning the best low-dimensional representation of the input audio signals.

In order to investigate the discriminative performance of both architectures, they were applied to raw signals and then the extracted features were sent to different classifiers for depression classification. Four typical classifiers of SVM, MLP, Gradient Boosting (GB), and Random Forest (RF) have been explored on the extracted features from both scenarios. All these classifiers are provided by *scikit-learn* library in *Python* and the details of parameter configuration for each of these classifiers is shown in Table 3. The resulting confusion matrices which are depicted in Fig. 6, illustrates that the accuracy of detecting both depressed and non-depressed samples using CNN AE as the automatic feature extractor is higher than Dense AE on raw signals. Comparing Fig. 6a and Fig. 6e shows that the combination of Dense AE and SVM clearly fails to discriminate depressed from non-depressed class when the input data is raw audio signals. Moreover, comparing the confusion matrices with same classifiers generally depicts the fact that CNN AE is performing better than the Dense AE in finding the most representative feature set. More analysis of these matrices helps to find the best combination of feature extractor and classifier for the raw audio signals, which is the CNN AE and SVM, respectively. This system depicts a more balanced performance in the recall rate for both classes and shows a good potential for being an end-to-end approach in depression detection. The evaluation metrics for both architectures are represented in Table 4 for further analysis.

To mention the computational time performance of the proposed pipeline, we have to state that the offline processing which includes preprocessing audio files and Training CNN AE for feature extraction took about 4 h and 25 min and this can be considered as a reasonable development time and can be considerably less with higher performance computing. The processing time is determined by Google Collaboratory

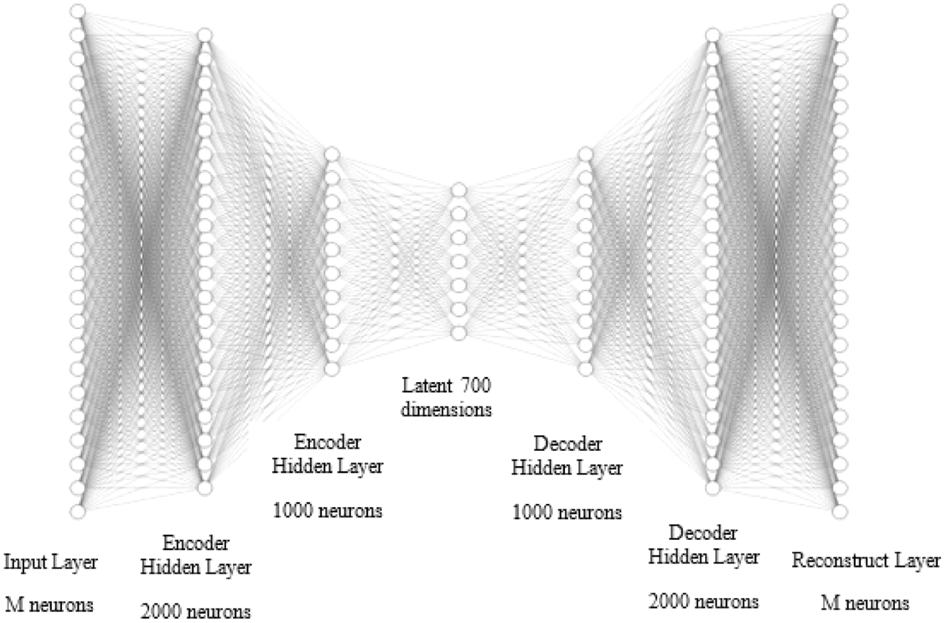

Fig. 4. Dense AE architecture utilized for reconstructing the input with size of M. The encoder component consists of two fully-connected layers. The first and the second layers have 2,000 and 1,000 hidden units and the activation functions are Tanh and ReLU, respectively. The bottleneck layer has 700 hidden units using the ReLU activation function, this means that the encoder transforms the original M−dimensional feature space to a 700-dimensional latent space. The decoder component has 1,000 and 2,000 hidden units and Tanh activation functions. Finally, the decoder reconstructs the output layer with the size of M and activation function of ReLU.

