Interspeech 2024
1-5 September 2024, Kos, Greece

A Cross-Attention Layer coupled with Multimodal Fusion Methods for Recognizing Depression from Spontaneous Speech

Loukas Ilias[1], *Dimitris Askounis*[1]

[1]DSS Laboratory, School of ECE, National Technical University of Athens, Greece
lilias@epu.ntua.gr, askous@epu.ntua.gr

Abstract

Depression is a serious mood disorder, which affects the way people feel and perform daily activities. Speech is a reliable biomarker for diagnosing depression, since depressed people present decreased verbal activity productivity and "lifeless" sounding speech. Existing methods employ unimodal models, use early, intermediate, or late fusion strategies to fuse the different modalities, rely on feature extraction, and perform their approaches only in the English language. This study presents a new method for identifying depression from spontaneous speech in the Italian language, which uses a cross-attention layer for capturing the cross-modal interactions, followed by a variety of multimodal fusion methods. We also perform a multi-task learning framework to explore whether the prediction of age, education level, and gender help in recognizing depression. Findings show that our approach yields multiple advantages over existing approaches reaching Accuracy up to 95.29%.

Index Terms: depression, spontaneous speech, cross-modal interaction, fusion methods, spectrogram, multitask learning

1. Introduction

The World Health Organization (WHO) states that approximately 280 million people suffer from depression globally [1]. Depression is a mental disorder, which is characterized by loss of interest and pleasure for a long period of time. If depression is not diagnosed early for allowing the person to receive the suitable treatment, it is a leading factor of suicide. Thus, early detection of depression appears to be imperative. Research has shown that speech constitutes a reliable biomarker for detecting depression [2]. Specifically, people with depression present anomalities in speech, including lower speech rates, less pitch variability and more self-referential speech. Depression affects language also [3]. For instance, depressed people use first-person singular pronouns, negative thinking, and self-focus. Therefore, employing both speech and transcripts in a multimodal setting is a hot research topic nowadays.

Existing research works rely on the extraction of handcrafted features and the train of traditional machine learning classifiers or deep learning approaches [4, 5, 6]. However, extracting features is a timely procedure requiring expertise on the specific topic. Additionally, the majority of research studies uses unimodal approaches for predicting depression using mainly speech [7]. Although there are studies employing multimodal models, these studies employ early [8, 9], intermediate [10, 11], or late fusion [12, 13] strategies. In the early fusion strategy, representation vectors of the modalities are concatenated at the input level, while in the intermediate fusion, the representation vectors are concatenated during training, thus equal importance is assigned to the modalities. In the late fusion strategy, unimodal models are trained independently and decision voting is applied, i.e., majority voting. The inter-modal interactions cannot be captured through these approaches. In addition, the majority of research works have tested their approaches only in English language, thus the acoustic and phonetic content of data might differ in other languages. Finally, to the best of our knowledge, no study has experimented with predicting depression, age, education level, and gender at the same time.

To tackle these limitations, we present a new method for detecting depression from spontaneous speech in the Italian language. Specifically, we feed each transcript into a pretrained Italian BERT model. Each speech signal is transformed into an image of three channels, namely log-Mel spectrogram, delta, and delta-delta. Each image is passed through a pretrained AlexNet [14] model. Next, the textual and image representations are passed through a cross-attention scaling layer. Finally, we employ and compare a variety of multimodal fusion methods, including Multimodal Factorized Bilinear Pooling (MFB), Multimodal Factorized High-order pooling (MFH) [15], and more, for fusing the outputs of the cross-attention scaling layer and predicting depression. Additionally, we introduce multitask learning (MTL) architectures to explore if gender, age, and education level as auxiliary tasks help the primary task (depression recognition). Results demonstrate the effectiveness of the proposed approach via an extensive ablation study, as well as multiple advantages over state-of-the-art approaches.

Our main contributions can be summarized as follows:

- We introduce a method which includes a cross-attention layer and multimodal fusion approaches.
- We perform multi-task learning experiments to explore whether the prediction of gender, age, and education level lead towards the increase of depression detection's performance.
- We compare our approaches with competitive baselines, including shallow machine learning classifiers and deep learning.
- We perform an extensive ablation study to verify the effectiveness of the proposed approach.

