End-to-End Multimodal Clinical Depression Recognition using Deep Neural Networks: A comparative Analysis

Muhammad Muzammel[a], Hanan Salam[b] and Alice Othmani[a,*]

[a] *Université Paris-Est Créteil (UPEC), LISSI, Vitry sur Seine 94400, France*
[b] *Emlyon, 23 Avenue Guy de Collongue, 69130 Écully, France*

ARTICLE INFO

Keywords:
multimodal depression recognition
features fusion
deep learning
biomedical information processing
biomedical informatics

ABSTRACT

Background and Objective: Major Depressive Disorder is a highly prevalent and disabling mental health condition. Numerous studies explored multimodal fusion systems combining visual, audio, and textual features via deep learning architectures for clinical depression recognition. Yet, no comparative analysis for multimodal depression analysis has been proposed in the literature.

Methods: In this paper, an up-to-date literature overview of multimodal depression recognition is presented and an extensive comparative analysis of different deep learning architectures for depression recognition is performed. First, audio features based Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) are studied. Then, early-level and model-level fusion of deep audio features with visual and textual features through LSTM and CNN architectures are investigated.

The performance of the proposed architectures using an hold-out strategy on the DAIC-WOZ dataset (80% training, 10% validation, 10% test split) for binary and severity levels of depression recognition is tested. Using this strategy, a set of experiments have been performed and they have demonstrated: (1) LSTM-based audio features perform slightly better than CNN ones with an accuracy of 66.25% versus 65.60% for binary depression classes. (2) the model level fusion of deep audio and visual features using LSTM network performed the best with an accuracy of 77.16%, a precision of 53% for the depressed class, and a precision of 83% for the non-depressed class. The given network obtained a normalized Root Mean Square Error (RMSE) of 0.15 for depression severity level prediction. Using a Leave-One-Subject-Out strategy, this network achieved an accuracy of 95.38% for binary depression detection, and a normalized RMSE of 0.1476 for depression severity level prediction. Our best-performing architecture outperforms all state-of-the-art approaches on DAIC-WOZ dataset.

Conclusions: The obtained results show that the proposed LSTM-based surpass the proposed CNN-based architectures allowing to learn temporal dynamics representations of multimodal features. Furthermore, model-level fusion of audio and visual features using an LSTM network leads to the best performance. Our best-performing architecture successfully detects depression using a speech segment of less than 8 seconds, and an average prediction computation time of less than 6ms; making it suitable for real-world clinical applications.

1. Introduction

Clinical depression (also known as Major Depressive Disorder) is a highly prevalent and disabling mental health condition. Nearly 4.4% of the world's population (i.e. 322 million people) are living with depression [1]. A survey paper reported the rise of depression among adults with the increase of age in Europe and its elevated risk for suicidal behavior [2].

Clinical depression diagnosis is a major challenge for health professionals. It lacks the biological gold standards [3, 4], and like most mental health conditions, it cannot be detected by a blood test or an imaging test. The most common approach for depression diagnosis constitutes clinical interviews [5, 6] and self-report scales and inventories (Self-RIs) [7, 8, 9, 10]. Clinical interviews scales are completed by a trained mental health professional in the context of a clinical interview. While such scales allow the assessment of various depression symptoms, however, their reliability is questioned [11, 12]. On the other hand, although Self-RIs alone are insufficient to support the diagnosis of depression, they are widely used screening tools in primary health care [13, 14]. Compared to clinical interviews, Self-RIs suffer from several shortcomings. For instance, such scales ignore the clinical significance of the reported symptoms, and do not account for individual personality characteristics, comorbid conditions, and life events that might have triggered depressive symptoms in the patient [15]. Additionally, both clinical interviews scales and Self-RIs are vulnerable to intentional or unintentional reporting bias including subjective bias, central tendency (avoiding extreme responses), and social acceptance [16].

Recently, automatic mental states and mental disorders recognition have attracted considerable attention from the Artificial Intelligence community. Several systems have been developed to automatically assess the person's emotions and mental state [17, 18, 19, 20, 21]. In these approaches, verbal and non-verbal behaviours were investigated.

Particularly, speech has been proven to be a robust indicator in the automatic diagnosis of clinical depression. In several works, it has been demonstrated that the speech modality outperforms visual and textual modalities in automatic clinical depression diagnosis. [22, 23]. However, the use of different depression benchmark datasets and estimation approaches makes it hard to conclude which acoustic features

[*] Corresponding author: Associate professor. Dr. Alice OTHMANI
✉ alice.othmani@u-pec.fr (A. Othmani)
ORCID(s):

have the most discriminating power for depression assessment [24, 25]. Yet, the Mel Frequency Cepstral Coefficients (MFCCs), which are considered as top audio features in applications like speech and speaker recognition [26], have proven their high efficiency in detecting clinical depression compared to other audio features in shallow-based approaches [24, 27]. Additionally, verbal information from speech transcription were also investigated as predictors of depression [23, 28, 29, 30].

Furthermore, depression is also conveyed through visual non-verbal cues. Several studies have reported promising results on the use of facial Action Units (AUs), introduced by Ekman et al. [31], for automatic depression assessment [18, 22, 32]. For instance, the first use of AUs for this task was performed in Cohn et al. [18], where statistical features such as the frequency of occurrence, mean duration, onset/total duration, and onset/offset ratios were extracted.

Moreover, numerous studies in the literature use multimodal fusion systems combining visual, audio and textual features for depression recognition [22, 32, 33]. The fusion of MFCC-based audio and video modalities leads to high-performance depression prediction results [32, 33]. Yet, these studies lack the information about model level fusion of MFCC features with other modalities. Also, they did not investigate the temporal dynamics' representations for multimodal depression recognition.

The efficiency of deep learning techniques in various intelligent applications has led to the exploration of different deep architectures for automatic depression assessment [34, 35]. Deep neural networks have significantly improved the performance on such task compared to shallow machine learning approaches due to their capability to automatically abstract both low and high level descriptors from the patient's verbal and non-verbal signals without the need of human intervention. Both, Convolutional Neural Networks (CNNs) [36, 37] and Recurrent Neural Networks (RNNs) [29, 33] based approaches were proposed in the literature for automatic depression recognition. CNNs were proven to be very efficient in modeling non-sequential visual data (by employing filters within convolutional layers to transform data). However, their incapacity to model temporal information leads to lower performances when used on sequential data. On the other hand, RNNs have the ability to interpret temporal information present in sequential data by reusing activation functions from preceding or succeeding data points in the sequence to influence the output and make better predictions. Despite all the deep-learning based approaches proposed in the literature for depression recognition, there is no reported work that investigates the performance of CNNs compared to RNNs for this task.

In this paper, a comparative analysis of several deep MFCC-based multimodal depression recognition frameworks is proposed. Learning the temporal dynamics from continuous and spontaneous data are considered. Two unimodal representations based on CNNs and RNNs allowing to learn high level MFCC audio features are proposed and compared. Moreover, an extensive study is performed to investigate the best suited multimodal fusion approach for clinical depression recognition. Early and model-level fusion of MFCC features with Word2Vec textual and AU visual features are investigated.

The remainder of this article is organised as follows. In section 2, a comprehensive review of related works in depression assessment and multimodal fusion is presented. The motivations and contributions of this paper are presented in section 3. In section 4, the proposed approach and deep learning architectures for depression assessment are presented. In section 5, details about the different performed experiments and results are presented. Finally, the findings are discussed in section 6, and section 7 concludes the paper.

2. Related Work

Traditionally, clinical interviews and self-report scales and inventories (Self-RIs) are used by health practitioners to assess clinical depression. The Diagnostic and Statistical Manual of Mental disorders (DMS) [5] and the Hamilton Depression Rating Scale (HDRS) are the most popular scales used in clinical interviews settings [6]. These scales are completed by a trained mental health professional in the context of a clinical interview. They allow the assessment of various depression symptoms such as mood swings, suicidal ideations, loss of interest in life, insomnia, anxiety, agitation, somatic symptoms, etc. Yet, both scales are widely criticized regarding their reliability [11, 12]. In addition, the assessment rely on clinicians subjective assessments which might present subjective biases [38] due to the clinician's skill.

On the other hand, self-report scales and inventories (Self-RIs) include Hospital Anxiety and Depression Scale (HADS) [7], Quick Inventory of Depression Symptomatology (QIDS) [8], Beck's Depression Inventory (BDI) [9], and the most commonly used Patient Health Questionnaire (PHQ) [10]. Although Self-RIs alone are insufficient to support the diagnosis of depression, they are widely used as screening tools in primary health care [13, 14]. Compared to clinical interviews, Self-RIs suffer from several shortcomings. For instance, such scales ignore the clinical significance of the reported symptoms, and do not account for individual personality characteristics, comorbid conditions, and life events that might have triggered depressive symptoms in the patient [15]. Additionally Self-RIs are vulnerable to intentional or unintentional reporting bias including subjective bias, central tendency (avoiding extreme responses), social acceptance [16], and the reliability of a patient's perception of their mental state [38].

Consequently automatic depression assessment approaches emerged which offer an objective way of mapping patient's verbal and non-verbal cues to a depression score. In the following, the existing methods for automatic depression assessment using redvisual-based, audio-based and text-based features are reviewed. Then, multimodal features fusion approaches for automatic depression recognition are reported. In addition, state-of-the-art approaches for learning temporal dynamics' representations are studied and its importance is highlighted.

2.1. Automatic Depression Assessment

Automatic depression assessment approaches are constituted mainly of three steps: 1) *data acquisition* where unimodal or multimodal data are acquired from various sources (audio, video [36, 37], text [29, 39], context [40], etc..) and the depression score ground truth is collected simultaneously using clinically validated scales, 2) *data processing* where data are pre-processed, and depression markers extracted (features), and 3) *prediction* where machine learning models are applied to predict the individual's depression state.

2.1.1. Visual-based Depression Assessment

Visual-based features approaches for depression assessment rely on the extraction of visual features from patients' image sequences or video. For a comprehensive review of depression recognition based on visual cues, please refer to the survey of [41]. Visual features learned for depression recognition concern mainly head and body features. While some approaches learn handcrafted features, other approaches investigate more complex and abstract features representations using deep learning [41].

Handcrafted head features include non-verbal cues extracted from the head pose or face. Previous studies extracted visual cues in the form of facial expressions [42], Action Units [18, 22, 23, 32, 43], gaze and pupil dilation [44, 45], eyelids movements and blinks [43, 45, 46], facial landmarks [30, 43], facial appearance [47] which can be extracted with several methods such as Local Binary Patterns (LBP), Edge Orientation Histogram (EOH) and Local Phase Quantization (LPQ) [48], and head motion and orientation [43, 49].

Handcrafted body features encode the body's orientation and movement, which carry significant information about an individual's mental state. For instance depressed individuals tend to move less. Few approaches have concentrated on the investigation of handcrafted body features for depression recognition. Example features used in the literature include Space-Temporal Interesting Point (STIP) features that describe the body movements [50] and the so-called "parts algorithm" which allows extracting the body's orientation and distance [51].

Deep visual features offer an automatic abstraction of facial image descriptors contributing to the manifestation of depression, which has proven to be a good diagnostic biomarker for depression recognition. The full face or specific facial regions are either used as input to deep architectures [48, 52, 53], or handcrafted visual features are extracted *a priori* and then used as input to the deep neural network [37]. For instance, in the approach of [37], spectral representations of a set of facial non-verbal behaviour features (AUs, gaze, head pose) were used as inputs to the deep learning model.