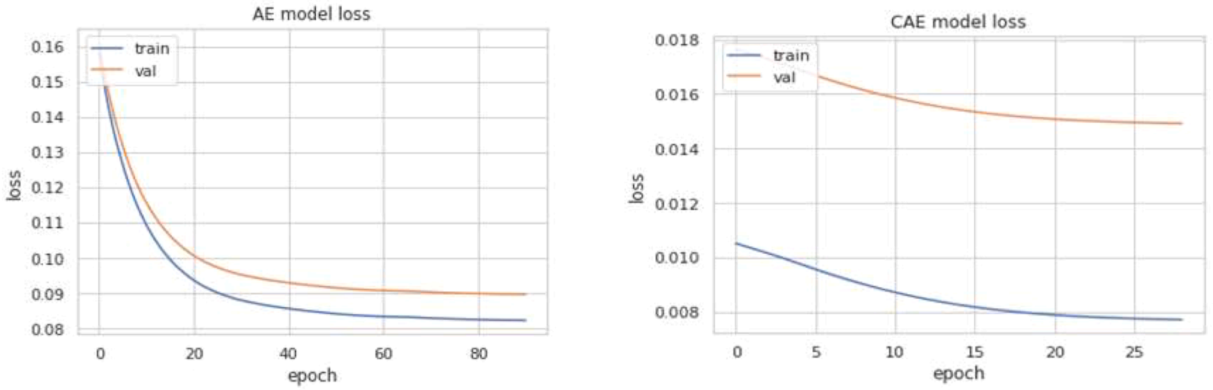

Fig. 5. Mean Squared Error (MSE) along the number of epochs for CNN AE (right side) and Dense AE (left side).

Table 3
The parameter settings of four common classifiers explored for depression detection.

Classifier	Parameter settings
Support Vector Machine (SVM)	Kernel = Radial Basis Function (RBF)
Multi-Layer Perceptron (MLP)	Maximum iteration = 100, Number of neurons in first hidden layer = 128, Number of neurons in second hidden layer = 100
Gradient Boosting (GB)	Number of estimators = 100, Max-depth = 10
Random Forest (RF)	Max-depth = 10

GPU, 12 GB RAM.

3.2. *Comparing an end-to-end approach to hand-crafted features*

3.2.1. *Hand-crafted feature extraction using openSMILE tool*

To compare the performance of the proposed methodology to other hand-crafted feature extraction methods, we build another dataset called dataset *B* for hand-crafted features. After preprocessing the audio files to extract everyone's voice, the vocal features are captured programmatically using *Python 3.8* and *open-source Speech and Music Interpretation by Large-space Extraction* (openSMILE) toolkit for each audio file. We extracted the AVEC 2013 (Valstar et al., 2013) feature set including 2,268 low-level descriptor (LLD) features commonly used in audio signal analysis and emotion detection (Valstar et al., 2013). They are composed of 32 energy and spectral related LLDs × 42 functionals, 6 voicing related LLD × 32 functionals, 32 delta coefficients of the energy and spectral LLD × 19 functionals, 6 delta coefficients of the voicing related LLD × 19 functionals, and 10 voiced or unvoiced durational features.

The functionals are composed of statistical functionals (maximum, mean, skewness, flatness, etc.), regression functionals (linear regression slope, etc.), and local minima/maxima-related functionals (mean and standard deviation of rising and falling slopes, etc.) (Demiroglu et al., 2020). These features were derived using overlapping short fixed windows of 20 s duration with the stride of 1 s shifting forward (Valstar et al., 2013). This results in multiple segments each of 20 s length from the whole audio. Due to the different lengths of each recording, it is better to choose 150 first sequential segments as *samples* with corresponding features. In this way, we approximately assume that the first 168 s of each participant's audio will prove sufficient for identifying the key characteristics for depression classification. This sampling technique assists us in obtaining an equal sized set of samples for each speaker and in turn to avoid model inclination to specific over-represented speakers. In addition, this method can be interpreted as a data augmentation method to train the model on larger dataset. Therefore, this model will be more general and accurate compared to the model trained on the original dataset.

3.2.2. *Comparison of different dimensionality reduction methods with the proposed model*

In this evaluation phase, we assessed the ability of different models, trading dimensionality reduction factor with classification performance. To perform this task, we evaluate the discriminative ability of the extracted latent representations in the proposed CNN AE compared to other dimensionality reduction methods developed on openSMILE features set. In addition, the Dense AE mentioned in the Section 3.1 is utilized for dimensionality reduction on hand-crafted features and compared to the other aforementioned methods. The classifiers are trained on the original feature set of openSMILE and a lower dimensional representation obtained from following popular techniques: PCA with 700 principal components (Jolliffe & Cadima, 2016) as an unsupervised technique, LDA (Balakrishnama & Ganapathiraju, 1998) as a supervised technique, and the proposed Dense and CNN Auto-encoders.