2. Related Work

Early Fusion. The study in [9] constructs a graph based on question-answering pairs. Specifically, a Graph Attention Network is trained. In terms of the multimodal fusion, the authors employ an early fusion approach. A multitask learning framework is adopted, which predicts the level of depression severity (regression) and classifies the subject as depressive or non-depressive. A similar approach is introduced by [8], where the

10.21437/Interspeech.2024-188

authors employ an early fusion approach and concatenate the representation vectors of audio, visual, and textual modalities. A multi-task learning framework is trained for classifying the level of disorder and predicting the disorder score.

Intermediate Fusion. The study in [16] converts speech signals into spectrogram and uses a VGG16 pretrained model followed by Gated Convolutional Neural Networks and one LSTM layer. The authors pass the BERT embeddings into CNN layers followed by LSTM layers. The representation vectors of the two modalities are concatenated for predicting the Patient Health Questionnaire (PHQ) score. In [4], the authors use articulatory coordination features (ACFs) derived from vocal tract variables. A staircase regression approach is used, where an ensemble of models is trained on multiple partitions of the same training data set. A hierarchical attention network (HAN) is used for extracting textual representation. Additional features representing the prosodic information are extracted. The abovementioned feature representations are concatenated for estimating the depression severity score. In [10], speech signals are represented as log-Mel spectrograms and fed into temporal CNNs, while text is passed through the encoder part of the transformer. Representation vectors of these two modalities are concatenated for predicting whether the individual has depression or not. Ref. [11] adopts a similar approach. DeepSpectrum features are obtained from speech signals and fed into Temporal Convolutional Networks (TCNs) followed by Attention and Dense Layers. The authors feed the word2vec embeddings into a transformer encoder. Finally, the audio and textual vectors are concatenated into a single vector. Ref. [17] proposes a multimodal neural network consisting of two branches of LSTMs for extracting textual and acoustic representations. These two representations are concatenated in one single vector. The authors in [18] concatenate the audio and transcript representations during training. The authors in [19] pass the MFCC features through CNN layers, while the visual and textual modalities are passed through dense layers. The two representations are concatenated into one feature vector.

Late Fusion. The authors in [13] use audio, videos, and transcripts and combine the respective representations via a late fusion approach, namely adaptive nonlinear judge classifier. A majority vote approach is adopted by [12].

Other approaches. In [20], the authors use sentence embeddings, log-Mel spectrograms, and facial expressions and employ ConvBiLSTMs. They fuse the representation vectors by using an attention layer and state that the proposed approach outperforms late fusion strategies. A different approach is proposed by [21], where feed-forward highway layers with gating units are used for controlling the information flow of the different modalities. This approach is compared with early and late fusion strategies. Results suggest that the proposed approach yields the highest results.

3. Dataset

We use the Androids corpus [22] for performing our experiments. Participants are performing two tasks, namely the reading and interview task. In our experiments, we use data from the interview task. Specifically, this task consists of 116 spontaneous speech samples. All experiments are person independent. Audio files are in Italian language. This dataset includes information about the gender, age, and education level of the individuals. The populations of depressed and non-depressed participants have the same distribution in terms of age, gender, and education. Due to the fact that manual transcripts are not provided, we use whisper large-v3 [23], in order to produce automatic transcripts.

4. Proposed Methodology

In this section, we describe our proposed methodology for recognizing depression from spontaneous speech. Fig. 1 illustrates our proposed architecture.

4.1. Single - Task Learning

Text Processing: Since data are in Italian language, we employ Italian BERT[1]. Firstly, each transcript is passed through the Italian BERT tokenizer, where input_ids and attention mask are returned. Transcripts are padded to a maximum length of 512 tokens, while transcripts with number of tokens greater than 512 are truncated. Next, the input_ids and attention mask are fed to the Italian BERT model. Let $f^t \in \mathbb{R}^{1\times d}$, corresponding to the [CLS] token, be the transcript representation, where $d = 768$.