Deep learning architectures, in particular, Convolutional Neural Networks (CNN) were used for depression recognition from images sequence. These include Deep Transformation Learning (DTL) [54], 2D-CNN [36, 37], Artificial Neural Network (ANN) [37], 3D-CNN [52], and 3D-CNN followed by RNN [55]. VGG-Face [56] and AlexNet [48], initially developed for facial recognition, were also used for depression recognition from images by [48]. Particular architectures tried to tackle the challenging aspects that might be present in visual data. For instance, to account for varying head poses and imaging conditions in facial images, [57] introduced the memory attention mechanism. In this approach, an attention module acts as a pooling layer by adaptively learning the weights emphasising or suppressing face images with varying poses and imaging conditions. On the other hand, the deep architecture of [58], DepressNet, attempted to model the depression patterns visually encoded in the face. This was done through the generation of a depression activation map, which allows to identify salient regions of the input image in terms of depression severity score.

2.1.2. Audio-based Depression Assessment

Audio-based approaches for depression assessment rely on the extraction of acoustic and prosodic markers from patients' speech segments. These include low and high level features designed and extracted from the audio signal in an attempt to model the characteristics of speech such as prosody, voice quality, frequency range, energy, etc. Extracted features are then fed to main-stream classifiers to predict depression [25, 32, 59, 60, 61].

Acoustic features used for depression assessment can be categorized into six categories: Prosodic, Source, Formant, Spectral [62], Cepstral and deep learning features. These features have demonstrated that they contain relevant information about the depressed speech. For a comprehensive review of audio-based depression recognition, please refer to [62].

Prosodic features represent phoneme-level variations in speech rate, rhythm, loudness, intonation and stress [25, 29, 60, 63]. Examples include the fundamental frequency (F0) and energy which represent the perception of pitch and loudness [62].

Source features capture information of the voice production source. Such features parameterise the air flow from the lungs through glottis via glottal features [25, 64], Teager Energy Operator (TEO) features [64] or vocal fold movements via voice quality features [32, 60, 63].

Formant features contain information concerning the physical vocal tract properties such as the muscle tension in the form of formant frequencies (F1, F2, F3) that are affected by the depression state of the patient [29, 65].

Spectral features characterise the speech spectrum which constitutes frequency distribution of the speech signal at a specific time instance [25, 29, 60, 63]. Examples of spectral features used in the literature include spectral flux, energy, slope and flatness [25, 64, 66].

Cepstral features are those based on a non-linear spectrum-of-a-spectrum representation. The most common used are Mel-Frequency Cepstral Coefficients (MFCC) [63] and Linear Prediction Cepstral Coefficients [25, 61, 64, 67].

Deep audio features can be either learned from raw audio input [68], or from acoustic features, extracted from the audio signal and used as input to the Deep Neural Network (DNN). For instance, [69] proposed a self-supervised pre-trained audio embedding method to extract audio descriptors for depression recognition. Other audio features used as input to DNN include Mel-scale filter bank [19], MFCC features [19, 33, 34, 35], spectogram [34, 70], prosodic features [35], spectral and voice quality features [23].

Proposed deep learning architectures include Feed-Forward Neural Network (FF-NN) [43], Convolutional Neural Network (CNN) [34, 70], Long Short-Term Memory Convolutional Neural Network (LSTM-CNN) [19, 68], Bidirectional Long Short-Term Memory Convolutional Neural Network (BLSTM-CNN) [35, 69], Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [29, 33], BLSTM-RNN [35], and Deconvolutional Neural Network Multiple Instance Learning (DNN-MIL) [35].

2.1.3. Text-based Depression Assessment

Text-based approaches for depression assessment rely on the extraction of descriptors from the patients' speech transcriptions. In the literature, two categories of textual features are reported for depression assessment: statistical and text-to-vector embedding.

Statistical features compute statistics on spoken utterances of the patient and/or the interviewer (in case of clinical interviews). Statistics can be calculated on the detected sentiment from the textual data such as arousal, valence [63], pleasure and dominance [43]. Statistics can also be extracted directly from the verbal responses of the participant or interviewer. Example features include the total number of sentences, spoken words, average words spoken in each sentence, laughter to spoken words ratio, depression related words to spoken words ratio [43].

Text-to-vector embedding features convert textual data to numerical vectors, which are then input to Machine Learning models. For instance, Paragraph Vectors (PV), also called Doc2Vec, which project text documents (phrases, paragraphs, etc.) into d-dimensional space were employed as textual modalities for depression assessment [23, 28, 29, 30].

2.2. Multimodal Features Fusion

Several approaches for depression recognition are based on features extracted from different modalities, such as visual, audio and text. An important step in multimodal recognition approaches is modalities fusion, which tackles fusion strategies of the different modalities, or even different features from the same modality. Four categories of strategies are reported in the literature: feature-level, decision-level, hybrid and model-level fusion [71, 72]. For a comprehensive survey of multimodal fusion strategies for multimedia analysis, please refer to [73]. Table 1 presents a summary of state-of-the-art approaches employing multimodal fusion for depression recognition, categorized by the fusion level.

Feature-level fusion, also called **Early-fusion**, is the most commonly used strategy for multimodal recognition systems. It concerns approaches which concatenate features extracted from different modalities into a single high-dimensional feature vector at an early stage, immediately after extraction. For instance, in the approach of [63], the high dimensional feature vector resulting from the early fusion of audio and textual features is fed to an SVM classifier. Similarly, the approach of [28] extracted session-level audiovisual features using a multimodal Deep Denoising Autoencoder (multiD-DAE) followed by a Fisher Vector encoding. The audiovisual features were then fused with textual features in a multi-task Deep Neural Network (DNN) framework.

Decision-level fusion or **Late-fusion** concerns approaches that perform fusion at the decision level. After obtaining a decision based on each of the modalities, an integration step is performed on these decisions by applying an algebraic combination rule (e.g. minimum, maximum, sum, etc.) of the multiple predicted class labels. For instance, [74] implemented a decision-level fusion method based on the decisions obtained by Random Forest regressors on visual, audio and text modalities. A modality-wise confidence score was computed using the standard deviation of the outcomes of all trees for each modality. Decisions were fused based on a winner-takes-all strategy were the modality prediction with the highest confidence score was chosen as the final result .

Hybrid fusion performs early fusion in addition to fusion of individual classification scores of each used modality at the decision level. For example, [23] fused the prediction results of audiovisual and text-based models using a multivariate regression model. Two audiovisual Deep Convolutional Neural Network (DCNN) models are trained separately to predict depression and non-depression scores, which are then fused with binary depression classification results obtained from two text-based models.In their text-based model, statistical utterance features were fused at an early level to detect a binary depression score with a Random Forest classifier.

Model-level fusion aims to learn a joint representation of the different modalities with an extended learning after unimodal features extraction and concatenation. Fusion at the model-level is the most adapted way for fusing multimodal features.

In this family of approaches, Multiple Kernel Learning (MKL) [75], graphical models [76], or also neural networks-based approaches [71, 77, 78, 29, 30, 33] have been employed. MKL is used to find an optimal combination of the input modalities' features by constructing a kernel for each feature type [75].

Graphical models are able to easily exploit the spatial

Table 1
State-of-the-art multimodal fusion systems for depression (and emotion) recognition

Fusion Level	Paper	Modalities	Model
Decision	[74]	Audio, Video, Text	Random Forests
Hybrid	[23]	Audio, Video, Text	DCNN-DNN
			SVM
			Random Forests
Feature	[28]	Audio, Video, Text	DNN
	[63]	Audio, Text	SVM
	[78]	Audio, Video	LBP-TOP
			CNN
			CNN-LTSM
Model	[29]	Audio, Text	LSTM per modality
	[30]	Audio, Video, Text	DCNN-DNN
	[33]	Audio, Video	LSTM
	[77]	Audio, Video	LTSM
	[71]	Audio, Video	CNN per modality
			DBN for fusion
			SVM for classification
	[78]	Audio, Video	Probabilistic Graphical Model
	[75]	Audio, Video	MKL per modality
			SVM

and temporal structure of the multimodal data by incorporating multiple simple predictors and exploiting the temporal correlation between multiple time points and multiple input modalities [76].

Neural Networks have been widely used for audiovisual fusion. Feature representations of the different modalities are first learned with a different architecture for each modality. Multimodal data are then fused by concatenating the learned feature representations followed by an additional hidden layer. Mainly, Recurrent Neural Networks were used due to their ability of incorporating temporal information. For instance, audiovisual segment features are fused using a Deep Belief Network (DBN) [71], LSTM [77] or Bi-directional LSTM (BLSTM) [78]. Similarly, in the approach of [29], audio and text modalities are trained separately using BLSTM. Then, a multimodal model is trained after combining these two modalities by merging the outputs into a final feed-forward network. In [30] high level features from video, text, and audio data are learned using DCNN-DNN networks and then fused at the model-level with a DNN network. Authors of [33] extracted MFCC-based audio features which were fed to a series of LSTM networks and then fused with AU descriptors with two fully connected layers and one output layer.

It has been reported in the literature that model-level strategy achieves better performances than other strategies. Particularly, Neural Networks-based approaches for modality fusion obtain good performance thanks to their capacity of learning from large amounts of data in addition to finding complex decision boundaries [76].

2.3. Learning Temporal Dynamics Representation

Multimodal data are characterized by a dynamic nature represented by the signals in two dimensions: spatial and temporal. Thus, three representations could be learned [79]:

Spatial-features' representation consider the spatial dependencies of extracted features from each single modality without considering the temporal correlation between time steps. Spatial information is learned in the literature with 2D CNN coupled with Feature Dynamic History Histogram (FDHH) [48] to map features variations, or through fusion of multiple 2D CNN on different facial regions [58]. In [57], the authors fused video features by employing an attention mechanism.

Temporal-features' representation where features are learned from temporal data sequences. Consequently, temporal dynamics are learned within the features' representations and the contextual information of the temporal dimension is considered. It has been demonstrated that Recurrent Neural Networks capture the temporal dynamics present in the data [80]. To represent the temporal dynamics of expressive behaviour in video-based automatic depression analysis, spectral representation of such behaviors is extracted [37]. The constructed spectral signals of all behaviours are then aligned and fed to CNN and ANN for depression analysis.

Joint spatial and temporal features representation where both spatial and temporal modeling of the data is performed. Once features are extracted from multiple modalities at different time points, they are fused using one of the modality fusion strategies [81]. Spatial and temporal information can be exploited separately with DNN by cascading 2D CNN with RNN. However, such approach can degenerate intrinsic spatio-temporal relationships [52]. Alternatively, 3D-CNN (C3D) permit to leverage spatio-temporal information [52, 55, 82]. In the work of [52], fusion of diverse C3D predictions is proposed, where spatio-temporal features are extracted from global and local regions of the face. The network integrates a 3D Global Average Pooling layer instead of fully-connected layers, allowing to efficiently summarize spatio-temporal features, together while reducing the number of the model parameters and avoiding overfitting. On the other hand, [36] propose a novel temporal pooling method to capture and to encode the spatio-temporal dynamics of video clips into an image map. Their method is based on two-stream model that performs late fusion of appearance and dynamic information.

3. Motivations and Contributions

In this paper, low-level and high-level audio features are used in Deep Neural Networks frameworks to assess Major Depressive Disorder from speech. Mel Frequency Cepstral Coefficients (MFCCs) audio features have proven their high efficiency in detecting clinical depression compared to other audio features in shallow-based approaches [24, 27, 33]. In addition, they are considered as top audio features in speech-based applications like speech and speaker recognition. To thoroughly explore the discriminative power of MFCC features for depression assessment, different MFCC-based deep learning architectures are proposed and compared in this work.