Table 4 illustrates the results of the classification experiments for different dimensionality reduction methods. Three metrics of F-measure, precision, and recall are shown for both of the depressed and non-depressed classes (the results for the non-depressed classes are shown in parentheses). Two important points are evident after thoroughly observing the results in Table 4. First, the average F-measure rate for both classes is higher when the input data is raw audio signal and there is an Autoencoder for feature extraction. This shows that automatic feature extraction is fairly more successful in depression detection compared to high dimensional openSMILE hand-crafted feature sets and their low dimensional projections. Second, by comparing all of the combinations of input data, dimensionality reduction method, and classifiers it is apparent that the performance of SVM trained on the 700-dimensional latent representation extracted by the proposed CNN AE on the raw audio signal is significantly better than the others. This indicates that the new compressed representation discriminates depressed (non-depressed) samples strongly. It can also be observed that CNN AE performs better than the other dimensionality reduction techniques. It should be noted that although the reduced feature space provided by the Dense AE on hand-crafted features performs worse than the CNN AE, we can observe it still performs well in capturing the differences between samples of both classes compared to other dimensionality reduction methods and original high dimensional feature set. This is also another evidence for the better ability in discriminative feature learning of Autoencoders compared to other dimensionality reduction methods.

As a confirmation to the findings, we mentioned above, we can investigate the confusion matrices of the SVM classifier for all the methods mentioned in Table 4. Based on the confusion matrices reported in Fig. 6 and Fig. 7, the hit rate in detecting both depressed and non-depressed samples using the CNN AE on raw signals as a hidden representation extractor is better than using all hand-crafted features