Speech Processing: We use the Python library librosa [24] for converting the speech signals into images consisting of three channels, namely log-Mel spectrogram, delta, and delta-delta. We use 224 Mel bands, hop length equal to 512, and a Hanning window. Each image is resized to (224×224) pixels. We pass each image through a pretrained *AlexNet* [14] model. Let $f^v \in \mathbb{R}^{1\times d}$ be the image representation, where $d = 768$.

Cross-Attention Layer: Motivated by [25], we design a cross-attention layer, which returns a pair of scalars, one for each modality. This pair of scalars allows for scaling the two modalities with respect to each other. One modality is used as a query for the attention of the other.

In terms of the textual modality, let $Q_i = FC_q^t\left(f^v\right)$, $K_t = FC_k^t\left(f^t\right)$, and $V_t = FC_v^t\left(f^t\right)$. The scaling value, denoted as S_t can be calculated as follows: $S_t = sigmoid\left(\frac{Q_i \cdot K_t^T}{\sqrt{d}}\right)$. In terms of the image modality, let $Q_t = FC_q^i\left(f^t\right)$, $K_i = FC_k^i\left(f^v\right)$, and $V_i = FC_v^i\left(f^v\right)$. The scaling value, denoted as S_i can be calculated as follows: $S_i = sigmoid\left(\frac{Q_t \cdot K_i^T}{\sqrt{d}}\right)$. The outputs of the attention mechanism can be calculated as $S_t \times V_t$ and $S_i \times V_i$. Note that $FC_q^t, FC_k^t, FC_v^t, FC_q^i, FC_k^i, FC_v^i \in \mathbb{R}^{d\times d}$.

Similar to [26], we use residual connections followed by layer normalization, as described via the equations below: $\hat{E}_t = LayerNorm\left(S_t \times V_t + f^t\right)$, $\hat{E}_i = LayerNorm\left(S_i \times V_i + f^v\right)$.

Next, we pass $\hat{E}_t$ and $\hat{E}_i$ through two shared fully connected feed-forward networks with a ReLU activation function in between, as follows: $\hat{E}_t{}' = LayerNorm\left(FC_m^n\left(ReLU\left(FC_p^q\left(\hat{E}_t\right)\right)\right)\right)$, $\hat{E}_i{}' = LayerNorm\left(FC_m^n\left(ReLU\left(FC_p^q\left(\hat{E}_i\right)\right)\right)\right)$, where $FC_p^q \in \mathbb{R}^{d\times 4d}$, $FC_m^n \in \mathbb{R}^{4d\times d}$.

Next, we concatenate $\hat{E}_t$ and $\hat{E}_t{}'$ (similarly $\hat{E}_i$ and $\hat{E}_i{}'$) into one single vector, i.e., $\hat{E}_t{}'' = [\hat{E}_t, \hat{E}_t{}']$, $\hat{E}_i{}'' = [\hat{E}_i, \hat{E}_i{}']$, where $\hat{E}_t{}'', \hat{E}_i{}'' \in \mathbb{R}^{2d}$.

Fusion Methods: Next, we employ a variety of fusion methods, which are described in detail below, so as to fuse $\hat{E}_t{}''$ and $\hat{E}_i{}''$ in one single vector:

- Concatenation: We concatenate $\hat{E}_t{}''$ and $\hat{E}_i{}''$ into one single vector, i.e., $z \in \mathbb{R}^{4d}$. We use a dropout layer with a rate of