Previous studies showed that Recurrent Neural Network attained significantly high performance in speech recognition [29, 35]. Therefore, we investigate the performance of the Short Long-Term Memory (LSTM) based networks and compare their performance with Convolutional Neural Networks for depression assessment. Therefore, two unimodal audio representations based on CNNs and RNNs are proposed and compared, allowing to learn high level MFCC representations from low level features and to classify depression.

Furthermore, deep model-level fusion of deep audio features with visual and textual features through LSTM and CNN architectures is investigated. This results in a higher performing deep neural network architecture for multimodal depression recognition.

The contributions of this research can be summarized as:

- Comparing different strategies of deep features learning and fusion for clinical depression recognition.
- A comparative analysis of several deep neural networks architectures for multimodal depression recognition.
- Learning temporal dynamics representations from multidimensional signals and modalities for clinical depression recognition.
- An automatic and high performing approach for detecting depression in less than 5.26 milliseconds, based on input segments of less than 8 seconds from multimodal data.

4. Methods

Audio, visual, and textual data are investigated in this work and they are used for unimodal and multimodal representations in several deep learning frameworks.

The proposed approach is based on clinical interviews data from the DIAC-WOZ dataset [83]. The data constitute conversations between participants (patients) and an Embodied Conversational Agent (ECA), playing the role of the interviewer. During the conversation, the patient responds to a clinically validated questionnaire assessing his/her depression level. A detailed description of this dataset is provided in section 5.1.

The proposed approach is constituted of five steps. First, a preprocessing step is performed (section 4.1) where patients' speeches are extracted and the corresponding audio signals are divided into fixed-size windows (section 4.1.1). Audio data are synchronized with visual and textual data, and textual data are pre-processed to clean noisy information (4.1.2). Visual features are provided in the used dataset, and thus, no preprocessing is needed for visual data. In a second step, audio data augmentation is performed to minimize overfitting and data scarcity problems relative to deep learning (section 4.2). Then, low or high-level features are extracted from the different modalities (section 4.3). Afterwards, multimodal features fusion strategies for automatic depression detection are studied in section 4.4. The last step of the proposed approach concerns the final classification of the learned multimodal representations (section 4.5).

4.1. Preprocessing

A preprocessing step is applied only to audio and to textual data prior to features extraction as follows. Visual features are provided in the used dataset, and thus, no preprocessing is needed for visual data.

4.1.1. Audio Preprocessing

The proposed approach is based on depression assessment from patients' responses to clinical questions asked by an interviewer. Audio recordings are preprocessed in order to separate the patient's speech from that of the interviewer. For each audio recording, timestamps relative to the interviewer and the participant's speech are provided. These timestamps are used to retrieve the participant's speech. The speech segments relative to the interviewer's speech are discarded and only the patient's speech is used for automatic depression detection. The participant's audio is then divided into small speech segments of size $n = 7.6$ seconds.

4.1.2. Text Preprocessing

The audio recordings of the clinical interviews are accompanied with speech transcriptions corresponding to the conversations between the participant and the interviewer. Table 2 presents an extract of a speech transcript from the DAIC-WOZ dataset. Contrary to the audio data, we use both the transcriptions of the interviewer and the participant. As a matter of fact, verbal reactions of the interviewer following those of the participant, might carry relevant information about the participant's emotions encoded in their responses. For instance, when the participant replies negatively to the interviewer's questions, the interviewer's responses include phrases like “that sucks” or “I'm sorry to hear that” which caries significant information about the depressive state of

Table 2
An extract of a speech transcript from the DAIC-WOZ dataset.

Start Time	Stop Time	Speaker	Value
81.03	82.23	Ellie	where are you from originally
82.72	83.69	Participant	los angeles
84.56	85.02	Ellie	really
86.47	88.41	Ellie	what are some things you really like about LA
89.85	90.62	Participant	um
92.59	94.81	Participant	well <laughter> that's a good question
95.84	100.56	Participant	um I like the familiarity with everything I know where everything is in the city
101.89	102.38	Ellie	mhm

the participant. We have made the choice of using solely the audio patterns of the participant because the audio patterns of the interviewer do not present signs of depression but his/her words can present signs of sympathy when the patient is depressed.

First, participant's speech transcriptions (text) are synchronized with her/his speech (audio). This is done using the timestamps provided with the text transcript files (cf. Table 2). Then, the participant's transcriptions are extracted for the same fixed size windows duration as the audio data segments (7.6 s). The corresponding interviewer's question is then identified and extracted from the start and stop time, and the "Speaker" value.

A set of preprocessing techniques is applied to clean the textual data. These include: (1) removal of numerical numbers, punctuation and white spaces, (2) grammar correction using the language python tool[1], and (3) removal of stop words, lower case conversion and lemmatization using NLTK toolbox [84]. Note that lemmatization is the process of reducing inflectional and derivationally related forms of a word to a common base form, known as the lemma. A lemma is referred to as the canonical, dictionary, or citation form of a word.

4.2. Audio Data Augmentation

The technique of artificially expanding labeled training sets by transforming data points while preserving class labels, known as data augmentation, is considered in this work. Such technique allows handling the labeled data scarcity problem, avoid overfitting relative to deep neural networks, and improve the performance of the proposed approach and its robustness to noise. Two types of audio augmentation techniques are then performed on the audio frames to perturb the raw audio signals and generate new ones [33].

- **Noise Injection**: a random white noise is added to the speech segments of participants. If y is the audio signal and α is the noise factor, then the noise augmented data x is given by: $x = y - \alpha \times rand(y)$. We use $\alpha = 0.01$, 0.02 and 0.03.
- **Pitch Augmentation**: audio frames pitch is lowered and the audio duration is kept unchanged. Audio frames pitch is lowered by 0.5, 2, 2.5 in semitones.

4.3. Data Encoding

In this section, the audio, visual, and textual features used in this work are presented.

4.3.1. Audio Features

A good representation of the audio signal can allow the discrimination between depressed and non depressed subjects. A negative emotion, such as when a person is sad or bored, is translated by slower speaking frequencies. Thus, low level audio features to describe the variation of low frequencies signal are needed. This makes the Mel Frequency Cepstral Coefficients (MFCC) good candidates for this task. MFCC are thus extracted from the patient's preprocessed audio signals. Following MFCC features extraction, high level deep audio features are learned from MFCC to further encode audio patterns for depression recognition. In the following, MFCC features extraction and deep audio features encoding are described.

MFCC Extraction – MFCC features describe the audio cepstrum energies in a non-linear scale known as the mel-scale. Tracking MFCC variations over time allows tracking the speech tone variation [85] which is largely affected in depressed speech. To extract MFCC features, the speech signal is first divided into frames by applying a Hamming window function of window length of 60 milliseconds. Let $s[n]$ be the original audio signal, $w[n]$ the hamming window function, then the sliced audio frame is given by:

$$x[n] = w[n]s[n] \tag{1}$$

with

$$w[n] = \alpha - \beta \cos\left(2\pi \frac{n}{N-1}\right) \tag{2}$$

where $\alpha = 0.54$, $\beta = 0.46$, N represents the length of the window, and $0 \leq n \leq N$.

Discrete Fourier Transform (DFT) is then computed for each frame to extract information in the frequency domain.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi}{N} kn} \tag{3}$$

[1] https://pypi.org/project/language-tool-python/

Mel-scale filter banks are then applied to the DFT power spectrum to map the frequency information to the Mel scale which approximates the human perception of pitch.

$$Y_t[m] = \sum_{k=1}^{N} W_m[k]|X_t[k]|^2 \tag{4}$$

where W_m represent the triangular Mel-scale filter banks, k is the DFT bin number, $1 \le k \le N$, and m represents the mel-filter bank number, $1 \le m \le N$.

This results in a power spectrum. The spectrum is then smoothed and 19 spectral components are collected in the Mel frequency scale.

Only the logarithm of the amplitude spectrum is retained and a cepstral feature vector is then generated. The components of the Mel-spectral vectors calculated for each frame are highly correlated. Therefore, the Karhunen-Loeve (KL) transform is applied which is approximated by the Discrete Cosine Transform (DCT).

$$y_t[n] = \sum_{m=0}^{M-1} log\left(Y_t[m]\right) cos\left(n(m+0.5)\frac{\pi}{M}\right) \tag{5}$$

This results in the mel frequency cepstral features matrix of 378 × 60 for each 7.6 seconds audio signal.

Deep Audio Features Encoding – Following MFCC features extraction, high level deep audio features are learned to further encode audio patterns for depression recognition. CNNs were proven to be very efficient in modeling non-sequential visual data (by employing filters within convolutional layers to transform data). However, their incapacity to model temporal information leads to lower performances when used on sequential data. On the other hand, RNNs have the ability to interpret temporal information present in sequential data by reusing activation functions from preceding or succeeding data points in the sequence to influence the output and make better predictions. To compare RNNs and CNNs for depression recognition, two unimodal representations of MFCC features are proposed in this work: (1) CNN-based and (2) RNN-based. Two deep neural networks architectures are studied based on these unimodal representations of the audio signal for clinical depression recognition. All implementation details of the proposed deep architectures are provided in section 5.3.

1. **MFCC-CNN** (Figure 1a). In the given network, high level MFCC-based descriptors are learned with an architecture consisting of 2 consecutive blocks, each includes consecutive convolutional and Relu layers, followed by dropout and flatten layers.
2. **MFCC-LSTM** (Figure 1b). In this network, MFCC-based high level descriptors are learned with an RNN architecture. It consists of 3 blocks. Each block is composed of an LSTM layer followed by batch normalization and dropout layers. There is flatten layer at the end of third block.

Table 3
The list of the Visual Action Units in the DAIC-WOZ dataset.

Face Part	AU	Description
Upper Face	$AU01$	inner brow raiser
	$AU02$	outer brow raiser
	$AU04$	brow lowerer
	$AU05$	upper lid raiser
	$AU06$	cheek raiser
	$AU45$	eyes blink
Lower Face	$AU10$	upper lip raiser
	$AU12$	lip corner puller
	$AU15$	lip corner depressor
	$AU09$	nose wrinkler
	$AU14$	dimpler
	$AU17$	chin raiser
	$AU20$	lip stretcher
	$AU25$	lips part
	$AU26$	jaw drop
	$AU11$	nasolabial deepener
	$AU15$	lip corner depressor
	$AU23$	lip tightener
	$AU28$	lip suck