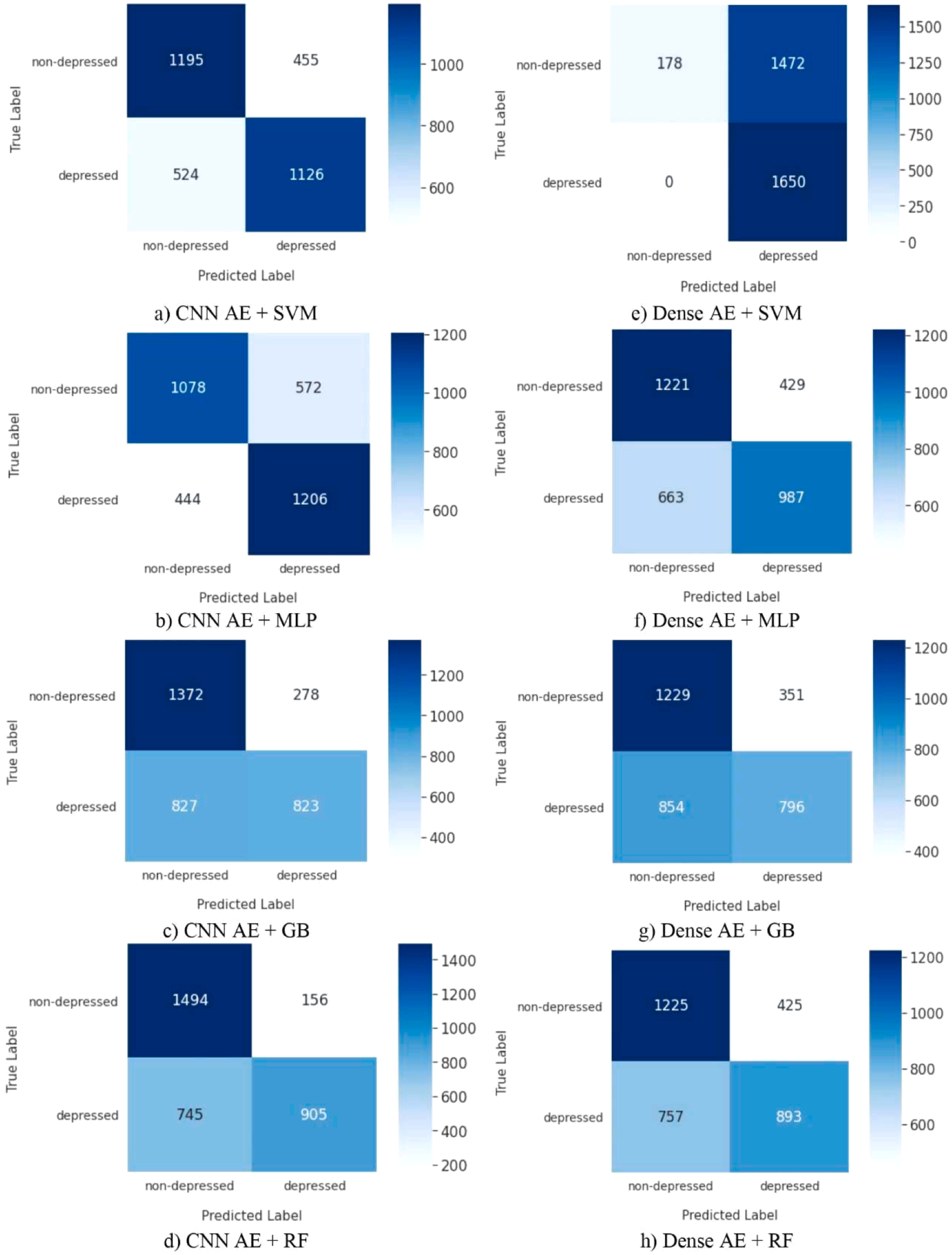

Fig. 6. The confusion matrices of the depression detection systems based on CNN AE (left side) and Dense AE (right side) on raw signal with different classifiers.

approaches. Based on this we can deduce that these automatic features extracted by the CNN AE have higher classifying performance and can be used as a good building block for an ADD.

Another important point that we need to highlight is that Fig. 6a and Fig. 7b show that CNN AE specifically outperforms Dense AE (on hand-crafted features) in finding a good representation for depressed class, which implies that this results in a system with better true positive prediction compared to other hand-crafted methodologies.

3.3. *Performance evaluation of different depression detection systems compared to proposed model*

We compare our best proposed depression detection system, to four previously proposed systems for detection based on vocal signals. Table 5 illustrates the performance of these methodologies and their feature extraction techniques. It is worth mentioning that these methods reported their performance on the testing set, and it is therefore entirely fair to compare our proposed method to them. The best proposed pipeline for classification according to the previous section is the CNN AE as the representation learning method and SVM as the classifier.

Table 4
Performance comparison of different classifiers on hand-crafted features and automatically extracted features (the results for the non-depressed classes are shown in parentheses).

Input data	Dimensionality reduction method	Classifier	Precision	Recall	F-measure
Hand-crafted openSMILE feature set	None	SVM	0.62 (0.58)	0.48 (0.71)	0.54 (0.64)
	PCA		0.62 (0.58)	0.47 (0.71)	0.54 (0.64)
	LDA		0.62 (0.57)	0.44 (0.73)	0.51 (0.64)
	Dense AE		0.69 (0.65)	0.60 (0.73)	0.64 (0.69)
	None	MLP	0.65 (0.59)	0.48 (0.74)	0.55 (0.65)
	PCA		0.65 (0.59)	0.48 (0.74)	0.55 (0.65)
	LDA		0.62 (0.57)	0.45 (0.73)	0.52 (0.64)
	Dense AE		0.70 (0.65)	0.60 (0.74)	0.64 (0.69)
	None	GB	0.81 (0.56)	0.26 (0.94)	0.40 (0.70)
	PCA		0.69 (0.57)	0.35 (0.84)	0.47 (0.68)
	LDA		0.63 (0.57)	0.44 (0.74)	0.52 (0.64)
	Dense AE		0.69 (0.60)	0.48 (0.79)	0.57 (0.68)
	None	RF	0.72 (0.57)	0.36 (0.86)	0.48 (0.69)
	PCA		0.70 (0.56)	0.33 (0.86	0.45 (0.68)
	LDA		0.63 (0.57)	0.44 (0.74)	0.52 (0.64)
	Dense AE		0.68 (0.62)	0.54 (0.74)	0.60 (0.67)
Raw speech data	Dense AE	SVM	0.53 (1.00)	1.00 (0.11)	0.69 (0.19)
	Proposed CNN AE		**0.71 (0.70)**	**0.68 (0.72)**	**0.70 (0.71)**
	Dense AE	MLP	0.77 (0.64)	0.52 (0.84)	0.62 (0.73)
	Proposed CNN AE		0.68 (0.71)	0.73 (0.65)	0.70 (0.68)
	Dense AE	GB	0.84 (0.65)	0.51 (0.90)	0.63 (0.75)
	Proposed CNN AE		0.75 (0.62)	0.50 (0.83)	0.60 (0.71)
	Dense AE	RF	0.81 (0.65)	0.53 (0.87)	0.64 (0.74)
	Proposed CNN AE		0.85 (0.67)	0.55 (0.91)	0.67 (0.77)