[1] https://github.com/dbmdz/berts

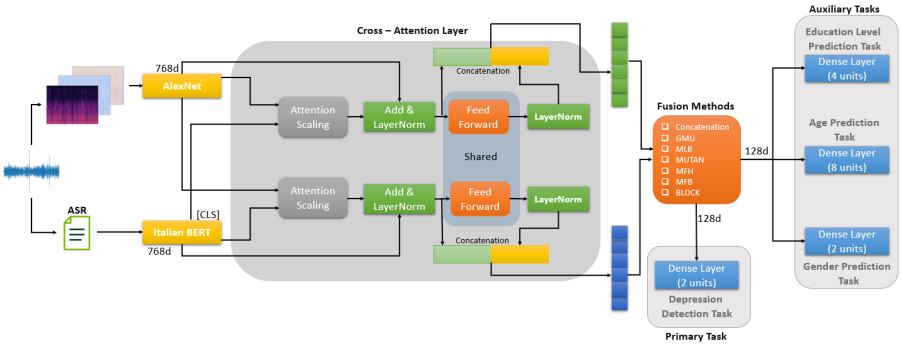

Figure 1: *Our Proposed Methodology*

0.4. We use a dense layer of 128 units.

- Gated Multimodal Unit (GMU): We adopt the method introduced in [27], which controls the information flow of the two modalities towards the final classification. The equations govering the GMU are described as follows: $h^t = \tanh\left(W^t \hat{E_t}'' + b^t\right), h^v = \tanh\left(W^v \hat{E_i}'' + b^v\right), z = \sigma(W^z[\hat{E_t}''; \hat{E_i}''] + b^z), h = z * h^t + (1 - z) * h^v$, where $W^t, W^v, W^z \in \mathbb{R}^{128}$ denote the learnable parameters, and [.;.] the concatenation operation. h is the output of the GMU.
- MUTAN decomposition [28]
- Multimodal Low-rank Bilinear (MLB) pooling [29]
- MFB [15]
- MFH [15]: It is based on cascading two MFB blocks.
- BLOCK [30]: This method is based on the block-term tensor decomposition [31] and combines the strengths of the Candecomp/PARAFAC (CP) [32] and Tucker decompositions.

The output of the aforementioned fusion methods corresponds to a vector with dimensionality accounting for 128.

Output Layer: Finally, we use a dense layer consisting of two units, which gives the final prediction. The cross-entropy loss function is minimized.

4.2. Multi - Task Learning

According to research, gender [33], age[2], and education level [34] are linked with depression. In this section, we design a multi-task learning framework consisting of a primary task, i.e., depression detection (binary classification), and auxiliary tasks, i.e., gender recognition (binary classification), estimation of education level (multiclass classification), and age prediction (multiclass classification). In this approach, we explore if the auxiliary tasks help the primary task in increasing its performance. As illustrated in Fig. 1, in terms of gender recognition, we add a dense layer consisting of two units. In terms of education level recognition, we add a dense layer consisting of four units. Regarding the age prediction, we define the following age groups: [19,25],[26,32],[33,39],[40,46],[47,53],[54,60],[61,67],[68,71]. Thus, we add a dense layer consisting of 8 units.

All the tasks are learnt simultaneously and updated by the following loss function:

$$L = (1 - \alpha - \beta - \gamma) \cdot L_{depression} + \alpha \cdot L_{gender} + \beta \cdot L_{education} + \gamma \cdot L_{age} \quad (1)$$

[2]https://www.nhs.uk/mental-health/conditions/depression-in-adults/causes

, where $L_{depression}, L_{gender}$, $L_{education}$, and L_{age} correspond to the cross-entropy loss function. α, β, γ are hyperparameters denoting the importance we place to each task.

5. Experiments and Results

Baselines. We compare our approaches with the following baselines:

- *Only transcript:* We use a pretrained Italian BERT model and a learning rate of 1e-5.
- *Only Speech signal:* Each speech signal is represented as an image and fed into a pretrained AlexNet model. A learning rate of 1e-5 is employed.
- *BS1* [22]: This approach segments the audio signal into analysis windows of 25ms length and extracts features per window. SVM classifier is trained.
- *BS2* [22]: After calculating the feature sets per analysis windows as above, this approach segments the speech signal into frames of length equal to 128 and passes each frame through an LSTM layer. A majority vote approach is adopted.
- eGeMAPSv02 features (functional): This method trains a SVM classifier. We use the openSMILE Python toolkit [35].
- ComParE_2016 features (functional): This method trains a SVM classifier. We use the openSMILE Python toolkit [35].