4.3.2. *Visual Features*

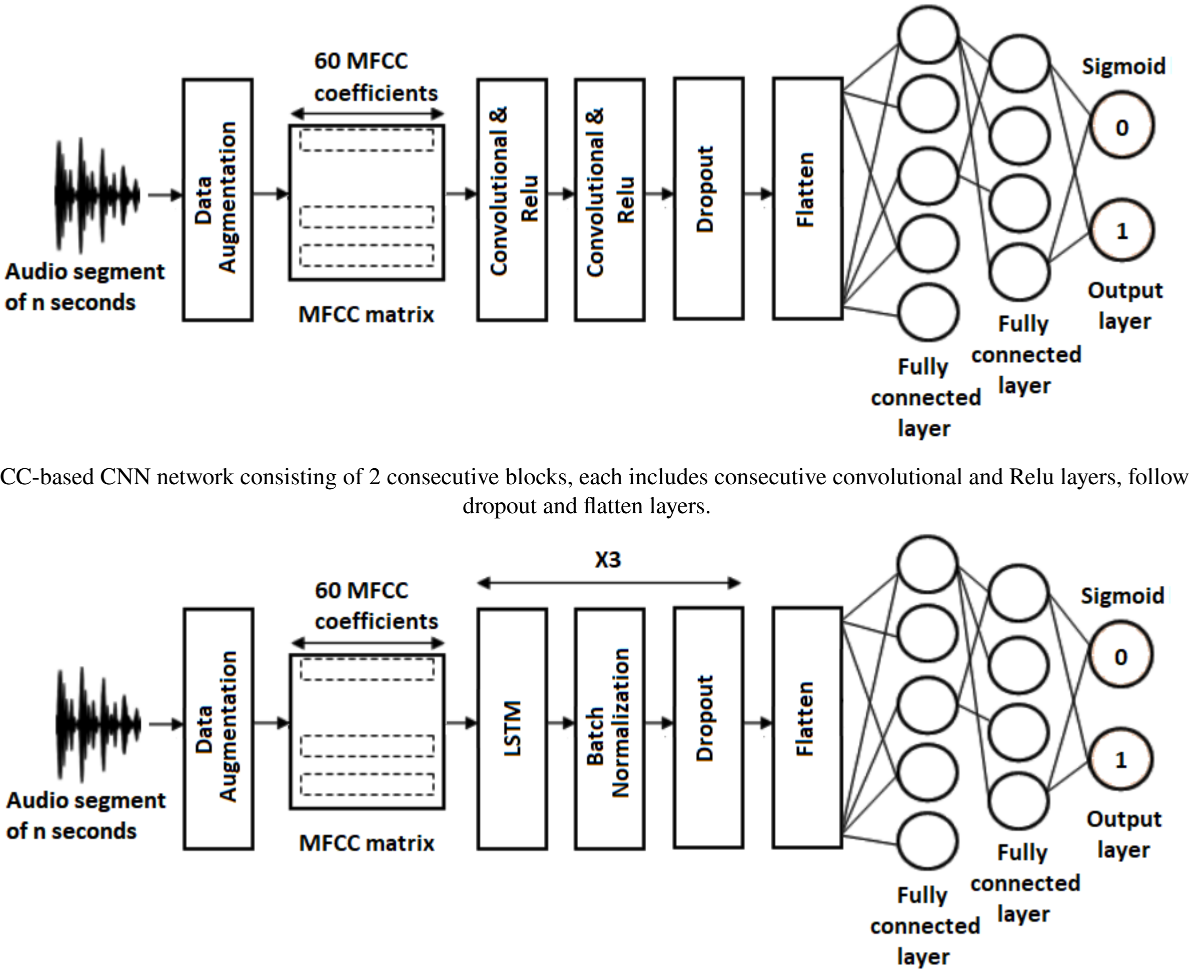
The publicly available baseline visual features in the DAIC-WOZ dataset are considered in this work. These visual features consist of facial Action Units (AUs), which were first introduced in the Facial Action Coding System (FACS) [86]. AUs refer to a set of facial muscle movements that correspond to a displayed emotion. Using FACS, any displayed emotion can be described in terms of a set of AUs. For ethical reasons, no raw video was made available in the DAIC-WOZ, which is the main reason behind limiting our work to this set of features. The visual features consist of 20 Action Units (AUs) [86] which are extracted from the upper and lower face of each subject using the OpenFace[2] Framework [87]. The extracted AUs are presented in Table 3. The participants' facial Action Units are synchronized with their speech and text transcripts using the provided timestamps.

4.3.3. *Textual Features*

As textual features, the words uttered by the participant and the interviewer are converted during each audio segment ($7.6sec$) to sequences of vectors using the word embedding approach, as shown in Figure 2.

Word embedding allows to transform words into dimensional vectors, where similar words or words that appear in the same context tend to be mapped to vectors which are close in the dimensional space. In particular, each frame transcript is represented by a matrix $E = (e_1, \dots, e_k, \dots e_{nw})$ where e_k is the word vector corresponding to the k^{th} word and nw is the number of words in the frame transcript. We use the fastText pretrained network [88], trained on Common Crawl[3] with sub-word information resulting in word vectors of size 300. Words that do not exist in the pretrained model

[2] https://github.com/TadasBaltrusaitis/OpenFace
[3] https://commoncrawl.org/2017/06



(a) MFCC-based CNN network consisting of 2 consecutive blocks, each includes consecutive convolutional and Relu layers, followed by dropout and flatten layers.

(b) MFCC-based LSTM network consisting of 3 blocks of an LSTM layer followed by batch normalization, dropout, and flatten layers.

Figure 1: MFCC-based deep unimodal representations for depression recognition. 60 MFCC coefficients are extracted from the audio segment of 7.6 seconds. Then, they are fed to a CNN (a) or to an LSTM (b) followed by two fully connected layers. The output layer is a dense layer of size 2 for binary depression recognition and of size 24 for PHQ-8 depression severity levels prediction with a sigmoid activation function.

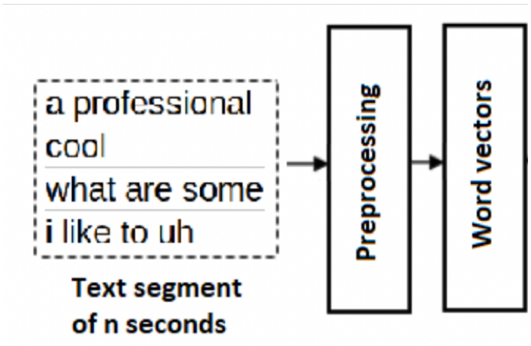


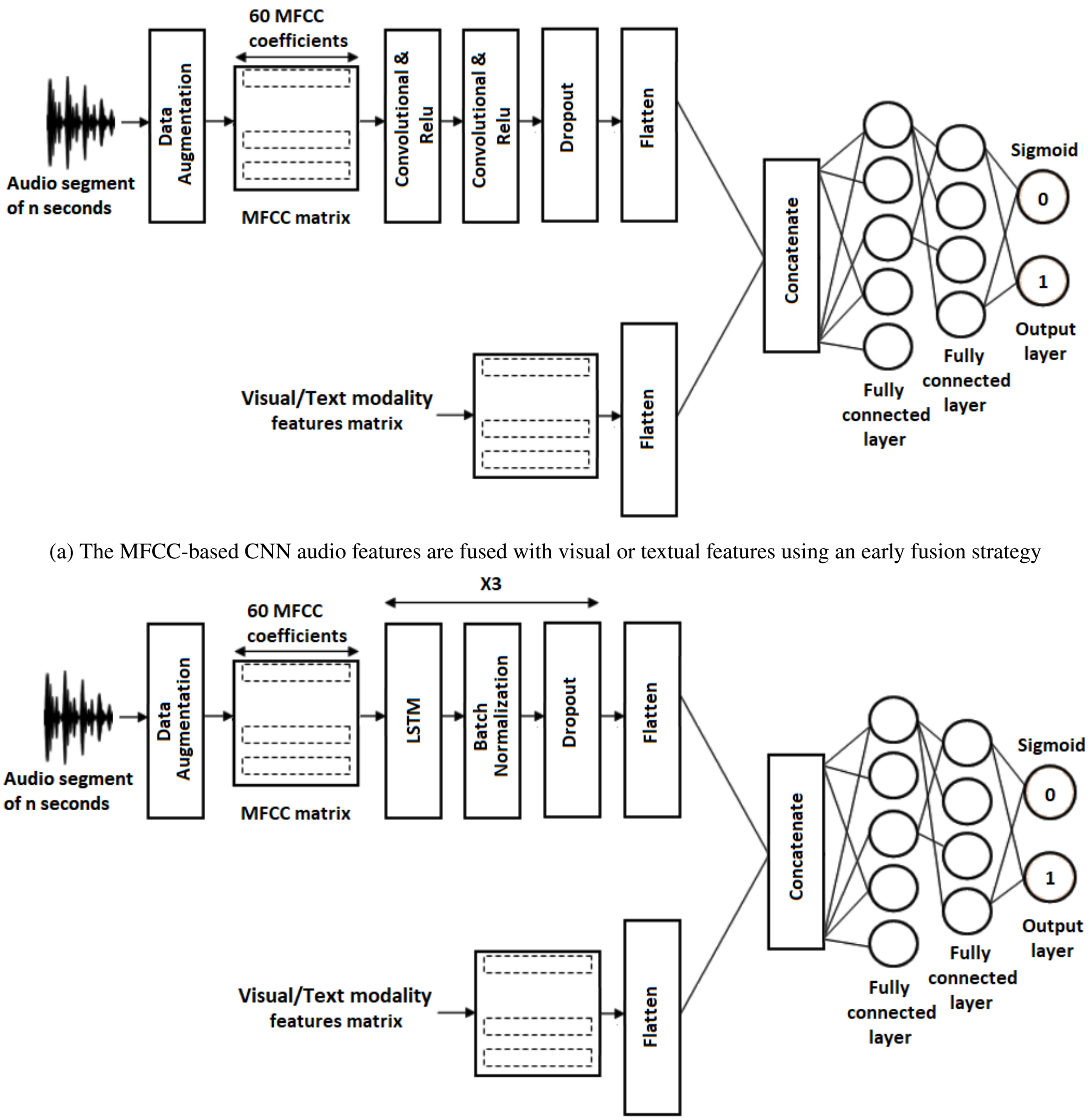
Figure 2: Process of textual features extraction. Text segments at first preprocessed to remove noisy information, then word vectors are extracted.

were replaced by their synonyms. Moreover, the resulting word vectors matrix for each transcript frame were resized to 378×9 where 378 corresponds to the MFCC coefficients size and 9 corresponds to the minimum number of words existing in one frame.

4.4. Multimodal Data Fusion

The deep encoded audio features are combined with textual and visual modalities following different data fusion strategies. The goal is to explore the best suited fusion strategy for multimodal depression assessment, as well as the best modality to be coupled with MFCC audio features for this task. Therefore, the CNN and LSTM based unimodal audio representations are fused with the textual and the visual modalities using two fusion strategies: (1) Early Fusion and (2) Deep Model-Level Fusion.

Early Fusion – The audio deep unimodal representation is concatenated with the visual or textual features, and the resulting vector is used to detect depression. The corresponding architecture is shown in Figure 3. Thus, four different



(a) The MFCC-based CNN audio features are fused with visual or textual features using an early fusion strategy

(b) The MFCC-based LSTM audio features are fused with visual or textual features using an early fusion strategy

Figure 3: Proposed multimodal early fusion strategy frameworks for depression recognition. After extrating the MFCC-based unimodal representations as described in Figure. 1, visual or text features are concatenated with the deep audio features into a single high-dimensional feature vector and then fed to two fully connected layers to predict depression. All implementations details are given in Section. 5.3.

deep neural networks are studied:

1. **MFCC-CNN AU** (Figure 3a). High level MFCC-based descriptors are learned with an architecture consisting of 2 consecutive blocks, each including consecutive convolutional and Relu layers, followed by dropout and flatten layers. Then, the obtained high level MFCC features are concatenated with the 20 visual Action Units.
2. **MFCC-LSTM AU** (Figure 3b). MFCC based high level features consisting of 3 blocks of LSTM layers are extracted. Then, the obtained high level MFCC features are concatenated with the 20 visual Action Units.
3. **MFCC-CNN Word2Vec** (Figure 3a). The network is the same as the MFCC-CNN AU. The AU visual features are replaced by the Word2Vec textual features.
4. **MFCC-LSTM Word2Vec** (Figure 3b). The network is the same as the MFCC-LSTM AU. The AU visual features are replaced by the Word2Vec textual features.