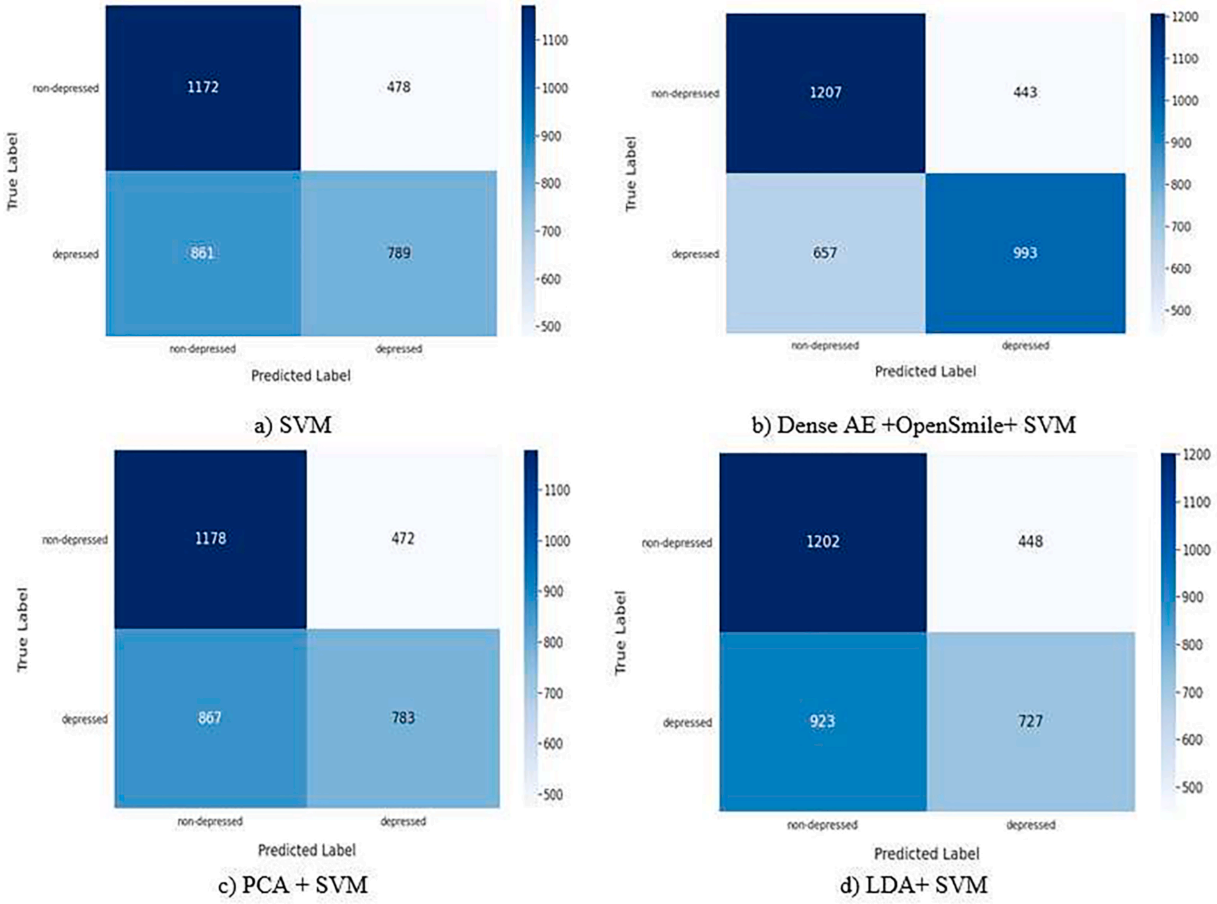

Fig. 7. Confusion matrices for hand-crafted features-based depression detection models on the testing set.

Therefore, the comparative analysis in this section will be based on a comparison to this model.

Two of the audio-based methods for comparison, are using hand-crafted features as an input to their traditional machine learning classifiers. The first method for comparison is the baseline model suggested by AVEC-2016 challenge. This performs an SVM classifier on the audio hand-crafted features extracted using COVAREP (Valstar et al., 2016). The second method proposed by Pampouchidou et al.

Table 5
Performance Comparison of different audio-based depression detection systems.

Method	Feature extraction technique	Precision	Recall	F-measure	F-measure avg
SVM (Baseline) (Valstar et al., 2016)	COVAREP	0.27 (0.94)	0.89 (0.42)	0.41 (0.58)	0.50
Audio + Gender (Pampouchidou et al., 2016)	Statistical descriptors extracted from COVAREP features	(N/A)	(N/A)	0.59 (**0.87**)	**0.73**
DepAudionet (Ma et al., 2016)	Log-spectogram features	0.44 (0.80)	0.60 (0.68)	0.50 (0.73)	0.62
Ensemble of CNNs (Vázquez-Romero & Gallardo-Antolín, 2020)	Log-spectogram features	0.52 (0.88)	0.79 (0.70)	0.63 (0.78)	0.71
Proposed CNN AE + SVM	Raw audio	0.71 (0.70)	0.68 (0.72)	**0.70** (0.71)	0.71