Experimental Setup. In [22], the split of the participants into subsets to be used for a 5-fold setup is provided. In our study, we repeat the experiments four times and report the average and standard deviation over four runs. For Italian BERT and AlexNet, the learning rate is set to 1e-5, while for the rest layers, the learning rate is set to 1e-4. We train our models for 40 epochs with a batch size of 4. In terms of the MTL setting, we set $\alpha = \beta = \gamma = 0.1$. We use PyTorch for performing our experiments. All experiments are performed on a single Tesla P100-PCIE-16GB GPU with the running time ranging from 1 hour to 1.5 hours. For significance testing, we use the Almost Stochastic Order (ASO) test [36, 37] as implemented by [38]. Specifically, the ASO test determines whether a stochastic order [39] exists between two models, i.e., A and B. A score (ϵ_{min}) is calculated representing how far the first is from being significantly better than the second. When $\epsilon_{min} = 0$, then A is truly stochastically dominant over B. When $\epsilon_{min} < 0.5$, A is almost stochastically dominant over B. For $\epsilon_{min} = 0.5$, no order can be determined.

Evaluation Metrics. Precision, Recall, F1-score, Accuracy, and Specificity are used to evaluate the performance of the introduced approaches.

Results. Results are reported in Table 1. We observe that the

usage of *BLOCK* as fusion method leads to the best performing model outperforming the rest approaches in Accuracy and F1-score by 1.21-21.99% and 1.32-22.23% respectively. Multimodal models perform better than unimodal ones verifying our initial hypothesis that the usage of multiple modalities improves detection performance. The concatenation mechanism achieves the worst results compared with the other fusion methods, since it assigns equal importance to each individual modality. We believe that MFB outperforms MFH, since the MFH method is developed by cascading two MFB blocks, thus appears to be complex for our limited dataset. We hypothesize that GMU achieves a poor performance, since it controls the information flow without capturing so effectively the cross-modal interactions. We observe that single-task learning settings perform better than multi-task learning ones. This can be justified by the fact that depression is a mental disorder, which can happen to anyone. There are many causes of depression, e.g. stressful events, personality, health issues (cancer), loneliness, etc. According to statistical test, our best performing model is almost stochastically dominant in terms of accuracy over all the approaches, except for *Only speech signal*, where $\epsilon_{min} = 0$. We are not able to perform statistical tests with [22], since the results obtained over individual folds are not available.

Table 1: *Performance comparison among proposed models and baselines. Reported values are mean ± standard deviation. Results are averaged across four runs (5-fold setting).* $(*)$ *means that* $\epsilon_{min} < 0.1$*,* $\dagger$ *means that* $\epsilon_{min} < 0.2$*,* $\ddagger$ *means that* $\epsilon_{min} < 0.3$*,* $**$ *means that* $\epsilon_{min} < 0.4$*, and* $\dagger\dagger$ *means that* $\epsilon_{min} < 0.5$*. We are not able to perform statistical test regarding baselines in [22], since the authors have not provided the results obtained over individual folds.*