Deep Model-Level Fusion – The unimodal representations of the different modalities are concatenated and a joint representation is learned with an extended deep learning

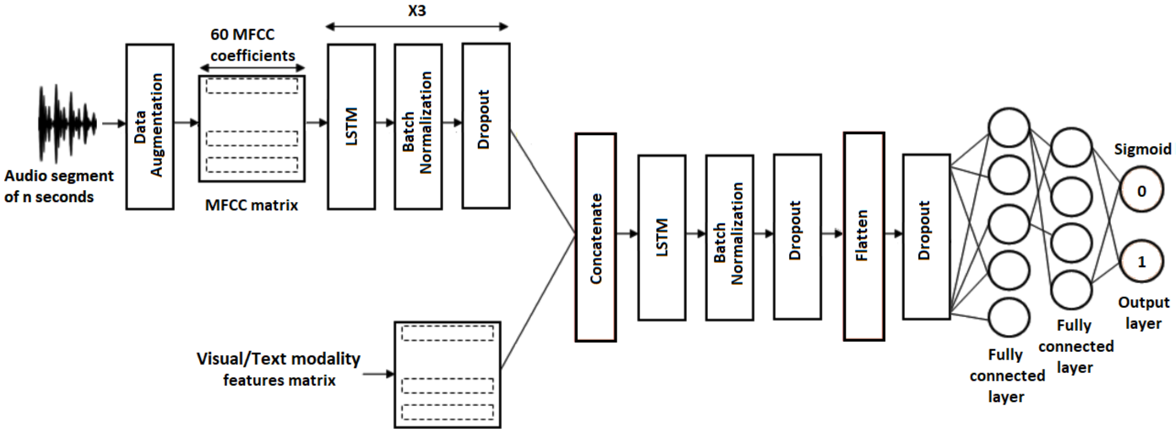


Figure 4: Proposed model-level fusion approach for depression recognition. The MFCC-based deep audio features are concatenated with the visual or the textual features and then fed to an LSTM-based deep neural network to learn a joint multimodal representation for depression recognition. All implementations details are given in Section. 5.3.

architecture. LSTM is used in order to learn a temporal representation of the two modalities. The architecture is shown in Figure 4. Two multimodal architectures fusing MFCC high level features with visual and textual features, respectively, are implemented as follows.

1. **MFCC-AU Model-Level Fusion**. MFCC based high level descriptors are learned with a deep architecture consisting of 3 blocks. Each block is composed of an LSTM layer followed by a batch normalization layer and a dropout layer. The obtained high level MFCC features are concatenated with 20 Action Units and fed to an LSTM layer followed by batch normalization, dropout, and flatten layers.
2. **MFCC-Word2Vec Model-Level Fusion**. The network is similar to that of the MFCC-AU model-level fusion. The AUs visual features are replaced by the Word2Vec textual features.

4.5. Depression Classification

All proposed frameworks and representations, whether unimodal or multimodal are fed to two fully connected layers followed by one output layer that implements the following equation:

$$\hat{l} = \sigma(W_2^T.(\beta(W_1^T.(\beta(W_0^T.f + b_0)) + b_1)) + b_2) \quad (6)$$

where f is the embedding vector, σ is the sigmoid function and β is a $tanh$ function. W_0, W_1, W_2 represent the first dense layer weights, second dense layer weights, and output layer weights, respectively. b_0, b_1, and b_1 are the bias vectors of the first dense layers, the second dense layers and the output layer, respectively. Both the weight matrices and the bias are learned through a training process and the classification is learned by optimizing the cross-entropy between the ground truth and classification outcome using the function defined by:

$$L = 1/N \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (7)$$

where L is the average loss for all the training samples, $\hat{Y}_i$ is the estimated depression label and Y_i is the real depression label.

5. Experiments and Results

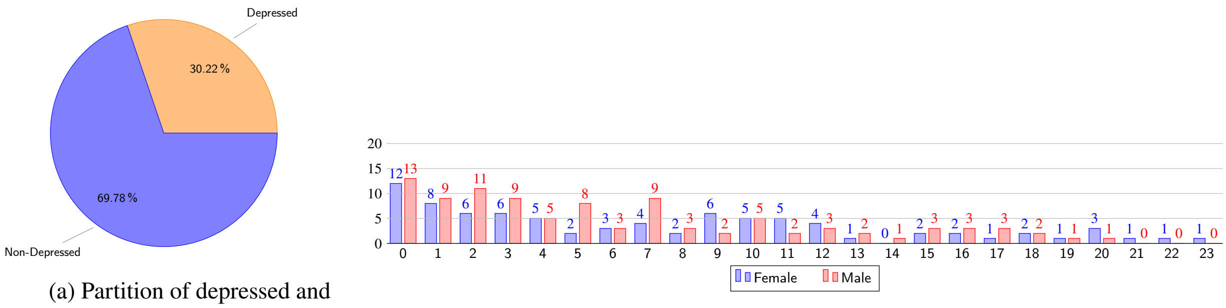
In this section, we present the evaluation results of the proposed approaches. A description of the used dataset is presented in section 5.1. The used evaluation metrics are presented in section 5.2. All the proposed deep networks implementation details are described in section 5.3. Finally, the performed experiments and corresponding results are presented and discussed in section 5.4.

5.1. Dataset

The performances of the proposed architectures are evaluated on the Distress Analysis Interview Corpus Wizard-of-Oz dataset (DAIC- WOZ) [83]. DAIC-WOZ was collected by the University of California and introduced in the Audio/Visual Emotion Challenge and Workshop in 2017 (AVEC 2017) [60]. The dataset is composed of clinical interviews between an interviewer and 189 participants aiming to investigate different psychological distress conditions such as depression, anxiety, and post-traumatic stress disorder. Due to technical reasons, in this work, only data of 182 participants have been used.

Multimodal data – The dataset includes the visual Action Units of the participants in addition to the audio recordings and the speech transcript files of the participants and the interviewer. In order to respect the privacy of the subjects who participated in the study, the participants' images are not included in the dataset. Transcript files and visual Action Units are timestamped. The average length of participants' audio recordings is 15 minutes obtained at a sampling rate of $16kHz$.

Depression labels – Participants' data are labeled in terms of depression severity level and a binary depression label



(a) Partition of depressed and non-depressed participants.

(b) Participants repartitions across the severity levels of the PHQ test.

Figure 5: Participants repartitions for binary depression and severity level scores within the DAIC-WOZ dataset. (a) Depressed versus Non-Depressed participants. (b) Participants repartitions across the twenty four depression severity levels given by the PHQ-8 questionnaire.

Table 4
Training, testing and validation DAIC-WOZ subjects distribution in our experiments following the hold-out strategy.

Dataset	Number of Subjects	Subjects (%)	Samples
Train	146	80	8191 (*57337)
Testing	18	10	1250
Validation	18	10	1064
Total	182	100	10505

*Including augmented samples

indicating if a participant is depressed or not. Depression in this dataset was assessed using a self-report depression test based on the Patient Health Questionnaire (PHQ). The PHQ is constituted of eight questions (PHQ-8) [10] giving a binary depression label (PHQ-8 binary) and 24 depression severity levels (PHQ-8 scores). The distribution of the number of participants across the depression labels sets (PhQ binary and PhQ-8 scores) for the DAIC-WOZ dataset is shown in Figure 5.

For all experiments, the dataset is randomly divided into 80% for training, 10% for validation, and 10% for testing. In order to keep a balanced ratio between the labels, depressed and non-depressed subjects are first separated before performing the random split. Different subjects were used for training, testing and validation (i.e. among 182 subjects, 146 were considered for training, 18 for validation, and the remaining 18 for testing). Further details are given in Table 4 regarding the number of subjects and augmented data samples in each partition.

5.2. Evaluation Metrics

The given approach is assessed using Accuracy, Precision, Recall, Root Mean Square Error (RMSE), Pearson Correlation Coefficient (CC) and Concordance Coefficient Correlation (CCC) [89]. CCC is usually used to evaluate inter-rater reliability and it measures the agreement between the predicted and true depression scores. Let Tr and Pr be the true and predicted depression score vectors, then $CC(Tr, Pr)$ and $CCC(Tr, Pr)$ are given by:

$$CC(Tr, Pr) = \frac{\sum_{i=1}^{n}(Pr_i - \mu_{Pr})(Tr_i - \mu_{Tr})}{\sum_{i=1}^{n}(Pr_i - \mu_{Pr})^2 \sum_{i=1}^{n}(Tr_i - \mu_{Tr})^2} \tag{8}$$

$$CCC(Tr, Pr) = \frac{2\sigma_{Tr}\sigma_{Pr}CC}{\sigma_{Tr}^2 + \sigma_{Pr}^2 + (\mu_{Tr} - \mu_{Pr})^2} \tag{9}$$

σ_{Tr} and σ_{Pr} represent the standard deviations of variables Tr and Pr, and μ_{Tr} and μ_{Pr} represent their respective means.

5.3. Network Implementation Details

In the following, we present the proposed networks implementation details.

- *Dense layers*: The dense layers' sizes are set to 15 and 10, respectively. In all experiments, a hyperbolic tangent activation function (tanh) is used for both dense layers.

- *Output layer*: For predicting the PHQ-8 binary score, the output layer is a dense layer of size 2 with a sigmoid activation function. To predict the 24 severity levels (PHQ-8 scores), the output layer's size is changed to 24 neurons.

- *Learning rate*: To prevent the models from training instability or failure caused by a large or tiny learning rate value, an adaptive learning rate is used. It is updated each epoch and decreases from the initial value to a minimum of 10^{-10} according to the estimated error. The LSTM models are trained using an initial learning rate of 10^{-5} and a decay of 10^{-6}. While, for convolutional models, an initial learning rate of 10^{-6} and a decay of 10^{-6} are used.

- *Batch size*: The batch size is set to 120 samples for all experiments.

- *Epochs*: A total number of 300 epochs is set for training where an early stopping is performed when the loss function stops improving after 10 epochs.

- *Optimizer:* The proposed models are trained with the RMSProp optimizer and Root Mean Square Error as a loss function.

CNN architectures – For both MFCC-based and audiovisual based CNN architectures, filters of sizes 128 and 64 are used for the convolutional layers. ReLU is used as an activation function for both convolutional layers. The filter and stride are set to 3 and 1, respectively. The dropout fraction value is set to 0.02%.

LSTM architectures – In all LSTM architectures, the following number of output cell units was used:

- *MFCC LSTM*: 60, 40, and 20 output cell units;

- *MFCC-AU LSTM*: 60, 40, and 20 output cell units;

- *MFCC-Word2Vec LSTM*: 60, 40, and 20 output cell units;

- *Model-level fusion LSTM*: 20 output cell units.

Batch normalization and dropout layers are assigned to every LSTM layer with fraction values of 0.02%. Furthermore, in the model-based fusion architecture, a dropout layer is added after the flatten layer with a fraction value of 0.02%. The following parameters are used for all LSTM layers:

- *Activation function:* all LSTM layers are activated with the hyperbolic tangent activation function (tanh).

- *Recurrent step:* the first two LSTM layers, recurrent step is activated with the hard sigmoid function, while in the remaining LSTM layers, recurrent step is activated with the sigmoid function only.

- *Recurrent dropout:* in order to prevent the recurrent state from overfitting, a recurrent dropout of 0.2% is applied.

- *Kernel weights:* in all LSTM layers, kernel weights are initialized using the *glorot_uniform* initializer.

- *Optimization:* penalties are applied over the bias vector using a regularization function to improve performances. The $l1$ and $l2$ imposed constraints are set to 0.0001 and 0.0001, respectively.

5.4. Experiments

In this section, the different performed experiments for depression recognition and assessment using speech, visual and text signals are presented. First, the performance of MFCC-based unimodal deep learning frameworks for PhQ-binary depression detection is evaluated (section 5.4.1). Then, the performances of early fusion (section 5.4.2) and model-level fusion (5.4.3) based frameworks for fusing MFCC features with visual and textual modalities are investigated. Additionally, further analysis is performed to depict the performances of the proposed networks for the detection of depression and non-depression classes (section 5.4.4). Moreover, the performance of the proposed architectures are compared for the prediction of depression severity levels (section 5.4.5). Furthermore, comparison with state-of-the-art benchmark approaches in automatic depression assessment is performed (section 5.4.7). Finally, the best performing architecture among the proposed ones is further evaluated with a Leave-One-Subject-Out strategy (section 5.4.6).