(2016) is a gender-based model utilizing the gender of participants as an additional information, to achieve better performances. The extracted statistical descriptors from pre-extracted features using COVAREP are utilized as the hand-crafted feature set and a Decision Tree as the classifier. The F-measure metrics reported for each of these methods in Table 5, proves the fact that our proposed automatic feature extraction method is performing better than other hand-crafted feature techniques such as COVAREP and openSMILE (investigated in the previous section) in depression detection. Although, the study presented by Pampouchidou et al. (2016) is considering gender as an extra feature and the average F-measure for both classes reported by them is better than our proposed method, it is clear that their gender-based model is failing in detecting depressed class and the inclination towards the non-depressed class is high. However, the interesting fact is that even though our model is gender-independent our depression detection system has a relatively high performance in both classes and represents a balanced F-measure result. This can be considered as an additional advantage of our model which is detecting depression in a gender-independent manner.

Two of the other methods, chosen for comparison in Table 5, are based on adopting deep learning techniques for depression detection. One of them is the DepAudionet model proposed by Ma et al. (2016). This consists of a spectrogram-based Convolutional and LSTM deep model for depression detection. Another method of interest is the one proposed by Vázquez-Romero and Gallardo-Antolín (2020) which takes the voice log-spectrograms of each individual as inputs and performs an ensemble of CNNs on them. Both methods (Ma et al., 2016; Vázquez-Romero & Gallardo-Antolín, 2020), perform a random sampling technique in order to handle the unbalanced data problem.

Two important issues are observable from this table. First, the methodologies based on log-spectrograms (by converting the audio signal to log-spectrograms (images)) and then performing CNN models generally results in better performance comparing to those using one dimensional signals. However, the interesting result is that our proposed CNN AE performs better in depression detection based on F-measure metric of depressed class. Another important issue to notice is that, converting an audio signal to spectrograms hinders the system performing real-time recognition, converting audio to large two-dimensional images is computational expensive. This argues against the spectrographic approach and emphasizes the necessity of building a system that performs immediate feature extraction resulting in a small one-dimensional feature set to perform depression detection. Our proposed method not only has better performance in depression detection compared to other techniques, but it also performs real-time feature extraction, which is an important advantage of our proposed CNN AE. Although, reported F-measure for non-depressed class in Vázquez-Romero and Gallardo-Antolín's work is higher than our proposed method, but the overall (average) of F-measures for both methods is approximately 0.71. This states the fact that the overall discriminative performance of our method is equal to a system with huge amount of preprocessing and can be considered an efficient and effective depression detection method.

4. Conclusion

In this paper, we proposed an end-to-end depression detection model using a CNN Auto-encoder for automatic feature extraction out of raw audio signals. Specifically, not only is the proposed CNN AE better than a simple Dense AE in extracting discriminative low dimensional features for depression detection, it also performs better than other proposed methods in the literature. The CNN AE approach for automatic feature extraction as a preprocessing step performed better than hand-crafted feature extraction methods. The results confirm the fact that the proposed automatic feature extraction methodology can be utilized as an accurate and efficient technique in any real-time depression detection system assisting the health care providers.

In this study, due to the memory and computational limitations, we couldn't utilize the whole samples from the dataset to train and test the proposed model. Therefore, it's worthwhile exploring the proposed model using high-performance computing facility. As in this study, we have presented that Convolutional AE is outstanding hand-crafted feature extraction techniques as well the proposed DL methods, therefore, it is recommended to explore other extensions of Autoencoder such as Adversarial AE for depression detection model. In addition, another future work that we can offer is to assess the performance of our proposed model on other datasets with different spoken languages to evaluate the sensitivity of the method to languages.

CRediT authorship contribution statement

Sara Sardari: Methodology, Formal analysis, Writing - original draft. **Bahareh Nakisa:** Conceptualization, Validation, Supervision, Writing - review & editing. **Mohammed Naim Rastgoo:** Data curation, Visualization, Writing - review & editing. **Peter Eklund:** Supervision, Project administration, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Al Hanai, T., Ghassemi, M. M., & Glass, J. R. (2018). September. Detecting Depression with Audio/Text Sequence Modeling of Interviews. *Interspeech*, 1716–1720.

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE, 2*(1), 1–18.

Banan, A., Nasiri, A., & Taheri-Garavand, A. (2020). Deep learning-based appearance features extraction for automated carp species identification. *Aquacultural Engineering, 89*, Article 102053.

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing, 18*(1998), 1–8.