Architecture	Evaluation metrics				
	Precision	**Recall**	**F1-score**	**Accuracy**	**Specificity**
Unimodal approaches					
Only transcript	$94.72^{\ddagger}$ ±5.38	91.78^{**} ±5.77	$93.04^{\dagger}$ ±3.77	$92.49^{\ddagger}$ ±3.97	93.51^{**} ±6.96
Only Speech signal	80.73^{*} ±12.12	85.70^{*} ±9.57	82.49^{*} ±8.51	80.52^{*} ±8.97	74.21^{*} ±16.87
eGeMAPSv02	79.05^{*} ±13.50	85.46^{*} ±7.92	81.67^{*} ±9.69	80.29^{*} ±10.11	76.64^{*} ±15.26
ComParE_2016	86.03^{*} ±8.92	92.29 ±3.96	88.82^{*} ±5.31	87.97^{*} ±4.93	$84.92^{\dagger}$ ±9.49
Baselines reported in [22]					
BS1	73.50 ±16.10	74.50 ±13.20	73.60 ±13.60	73.30 ±10.60	– –
BS2	85.80 ±3.10	86.10 ±2.70	84.70 ±0.90	83.90 ±1.30	– –
Single - Task Learning					
Concatenation	91.51^{*} ±8.74	93.35 ±5.99	$92.11^{\dagger}$ ±5.54	$91.46^{\dagger}$ ±6.05	$90.91^{\dagger}$ ±10.48
GMU	94.10^{**} ±9.51	93.41 ±6.61	93.38^{**} ±6.25	$92.34^{\ddagger}$ ±7.22	92.33^{**} ±11.91
MLB	95.95 ±7.69	91.82^{**} ±6.31	93.57^{**} ±5.37	92.96^{**} ±5.94	95.33 ±9.71
MUTAN	$93.75^{\ddagger}$ ±8.76	94.46 ±5.57	93.82^{**} ±5.71	92.75^{**} ±6.79	90.78^{**} ±13.07
MFH	95.04^{**} ±6.62	$92.79^{\dagger\dagger}$ ±5.01	93.75^{**} ±4.46	$92.94^{\ddagger}$ ±5.56	91.28^{**} ±17.76
MFB	94.68^{**} ±8.19	93.63 ±4.63	93.95^{**} ±5.32	93.18^{**} ±6.13	92.53^{**} ±10.66
BLOCK	**97.30** ±4.43	**94.52** ±4.52	**95.83** ±3.81	**95.29** ±4.23	**96.42** ±6.04
Multi-Task Learning					
Gender, Education, Age	96.14 ±5.02	93.24 ±6.95	$94.38^{\dagger\dagger}$ ±3.65	$94.08^{\dagger\dagger}$ ±3.45	96.31 ±4.86
Gender, Education	97.22 ±5.14	$92.28^{\dagger\dagger}$ ±6.82	$94.51^{\dagger\dagger}$ ±4.65	$94.07^{\dagger\dagger}$ ±5.03	95.95 ±9.35
Education, Age	94.41^{**} ±7.24	93.63 ±5.97	93.74^{**} ±4.52	$93.62^{\dagger\dagger}$ ±4.48	$93.56^{\dagger\dagger}$ ±8.05
Gender, Age	96.55 ±4.87	$92.51^{\dagger\dagger}$ ±6.09	$94.30^{\dagger\dagger}$ ±3.72	$93.84^{\dagger\dagger}$ ±4.25	94.53 ±13.05
Gender	94.61^{**} ±9.28	93.29 ±7.18	93.61^{**} ±6.51	93.20^{**} ±6.81	$93.68^{\dagger\dagger}$ ±10.63
Education	94.22^{**} ±9.16	93.04 ±7.27	$93.44^{\ddagger}$ ±7.34	93.00^{**} ±7.31	92.03^{**} ±12.41
Age	94.99^{*} ±7.46	$92.32^{\dagger\dagger}$ ±6.72	$93.34^{\ddagger}$ ±5.09	$92.56^{\ddagger}$ ±5.85	$93.42^{\dagger\dagger}$ ±10.79

Ablation Study. In this section, we perform a series of ablation experiments to explore the effectiveness of the best performing architecture. Results are reported in Table 2. Firstly, we experiment with removing both the cross-attention layer and the fusion methods. Results show that a decrease of Accuracy ($\epsilon_{min} = 0.25$) and F1-score by 3.14% and 2.59% ($\epsilon_{min} = 0.27$) respectively. Secondly, we remove the cross-attention layer and pass the outputs of Italian BERT and AlexNet through the fusion methods. Findings suggest that Accuracy and F1-score drop by 3.19% ($\epsilon_{min} = 0.16$) and 3.01% ($\epsilon_{min} = 0.16$). Thirdly, we replace the shared layer with two non-shared ones and observe that Accuracy presents a decrease accounting for 2.36% ($\epsilon_{min} = 0.31$), while F1-score is decreased by 2.26% ($\epsilon_{min} = 0.26$). Next, we remove the concatenation mechanisms in the cross-attention layer and pass the outputs of LayerNorm through the fusion methods. Findings suggest that Accuracy and F1-score are decreased by 1.68% ($\epsilon_{min} = 0.45$) and 1.69% ($\epsilon_{min} = 0.38$) respectively. Finally, we remove the shared layer followed by LayerNorm and thus pass the outputs of Add & LayerNorm directly through fusion methods. Results show that Accuracy drops by 1.29% ($\epsilon_{min} = 0.45$).