In the following, details concerning the performed experiments are given.

5.4.1. MFCC-based Deep Learning for Audio Depression Detection

Table 5 summarizes the resulting performances of the proposed MFCC-based networks on the testing set for PHQ-binary detection. As the table shows, comparing MFCC-based CNN and LSTM architectures, the LSTM architecture performs slightly better than the CNN one with an accuracy of 66.25% vs. 65.60%. The slight augmentation of accuracy might be an indicator that RNN models are better suited for audio-based depression detection than CNN due to their ability to model the temporal dynamics of audio signals. Thus, the fusion of MFCC high level descriptors with other modalities needs to be investigated.

5.4.2. Early Fusion Strategy for Multimodal Depression Detection

The performance of early fusion strategy of MFCC high level descriptors with visual AUs and textual Word2Vec features for binary depression detection are also presented in Table 5. The fusion of both MFCC unimodal representations, learned with CNN and LSTM models, with the other modalities is compared.

Using the early fusion strategy, fusing CNN and LSTM based MFCC high level descriptors with Word2Vec textual features slightly degrades the results compared to unimodal audio deep models.

On the other hand, fusion with AUs, achieves better results with respect to using single audio modality in both, the CNN and LSTM architectures. An accuracy increment of 2.72% is obtained with the MFCC-AU CNN with respect to the MFCC CNN. Similarly, an accuracy increment of 5.08% is obtained with the MFCC-AU LSTM network compared to the MFCC LSTM network. Moreover, the audiovisual fusion LSTM architecture achieves better performance than the CNN one (accuracy of 71.33% vs. 68.32%). The AUC scores and other performance metrics presented also in Table 5 confirm that the best performing deep neural network is the MFCC-AU LTSM.

This confirms that audiovisual fusion models perform better than audio-based models for binary depression detection, regardless of the architecture type (RNN vs. CNN).

Table 5
Performance of proposed deep learning models for binary depression recognition in terms of AUC score, Accuracy, RMSE, CC, and CCC on the testing set.

Fusion	Features	Deep Audio Features	AUC Score	Acc. (%)	RMSE	CC	CCC
	MFCC	CNN	0.4866	65.60	0.49	0.149	0.06
	MFCC	LSTM	0.4816	66.25	0.49	0.154	0.07
Early	MFCC-Word2Vec	CNN	0.4740	64.29	0.49	0.12	0.06
	MFCC-Word2Vec	LSTM	0.4678	65.98	0.48	0.13	0.07
	MFCC-AU	CNN	0.5059	68.32	0.47	0.22	0.09
	MFCC-AU	LSTM	0.5391	71.33	0.46	0.35	0.17
Model	MFCC-Word2Vec	LSTM	0.4690	68.79	0.46	0.20	0.12
	MFCC-AU	LSTM	**0.6575**	**77.16**	**0.42**	**0.54**	**0.34**

Table 6
Performance of the proposed networks for binary depression recognition in terms of Precision, Recall, and F1-Score. ND: Non-Depressed, D: Depressed

Fusion	Features	Deep Audio Features	Precision		Recall		F1-Score	
			D	ND	D	ND	D	ND
	MFCC	CNN	21	75	16	81	18	78
	MFCC	LSTM	20	75	13	83	16	79
Early	MFCC-AU	CNN	25	76	16	85	20	80
	MFCC-AU	LSTM	34	78	20	88	25	82
	MFCC-Word2Vec	CNN	19	75	15	80	17	77
	MFCC-Word2Vec	LSTM	16	75	10	84	12	79
Model	MFCC-Word2Vec	LSTM	12	75	05	**89**	07	81
	MFCC-AU	LSTM	**53**	**83**	**44**	88	**48**	**85**

5.4.3. *Model-level Fusion Strategy for Multimodal Depression Detection*

The model-level fusion of MFCC-LSTM descriptors with textual and visual modalities are summarized in Table 5. Model-level fusion of audiovisual signals achieved a higher accuracy and a lower RMSE compared to all other models. In this high performing model, two LSTMs are applied: (1) in the first one, MFCC features are fed to LSTM to extract high level representation and more deeper features from the audio signal. (2) The deep audio and the visual features are fused through a second LSTM to learn a deep joint representation of the audiovisual signal. This model-level fusion of audio and visual features boosts the performance for PHQ-8 binary where an accuracy of 77.16% is achieved. The fusion model accuracy is improved by 11.56%, 10.91%, 8.84%, and 5.83%, with respect to MFCC CNN, MFCC LSTM, MFCC-AU CNN, and MFCC-AU LSTM respectively.

Concerning fusion with textual features, using the same model-level fusion architecture, fusing MFCC with word embedding features degrades the performance by 8.37% with respect to fusion with Action Units. On the other hand, model-level fusion with textual features improves the performance with respect to early fusion of audio features with the same modality.

5.4.4. *Performance Analysis for Binary Depression Classes Detection*

To better analyze the performance of the proposed architectures for detecting PHQ-8 binary classes, Table 6 summarizes the resulting performances of all experiments in terms of Precision, Recall and F1-Score, on the testing set, for both Depression and Non-Depression binary classes.

In model-level fusion, an increment is observed in Precision, Recall and F1-Score for both binary depression classes of Depressed and Non-Depressed as compared to all other methods. This can be explained by the fact that further high level features are learned using model-level fusion. In addition, adding an LSTM layer which combines the MFCC-based deep features with AU features allows the network to learn temporal dynamics of the audiovisual signal, resulting in better modeling of depression.

An F1-score of 85% is obtained for the Non-Depressed class, corresponding to 3% and 6% increment compared to MFCC-AU LSTM and MFCC LSTM models, respectively. Similarly, an increase in F1-score of 5% and 7% were noted for the Non-Depressed class compared to MFCC-AU CNN and MFCC CNN experiments, respectively. While, for the Depressed class an increment of 23%, 28%, 32% and 30% in F1-score has been observed as compared to the other four models. Furthermore, all LSTM based models performed better as compared to CNN based models. This could be due to temporal dynamic based learning of LSTM layers.

Table 7
Performance of the proposed deep learning models for PHQ-scores severity levels of depression prediction in terms of RMSE on the testing set. (^{N}): Normalized RMSE

Fusion	Features	Deep Audio Features	RMSE
	MFCC	CNN	0.2041^{N}/4.69
	MFCC	LSTM	0.2093^{N}/4.81
Early	MFCC-Word2Vec	CNN	0.2175^{N}/5.00
	MFCC-Word2Vec	LSTM	0.2109^{N}/4.85
	MFCC-AU	CNN	0.1978^{N}/4.55
	MFCC-AU	LSTM	0.1862^{N}/4.28
Model	MFCC-Word2Vec	LSTM	0.1945^{N}/4.47
	MFCC-AU	LSTM	**0.1519^{N}/ 3.49**

On the other hand, Action Units play a notable role in the performances of the models. For instance, MFCC-AU LSTM and MFCC-AU CNN models achieved 9% and 2% higher F1-score for the Depressed class compared to MFCC LSTM and MFCC CNN, respectively.

Figure 6 shows the confusion matrices for all the proposed architecture in terms of PHQ-binary classes detection. From the figure 6, one can notice that the concatenation of AU with audio features considerably improves the performance of CNN and LSTM based deep learning networks. For Non-Depression detection, an increment of 3.47% and 4.46% is achieved in both audiovisual based CNN and LSTM networks, respectively. Similarly, for Depression class detection, an increment of 0.39% and 7.03% is observed for audiovisual based CNN and LSTM networks, respectively. These results show that fusion of AU features in LSTM network considerably improves its performance for depression detection compared to CNN-based deep learning network. Furthermore, in audiovisual model based fusion, the addition of LSTM layer after concatenation plays an important role for the detection of depression. 23.44% and 30.47% depressed samples are predicted more accurately as compared to early fusion of audiovisual modalities with LSTM network and MFCC-based LSTM network, respectively.

5.4.5. *Evaluation of the Depression Severity Levels Prediction*

Table 7 summarizes the resulting performances of the proposed frameworks on the testing set for the PHQ-score prediction task. Similar to the binary depression detection task, the audiovisual model-level fusion approach outperforms all the other approaches with a normalized RMSE value of 0.15. The model that performed the worst is the one implementing early fusion of CNN-based audio descriptors with Word2Vec features (MFCC-Word2Vec CNN) which obtained the highest normalized RMSE value of 0.2175.

5.4.6. *Leave-One-Subject-Out Experiment for MFCC-AU LSTM*

As mentioned in section 5.1, all of the above experiments were performed using a hold-out strategy with an 80/10/10 percentage split. Leave-One-Subject-Out (LOSO) is generally a better strategy for measuring generalization performance of Machine Learning models. Thus, Leave-One-Subject-Out (LOSO) evaluation has been performed for the best performing architecture (MFCC-AU LSTM). Results of this experiment are presented in Table 8 for binary PHQ detection and PHQ-score prediction. As shown in the table 8, an overall accuracy of 95.38% for the binary depression assessment task was obtained by the MFCC-AU LSTM model. On the other hand, the obtained AUC score, RMSE, and CCC values are 0.94 , 0.22, and 0.89 respectively.

Moreover, the corresponding confusion matrix is given in Figure 7 for PHQ binary recognition task. For LOSO experiment, the proposed model achieved a higher precision for both depressed and non-depressed classes as compared to the 80/10/10 hold-out strategy. Furthermore, regarding PHQ-score prediction, a decrease in RMSE value has been observed with the LOSO strategy (0.15). Overall, the high performance of MFCC-AU LSTM architecture measured using a LOSO strategy for both PHQ assessment tasks (binary and severity level) confirms the generalization power of the proposed architecture.

5.4.7. *Comparison with State-of-the-Art Methods*

In the following, a comparison of performance with state-of-the art benchmark approaches in clinical depression assessment from speech are presented for both PHQ-binary and PHQ-Score prediction tasks.

PHQ-binary – Table 9 compares the performance of the given deep audiovisual model-level fusion approach with existing state-of-the-art methods for PHQ-8 binary (depressed and non-depressed) classes in terms of Precision, Recall and F1-score. The table also presents the average F1-score, Accuracy, RMSE and CC when available.

The presented MFCC-AU LSTM model-level fusion obtained an overall higher accuracy as compared to EmoAudioNet [34] and MFCC-based Recurrent Neural Networks proposed in [33].

The authors of DepAudioNet [19] only provided Precision, Recall, and F1-score for depressed and non-depressed classes. Table 9 shows that the Precision of the proposed MFCC-AU LTSM model for non-depressed class is less than that obtained by DepAudioNet [19]. However, the given MFCC-AU LTSM model achieved a higher precision for depressed class as compared to the DepAudioNet approach. Along with that, our model achieved a higher Precision for both depressed and non-depressed classes as compared to EmoAudioNet and MFCC-based Recurrent Neural Networks proposed in [33, 34].

Among the state-of-the-art approaches presented in Table 9, only the approach of Salekin et al. (2018) [35] was evaluated with a LOSO strategy. With their BLSTM-MIL model, the authors reported an accuracy of 96.7% and an F1-Score of 85.44%. For our LOSO experiment, the accuracy is slightly lower as compared to BLSTM-MIL model, yet a notable increment has been found in average F1-Score.

Figure 6: Confusion matrices of the different proposed deep neural networks architectures for binary depression recognition task. ND: Non-Depressed, D: Depressed.