Beck, A. T., Steer, R. A., Brown, G. K., 1996. Beck depression inventory.

Braga, D., Madureira, A. M., Coelho, L., & Ajith, R. (2019). Automatic detection of Parkinson's disease based on acoustic analysis of speech. *Engineering Applications of Artificial Intelligence, 77*, 148–158.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., … Gill, M. P. (2020). Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7124–7128). IEEE.

Chernykh, V., & Prikhodko, P. 2017. Emotion recognition from speech with recurrent neural networks. *arXiv preprint arXiv:1701.08071*.

Chollet, F. 2015. Keras. Available online at: https://github.com/fchollet/keras.

Cohn, J. F., Cummins, N., Epps, J., Goecke, R., Joshi, J. Scherer, S., 2018. Multimodal assessment of depression from behavioral signals. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pp. 375-417.

Chorowski, J., Weiss, R. J., Bengio, S., & van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing, 27*(12), 2041–2053.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication, 71*, 10–49.

Demiroglu, C., Beşirli, A., Ozkanca, Y., Çelik, S., 2020. Depression-level assessment from multi-lingual conversational speech data using acoustic and text features. Journal on Audio, Speech, and Music Processing. 2020, 17 (2020). 10.1186/s13636-020-00182-4.

Fan, Y., Xu, K., Wu, H., Zheng, Y., & Tao, B. (2020). Spatiotemporal modeling for nonlinear distributed thermal processes based on KL decomposition, MLP and LSTM network. *IEEE Access, 8*, 25111–25121.

Giannakopoulos, T. (2015). Pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one, 10*(12), Article e0144610.

Gogoi, M., & Begum, S. A. (2017). Image Classification Using Deep Autoencoders. In *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1–5). IEEE.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language, 53*, 181–197.

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D. R., 2014, May. The distress analysis interview corpus of human and computer interviews. LREC, pp. 3123-3128.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders, 114*(1–3), 163–173.

Lee, H., Kim, J., Kim, B., Kim, S., 2018, December. Convolutional Autoencoder Based Feature Extraction in Radar Data Analysis. In *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)* pp. 81-84. IEEE.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research, 18*(1), 559–563.

Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences, 409*, 17–26.

Lopez-de-Ipina, K., Martinez-de-Lizarduy, U., Calvo, P. M., Mekyska, J., Beitia, B., Barroso, N., … Ecay-Torres, M. (2018). Advances on automatic speech analysis for early detection of Alzheimer disease: A non-linear multi-task approach. *Current Alzheimer Research, 15*(2), 139–148.

Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 35–42).

Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks* (pp. 52–59). Berlin, Heidelberg: Springer.

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine, 3*(11), Article e442.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O., 2015, July. librosa: Audio and music signal analysis in python. Proceedings of the 14th python in science conference, Vol. 8, pp. 18- 25.

McIntyre, G., Göcke, R., Hyett, M., Green, M., & Breakspear, M. (2009). An approach for automatically measuring facial activity in depressed subjects. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–8). IEEE.

Mou, L., Zhou, C., Zhao, P., Nakisa, B., Rastgoo, M. N., Jain, R., & Gao, W. (2021). Driver stress detection via multimodal fusion using attention-based CNN-LSTM. *Expert Systems with Applications, 173*, Article 114693.

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2020). Automatic Emotion Recognition Using Temporal Multimodal Deep Learning. *IEEE Access, 8*, 225463–225474.

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2018b). Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. *IEEE Access, 6*, 49325–49338.

Nakisa, B., Rastgoo, M. N., Tjondronegoro, D., & Chandran, V. (2018a). Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. *Expert Systems with Applications, 93*, 143–155.

Nanni, L., Maguolo, G., & Paci, M. (2020). Data augmentation approaches for improving animal audio classification. *Ecological Informatics, 57*, Article 101084.

Nogas, J., Khan, S. S., & Mihailidis, A. (2020). Deepfall: Non-invasive fall detection with deep spatio-temporal convolutional autoencoders. *Journal of Healthcare Informatics Research, 4*(1), 50–70.

Ortiz-Rodriguez, J. M., Martinez-Blanco, M. R, Cervantes-Viramontes, J. M., Vega-Carrillo, H. R., 2013. Robust design of artificial neural networks methodology in neutron spectrometry. In Artificial Neural Networks – Architectures and Applications – Edition 1. Chapter 4, INTECH.

Ozkanca, Y., Demiroglu, C., Besirli, A., & Celik, S. (2018). Multi-lingual depression-level assessment from conversational speech using acoustic and text features. *Proceedings of Interspeech, 2018*, 3398–3402.

Palylyk-Colwell, E., & Argáez, C. (2018). Telehealth for the Assessment and Treatment of Depression, Post-Traumatic Stress Disorder, and Anxiety: Clinical Evidence. In O. N. Ottawa (Ed.), *Canadian Agency for Drugs and Technologies in Health* (pp. 8–9). Canada.

Pampouchidou, A., Simantiraki, O., Fazlollahi, A., Pediaditis, M., Manousos, D., Roniotis, A., … Yang, F. (2016). Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (pp. 27–34).

Qureshi, S. A., Hasanuzzaman, M., Saha, S., Dias, G., 2019. The Verbal and Non Verbal Signals of Depression–Combining Acoustics, Text and Visuals for Estimating Depression Level. arXiv preprint arXiv:1904.07656.

Rastgoo, M. N., Nakisa, B., Maire, F., Rakotonirainy, A., & Chandran, V. (2019). Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications, 138*, Article 112793.

Rastgoo, M. N., Nakisa, B., Rakotonirainy, A., Chandran, V., & Tjondronegoro, D. (2018). A critical review of proactive detection of driver stress levels based on multimodal measurements. *ACM Computing Surveys (CSUR), 51*(5), 1–35.

Ribeiro, A., Matos, L. M., Pereira, P. J., Nunes, E. C., Ferreira, A. L., Cortez, P., Pilastri, A., 2020. Deep Dense and Convolutional Autoencoders for Unsupervised Anomaly Detection in Machine Condition Sounds. arXiv preprint arXiv:2006.10417.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A. and Çiftçi, E., 2018, October. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. Proceedings of the 2018 on audio/visual emotion challenge and workshop, pp. 3-13.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M., 2017, October. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 3-9.

Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W. and Espy-Wilson, C., 2018. Adversarial auto-encoders for speech based emotion recognition. arXiv preprint arXiv:1806.02146.

Shamshirband, S., Rabczuk, T., & Chau, K. W. (2019). A survey of deep learning techniques: Application in wind and solar energy resources. *IEEE Access, 7*, 164650–164666.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R. and Pantic, M., 2016, October. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th international workshop on audio/visual emotion challenge, pp. 3-10.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R. and Pantic, M., 2014, November. Avec 2014: 3d dimensional affect and depression recognition challenge. Proceedings of the 4th international workshop on audio/visual emotion challenge, pp. 3-10.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R. and Pantic, M., 2013, October. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 3-10.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative. *J Mach Learn Res, 10*(66–71), 13.

Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Eskofier, B., Klucken, J., & Nöth, E. (2018). Multimodal assessment of Parkinson's disease: A deep learning approach. *IEEE journal of biomedical and health informatics, 23*(4), 1618–1630.

Vázquez-Romero, A., & Gallardo-Antolín, A. (2020). Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks. *Entropy, 22*(6), 688.

Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports, 11* (1), 1–13.

Wen, T., & Zhang, Z. (2018). Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals. *IEEE Access, 6*, 25399–25410.

World Health Organization (WHO). (2017). *Depression and other common mental disorders: Global health estimates*. Geneva: World Health Organization.

Wroge, T. J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., & Ghomi, R. H. (2018). Parkinson's disease diagnosis using machine learning and voice. In *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1–7). IEEE.

Xia, R., & Liu, Y. (2013). Using denoising autoencoder for emotion recognition. *Interspeech*, 2886–2889.

Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., & Sahli, H. (2016). Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 89–96).

Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C., & Jiang, D. (2017). 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).

Zhang, F., Haddad, S., Nakisa, B., Rastgoo, M. N., Candido, C., Tjondronegoro, D., & de Dear, R. (2017). The effects of higher temperature setpoints during summer on office workers' cognitive load and thermal comfort. *Building and Environment, 123*, 176–188.

Zhang, X., Shen, J., ud Din, Z., Liu, J., Wang, G., & Hu, B. (2019). Multimodal depression detection: Fusion of electroencephalography and paralinguistic behaviors using a novel strategy for classifier ensemble. *IEEE Journal of Biomedical and Health Informatics, 23*(6), 2265–2275.

Zhao, Y., Dong, Q., Chen, H., Iraji, A., Li, Y., Makkie, M., … Liu, T. (2017). Constructing fine-granularity functional brain network atlases via deep convolutional autoencoder. *Medical Image Analysis, 42*, 200–211.

Zlotnik, A., Montero, J.M., San-Segundo, R. and Gallardo-Antolín, A., 2015. Random forest-based prediction of Parkinson's disease progression using acoustic, ASR and intelligibility features. INTERSPEECH-2015, 503- 507.