(a) MFCC CNN

		Actual		
		ND	D	total
Predicted	ND	657 (61.75)	215 (20.21)	75.34 24.66
	D	151 (14.19)	41 (3.85)	78.65 21.35
	total	81.31 18.69	83.98 16.02	

(b) MFCC LSTM

		Actual		
		ND	D	total
Predicted	ND	671 (63.06)	222 (20.86)	75.14 24.86
	D	137 (12.88)	34 (3.20)	80.12 19.88
	total	83.04 16.96	86.72 13.28	

(c) MFCC-AU CNN Early Fusion

		Actual		
		ND	D	total
Predicted	ND	685 (64.38)	214 (20.11)	76.20 23.80
	D	123 (11.56)	42 (3.95)	74.55 25.45
	total	84.78 15.22	83.59 16.41	

(d) MFCC-AU LSTM Early Fusion

		Actual		
		ND	D	total
Predicted	ND	707 (64.45)	204 (19.17)	77.61 22.39
	D	101 (9.49)	52 (4.89)	66.01 33.99
	total	87.50 12.50	79.69 20.31	

(e) MFCC-Word2Vec CNN Early Fusion

		Actual		
		ND	D	total
Predicted	ND	646 (60.71)	218 (20.49)	74.77 25.23
	D	162 (15.23)	38 (3.57)	81.00 19.00
	total	79.95 20.05	85.16 14.84	

(f) MFCC-Word2Vec LSTM Early Fusion

		Actual		
		ND	D	total
Predicted	ND	677 (63.63)	231 (21.71)	74.56 25.44
	D	131 (12.31)	25 (2.35)	83.97 16.03
	total	83.79 16.21	90.23 9.77	

(g) MFCC-Word2Vec LSTM Model-Level Fusion

		Actual		
		ND	D	total
Predicted	ND	720 (67.67)	244 (22.93)	74.69 25.31
	D	88 (8.27)	12 (1.13)	88.00 12.00
	total	89.11 10.89	95.31 4.69	

(h) MFCC-AU LSTM Model-Level Fusion

		Actual		
		ND	D	total
Predicted	ND	709 (66.64)	144 (13.53)	83.12 16.88
	D	99 (9.30)	112 (10.53)	46.92 53.08
	total	87.75 12.25	56.25 43.75	

Table 8
Leave-One-Subject-Out (LOSO) experiment for the best performing architecture (MFCC-AU LSTM) for both depression assessment tasks (binary and severity level) in terms of Accuracy, RMSE, CC, and CCC. (^{Norm}): Normalized RMSE

Assessment task	Metric	Value
Binary	**AUC Score**	0.94
	Acc. (%)	95.38
	RMSE	0.22
	CC	0.94
	CCC	0.89
Severity Level	**RMSE**	**0.15**^{Norm} / 3.40

Figure 7: Confusion matrix of the best performing architecture (MFCC-AU LSTM) using Leave-One-Subject-Out evaluation strategy. ND: Non-Depressed, D: Depressed.

(a) MFCC AU

		Actual		
		ND	**D**	**total**
Predicted	**ND**	7190 (68.44)	357 (3.40)	95.27 4.73
	D	128 (1.22)	2830 (26.94)	95.67 4.33
	total	98.25 1.75	88.80 11.20	

PHQ-Score – Table 10, presents state-of-the-art comparison results for predicting the depression severity levels in terms of RMSE. The proposed audiovisual model-level fusion approach performs better (0.15) than the MFCC-based Recurrent Neural Network architecture [33] (0.17), EmoAudioNet (fusion of MFCC and Spectorgram based CNN Networks) [34] (0.18), [23] (1.46 for depressed male) and other state-of-the-art approaches [30, 32] for the PHQ-8 scores prediction.

5.4.8. Computational Complexity

The proposed deep neural networks architectures were implemented on an Intel Xeon E-2124G @ 3.40GHz Processor (with 32.GB memory and an Nvidia Quadro P4000 graphics card). The computational complexity for the best performing deep neural network (MFCC-AU LSTM) is evaluated on the validation set consisting of 1064 samples. For binary depression detection, the average computation time for prediction of one sample is 5.26 milliseconds. On the other hand, for depression severity levels prediction, the average computation time for prediction of one sample is 5.37 milliseconds. We can conclude that our proposed approach is high performing with a reasonable computation time, and thus satisfies the computational requirements of clinical and real-world applications.

Table 9
Comparison of proposed network with state of the art methods for PHQ-8 binary in terms of Precision, Recall and F-Score (%) for Depressed (D) and Non-Depressed (ND) classes. The table also summrizes the average F-score, Accuracy, RMSE and CC. The best performances are highlighted in bold. (*Evaluation with Leave-One-Subject-Out strategy.)

Method	Precision		Recall		F1-Score			Acc.	RMSE	CC
	D	ND	D	ND	D	ND	Av.			
DepAudioNet [19]	35	**100**	**100**	54	52	70	—	—	—	—
EmoAudioNet [34]	52	80	46	84	49	82	72.89	73.25	0.47	—-
MFCC-based RNN [33]	**69**	78	35	**94**	46	**85**	73.65	76.27	**0.41**	—
MFCC-AU LSTM	53	83	44	88	48	**85**	76.09	77.16	0.42	**0.54**
BLSTM-MIL* [35]	—	—	—	—	—	—	85.44	**96.7**	—	—
MFCC-AU LSTM*	96	95	89	98	92	97	**95.48**	95.38	0.22	0.94

Table 10
Performance comparison of our best performing proposed deep neural network and state-of-the-art depression severity levels prediction methods on the DAIC-Woz dataset. (^{DM}) : Depressed Male, which refers the depression value of males only. (^{Norm}): Normalized RMSE.

Method	RMSE
Valstar et al. (2016) [32]	7.78
Yang et al. (2017)[30]	5.59^{DM}
Yang et al. (2017b)[23]	1.46^{DM}
Othmani et al. (2020)[34]	0.18^{*}
Rejaibi et al. (2019)[33]	0.17^{*}
MFCC-AU LSTM	0.15^{Norm} / 3.49

6. Discussion

In this research, a comparative analysis of different Deep Neural Networks is conducted for multimodal depression recognition. The obtained results show that Recurrent Neural Networks are better suited than Convolutional Neural Networks for clinical depression detection due to their temporal dynamics characteristics [90]. Furthermore, the fusion of deep audio features with visual features leads to better performance comparing to their fusion with word embedding textual features. Surprisingly, CNN based audio features alone provided better results as compared to their fusion with textual features. This degradation in performance does not necessarily indicate that fusing audio and textual features is not suitable for depression assessment. It is possible that the way textual information is encoded in this work with Word2Vec embedding is not the optimal way for encoding the patient's and interviewer's responses. More suitable representations of textual features should be investigated.

According to our experiments, the best outperforming model is based on the model-level fusion using an LSTM network of the deep audio and the visual features. This confirms that the facial information encodes relevant patterns about depression. In this work, the only considered visual features are Actions Units, because the raw videos were not made available in the DAIC-WOZ for ethical reasons. Research and development of clinical real-world applications for depression diagnosis is plagued by ethical and privacy issues. In fact, the use of several modalities, as shown in this work, improves the prediction performance. However, strict limitations are present because of: (1) the lack and the small size of the available datasets. (2) the limited available data/modalities and the absence of interesting data/modalities like the facial images, raw images of silhouette and other clinical and sociological factors that could be relevant for depression diagnosis. Artificial Intelligence can be used as a tool to improve data privacy without being a threat and by making profit of the potential of many modalities [91]. Thus, efforts and innovation from scientific communities is needed.

7. Conclusion and Future Work

This paper presents experiments aimed at showing robust multimodal features and the best strategy to fuse them for depression detection and assessment. Two unimodal representations based on CNNs and RNNs allowing to learn high level audio features from MFCC features are proposed and compared. Temporal dynamics representations of multimodal data are learned with Short Long-Term Memory (LSTM) Recurrent Neural Networks. Moreover, an extensive study is performed to investigate the best suited multimodal fusion approach of MFCC-based deep audio features with other modalities for clinical depression recognition. Early and model-level fusion strategies of MFCC-based deep audio features with word embedding textual and Action Units visual features are evaluated on the DAIC-WOZ corpus. Model-level fusion of audiovisual features improves significantly the results and a notable increment in performance is observed. Further, a comparison with state of the art benchmark approaches is performed. Our model attains state-of-the-art performance on binary depression detection, and outperforms all existing approaches in the recognition of depression severity level.

In future work, we aim to apply the proposed best performing deep neural network in a real-world application that can assist clinicians in making more accurate diagnosis and better follow-up and monitoring of patients. Follow-up and assisting systems are needed to prevent the onset of a mental health crisis by seeking the health practitioner's aid at the right time and recommend personalized interventions.

Declaration of Competing Interest

The author(s) declare(s) that there is no conflict of interest.

References

[1] M. Friedrich, Depression is the leading cause of disability around the world, Journal of the American Medical Association 317 (2017) 1517–1517.

[2] J. Lutz, K. Morton, N. A. Turiano, A. Fiske, Health conditions and passive suicidal ideation in the survey of health, ageing, and retirement in europe, Journals of Gerontology Series B: Psychological Sciences and Social Sciences 71 (2016) 936–946.

[3] R. Thomas-MacLean, J. Stoppard, B. B. Miedema, S. Tatemichi, Diagnosing depression: there is no blood test, Canadian Family Physician 51 (2005) 1102–1103.

[4] S. Kapur, A. G. Phillips, T. R. Insel, Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?, Molecular psychiatry 17 (2012) 1174–1179.

[5] S.-I. Validity, The structured clinical interview for dsm-iv axis i disorders (scid-i) and the structured clinical interview for dsm-iv axis ii disorders (scid-ii), Comprehensive handbook of psychological assessment, volume 2: Personality assessment 2 (2004) 134.

[6] M. Hamilton, A rating scale for depression, Journal of neurology, neurosurgery, and psychiatry 23 (1960) 56.

[7] A. S. Zigmond, R. P. Snaith, The hospital anxiety and depression scale, Acta psychiatrica scandinavica 67 (1983) 361–370.

[8] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, et al., The 16-item quick inventory of depressive symptomatology

(qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression, Biological psychiatry 54 (2003) 573–583.

[9] P. Olaya-Contreras, T. Persson, J. Styf, Comparison between the beck depression inventory and psychiatric evaluation of distress in patients on long-term sick leave due to chronic musculoskeletal pain, Journal of multidisciplinary healthcare 3 (2010) 161.

[10] L. S. Williams, E. J. Brizendine, L. Plue, T. Bakas, W. Tu, H. Hendrie, K. Kroenke, Performance of the phq-9 as a screening tool for depression after stroke, stroke 36 (2005) 635–638.

[11] M. Chmielewski, L. A. Clark, R. M. Bagby, D. Watson, Method matters: Understanding diagnostic reliability in dsm-iv and dsm-5, Journal of Abnormal Psychology 124 (2015) 764.

[12] R. M. Bagby, A. G. Ryder, D. R. Schuller, M. B. Marshall, The hamilton depression rating scale: has the gold standard become a lead weight?, American Journal of Psychiatry 161 (2004) 2163–2177.

[13] S. Gilbody, T. Sheldon, A. House, Screening and case-finding instruments for depression: a meta-analysis, Cmaj 178 (2008) 997–1003.

[14] Y. Ren, H. Yang, C. Browning, S. Thomas, M. Liu, Performance of screening tools in detecting major depressive disorder among patients with coronary heart disease: a systematic review, Medical science monitor: international medical journal of experimental and clinical research 21 (2015) 646.

[15] P. Pichot, Self-report inventories in the study of depression, in: New Results in Depression Research, Springer, 1986, pp. 53–58.

[16] Y. S. Ben-Porath, Assessing personality and psychopathology with self-report inventories, Handbook of psychology (2003) 553–577.

[17] F. Dornaika, B. Raducanu, Inferring facial expressions from videos: Tool and application, Signal Processing: Image Communication 22 (2007) 769–784.

[18] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, F. De la Torre, Detecting depression from facial actions and vocal prosody, in: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, pp. 1–7.

[19] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, Depaudionet: An efficient deep model for audio based depression classification, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, pp. 35–42.

[20] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, D. P. Subha, Automated eeg-based screening of depression using deep convolutional neural network, Computer methods and programs in biomedicine 161 (2018) 103–113.

[21] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, M. Pantic, A survey of multimodal sentiment analysis, Image and Vision Computing 65 (2017) 3–14.

[22] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, T. F. Quatieri, Detecting depression using vocal, facial and semantic communication cues, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 11–18.

[23] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, 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.

[24] P. Lopez-Otero, L. Dacia-Fernandez, C. Garcia-Mateo, A study of acoustic features for depression detection, in: Proceedings of the 2nd International Workshop on Biometrics and Forensics, IEEE, pp. 1–6.

[25] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, H. Kang, Detecting depression using an ensemble logistic regression model based on multiple speech features, Computational and mathematical methods in medicine 2018 (2018).

[26] M. Cutajar, E. Gatt, I. Grech, O. Casha, J. Micallef, Comparative study of automatic speech recognition techniques, IET Signal Processing 7 (2013) 25–46.

[27] N. Cummins, J. Epps, M. Breakspear, R. Goecke, An investigation of depressed speech detection: Features and normalization, in: Proceedings of the 12th Annual Conference of the International Speech Communication Association.

[28] Z. Zhang, W. Lin, M. Liu, M. Mahmoud, Multimodal deep learning framework for mental disorder recognition, in: Proceedings of the 15th IEEE International Conference on Automatic Face & Gesture Recognition.

[29] T. Al Hanai, M. M. Ghassemi, J. R. Glass, Detecting depression with audio/text sequence modeling of interviews., in: Proceedings of INTERSPEECH, volume 2522, pp. 1716–1720.

[30] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, H. Sahli, Multimodal measurement of depression using deep learning models, in: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, pp. 53–59.

[31] P. Ekman, W. Friesen, J. Hager, Facial action coding system: The manual, Salt Lake City, Research Nexus (2002).

[32] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, 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.

[33] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, A. Othmani, Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech, arXiv preprint arXiv:1909.07208 (2019).

[34] A. Othmani, D. Kadoch, K. Bentounes, E. Rejaibi, R. Alfred, A. Hadid, Towards robust deep neural networks for affect and depression recognition from speech, In CAIHA ICPR workshop (2020).

[35] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, J. A. Stankovic, A weakly supervised learning framework for detecting social anxiety and depression, Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 2 (2018) 81.

[36] W. C. de Melo, E. Granger, M. B. Lopez, Encoding temporal information for automatic depression recognition from facial analysis, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp. 1080–1084.

[37] S. Song, S. Jaiswal, L. Shen, M. Valstar, Spectral representation of behaviour primitives for depression analysis, IEEE Transactions on Affective Computing (2020).

[38] M. Blais, L. Baer, Understanding rating scales and assessment instruments, in: Handbook of clinical rating scales and assessment in psychiatry and mental health, Springer, 2009, pp. 1–6.

[39] C. Karmen, R. C. Hsiung, T. Wetter, Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods, Computer methods and programs in biomedicine 120 (2015) 27–36.

[40] S. Ware, C. Yue, R. Morillo, J. Lu, C. Shang, J. Bi, J. Kamath, A. Russell, A. Bamis, B. Wang, Predicting depressive symptoms using smartphone data, Smart Health 15 (2020) 100093.

[41] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, M. Tsiknakis, Automatic assessment of depression based on visual cues: A systematic review, IEEE Transactions on Affective Computing 10 (2017) 445–470.

[42] G. Stratou, S. Scherer, J. Gratch, L.-P. Morency, Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender, Journal on Multimodal User Interfaces 9 (2015) 17–29.

[43] S. Dham, A. Sharma, A. Dhall, Depression scale recognition from audio, visual and text analysis, arXiv preprint arXiv:1709.05865 (2017).

[44] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, M. Breakspear, Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors, IEEE Transactions on Affective Computing 9 (2016) 478–490.

[45] D. Zhou, J. Luo, V. M. Silenzio, Y. Zhou, J. Hu, G. Currier, H. Kautz, Tackling mental health by integrating unobtrusive multimodal sensing, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence.

[46] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, M. Breakspear, Cross-cultural detection of depression from nonverbal behaviour, in: Proceedings of the 11th IEEE International Conference

and Workshops on Automatic Face and Gesture Recognition, volume 1, IEEE, pp. 1–8.

[47] K. Ooi, Early prediction of clinical depression in adolescents using single-channel and multi-channel classification approach (2014).

[48] A. Jan, H. Meng, Y. F. B. A. Gaus, F. Zhang, Artificial intelligent system for automatic depression level analysis through visual and vocal expressions, IEEE Transactions on Cognitive and Developmental Systems 10 (2017) 668–680.

[49] H. Dibeklioğlu, Z. Hammal, Y. Yang, J. F. Cohn, Multimodal detection of depression in clinical interviews, in: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 307–310.

[50] L. He, D. Jiang, H. Sahli, Multimodal depression recognition with dynamic visual and audio cues, in: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pp. 260–266.

[51] J. Joshi, A. Dhall, R. Goecke, J. F. Cohn, Relative body parts movement for automatic depression analysis, in: Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, pp. 492–497.

[52] W. C. de Melo, E. Granger, A. Hadid, Combining global and local convolutional 3d networks for detecting depression from facial expressions, in: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, pp. 1–8.

[53] W. C. De Melo, E. Granger, A. Hadid, Depression detection based on deep distribution learning, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, pp. 4544–4548.

[54] Y. Kang, X. Jiang, Y. Yin, Y. Shang, X. Zhou, Deep transformation learning for depression diagnosis from facial images, in: Proceedings of the Chinese Conference on Biometric Recognition, Springer, pp. 13–22.

[55] M. Al Jazaery, G. Guo, Video-based depression level analysis by encoding deep spatiotemporal features, IEEE Transactions on Affective Computing (2018).

[56] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition (2015).

[57] X. Zhou, P. Huang, H. Liu, S. Niu, Learning content-adaptive feature pooling for facial depression recognition in videos, Electronics Letters 55 (2019) 648–650.

[58] X. Zhou, K. Jin, Y. Shang, G. Guo, Visually interpretable representation learning for depression recognition from facial images, IEEE Transactions on Affective Computing (2018).

[59] Z. Liu, B. Hu, L. Yan, T. Wang, F. Liu, X. Li, H. Kang, Detection of depression in speech, in: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, IEEE, pp. 743–747.

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

[61] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, T. Gedeon, M. Breakspear, G. Parker, A comparative study of different classifiers for detecting depression from spontaneous speech, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp. 8022–8026.

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

[63] M. Asgari, I. Shafran, L. B. Sheeber, Inferring clinical depression from speech and spoken utterances, in: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, IEEE, pp. 1–5.

[64] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, N. B. Allen, Detection of clinical depression in adolescents' speech during family interactions, IEEE Transactions on Biomedical Engineering 58 (2010) 574–586.

[65] N. Cummins, B. Vlasenko, H. Sagha, B. Schuller, Enhancing speech-based depression detection through gender dependent vowel-level formant features, in: Proceedings of the Conference on Artificial Intelligence in Medicine in Europe, Springer, pp. 209–214.

[66] S. Song, L. Shen, M. Valstar, Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features, in: Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, pp. 158–165.

[67] P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, Assessing speaker independence on a speech-based depression level estimation system, Pattern Recognition Letters 68 (2015) 343–350.

[68] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp. 5200–5204.

[69] H. Dinkel, P. Zhang, M. Wu, K. Yu, Depa: Self-supervised audio embedding for depression detection, arXiv preprint arXiv:1910.13028 (2019).

[70] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, A. Othmani, Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis, Machine Learning with Applications 2 (2020) 100005.

[71] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, IEEE Transactions on Circuits and Systems for Video Technology 28 (2017) 3030–3043.

[72] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017) 98–125.

[73] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia systems 16 (2010) 345–379.

[74] A. Samareh, Y. Jin, Z. Wang, X. Chang, S. Huang, Predicting depression severity by multi-modal feature engineering and fusion, arXiv preprint arXiv:1711.11155 (2017).

[75] J. Chen, Z. Chen, Z. Chi, H. Fu, Emotion recognition in the wild with feature fusion and multiple kernel learning, Proceedings of the 16th International Conference on Multimodal Interaction, pp. 508–513.

[76] T. Baltrušaitis, A. Chaitanya, M. Louis-Philippe, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[77] P. Tzarakis, G. Trigeorgis, M. Nicolaou, S. Björn, Z. Stefanos, End-to-end multimodal emotion recognition using deep neural networks, IEEE Journal of Selected Topics in Signal Processing 11 (2017) 1301–1309.

[78] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, S. Han, P. Liu, M. Chen, Y. Tong, Feature-level and model-level audiovisual fusion for emotion recognition in the wild, Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, pp. 443–448.

[79] P. V. Rouast, A. Marc, C. Raymond, Deep learning for human affect recognition: Insights and new developments, IEEE Transactions on Affective Computing. (2019).

[80] D. H. Kim, W. J. Baddar, J. Jang, Y. M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition, IEEE Transactions on Affective Computing 10 (2017) 223–236.

[81] F. Ringeval, F. Eyben, E. Kroubi, J. P. Thiran, T. Ebrahimi, D. Lalanne, B. Schuller, Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data, Pattern Recognition Letters 66 (2015) 22–30.

[82] W. C. de Melo, E. Granger, A. Hadid, A deep multiscale spatiotemporal network for assessing depression from facial dynamics, IEEE Transactions on Affective Computing (2020).

[83] J. Gratch, R. Artstein, G. M. 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., in: LREC, pp. 3123–3128.

[84] J. Perkins, Python 3 text processing with NLTK 3 cookbook, Packt Publishing Ltd, 2014.

[85] P. V. Janse, S. Magre, P. Kurzekar, R. Deshmukh, A comparative

study between mfcc and dwt feature extraction technique, International Journal of Engineering Research and Technology 3 (2014) 3124–3127.

[86] P. Lewinski, T. M. Den Uyl, C. Butler, Observer-based measurement of facial expression with the facial action coding system, The handbook of emotion elicitation and assessment 3 (2007) 203–221.

[87] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, IEEE, pp. 1–10.

[88] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the International Conference on Language Resources and Evaluation.

[89] I. Lawrence, K. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics (1989) 255–268.

[90] G. Kapidis, R. W. Poppe, E. A. van Dam, R. C. Veltkamp, L. P. Noldus, Where am i? comparing cnn and lstm for location classification in egocentric videos, in: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), IEEE, pp. 878–883.

[91] R. Kusters, D. Misevic, H. Berry, A. Cully, Y. Le Cunff, L. Dandoy, N. Díaz-Rodríguez, M. Ficher, J. Grizou, A. Othmani, T. Palpanas, M. Komorowski, P. Loiseau, C. Moulin Frier, S. Nanini, D. Quercia, M. Sebag, F. Soulié Fogelman, S. Taleb, L. Tupikina, V. Sahu, J.-J. Vie, F. Wehbi, Interdisciplinary research in artificial intelligence: Challenges and opportunities, Front. Big Data (2020) 577974.