Table 2: *Ablation Study.* $(*)$ *means that* $\epsilon_{min} < 0.1$*,* $\dagger$ *means that* $\epsilon_{min} < 0.2$*,* $\ddagger$ *means that* $\epsilon_{min} < 0.3$*,* $**$ *means that* $\epsilon_{min} < 0.4$*, and* $\dagger\dagger$ *means that* $\epsilon_{min} < 0.5$ *.*

Architecture	Evaluation metrics				
	Precision	**Recall**	**F1-score**	**Accuracy**	**Specificity**
– Cross-Attention and Fusion Methods	$92.77^{\ddagger}$ ±11.29	94.66 ±5.41	$93.24^{\ddagger}$ ±6.87	$92.15^{\ddagger}$ ±8.15	$90.35^{\ddagger}$ ±15.75
– Cross-Attention	$92.99^{\dagger}$ ±8.22	93.16 ±5.68	$92.82^{\dagger}$ ±5.19	$92.10^{\dagger}$ ±5.47	$92.08^{\ddagger}$ ±9.44
Not shared	96.21 ±5.51	91.66^{**} ±8.01	$93.57^{\ddagger}$ ±4.80	$92.93^{\ddagger}$ ±5.56	94.69 ±7.75
– Concatenation in Cross-Attention Layer	95.49^{**} ±7.73	93.33 ±5.39	94.14^{**} ±4.72	$93.61^{\dagger\dagger}$ ±5.19	95.03 ±8.88
– Shared feed forward and LayerNorm	94.36^{**} ±8.21	95.52 ±4.88	94.60 ±4.45	$94.00^{\dagger\dagger}$ ±5.04	92.29^{**} ±11.01
Proposed Approach	**97.30** ±4.43	**94.52** ±4.52	**95.83** ±3.81	**95.29** ±4.23	**96.42** ±6.04

6. Conclusion and Future Work

In this paper, we present the first study utilizing a cross-attention scaling layer and multimodal fusion methods in a single neural network for detecting depression from spontaneous speech in the Italian language through speech and automatic transcripts. This is also the first study experimenting with a multi-task learning setting to investigate if the prediction of gender, age, and education level as auxiliary tasks aid the depression detection task (primary task) in increasing its performance. Results show that our introduced approach improves competitive baselines in Accuracy by 1.21-21.99% and in F1-score by 1.32-22.23%. Results also show that the introduced single-task learning model outperforms the multitask learning ones. Finally, we perform an ablation study, where we remove several parts of the proposed architecture and observe differences in performance. Findings show degradation in performance in terms of Accuracy by 1.29-3.19%. However, this study comes with some limitations. We did not perform hyperparameter tuning due to limited access to GPU resources. Additionally, we tested our approaches only in one dataset. In the future, we plan to use Neural Architecture Search methods for finding automatically the best performing architecture. Also, explainability and self-supervised learning are some of our future plans.

7. References

[1] W. H. O.(2023)., "*Depressive disorder (depression)*," Available online at: https://www.who.int/news-room/fact-sheets/detail/depression, Accessed: 2024-02-15.

[2] S.Koops, G. S.Brederoo, N. J.de Boer, G. F.Nadema, E. A.Voppel, and E. I.Sommer, "Speech as a biomarker for depression," *CNS & Neurological Disorders - Drug Targets*, vol. 22, no. 2, pp. 152–160, 2023.

[3] J. D.Bernard, J. L.Baddeley, B. F.Rodriguez, and P. A.Burke, "Depression, language, and affect: An examination of the influence of baseline depression and affect induction on language," *Journal of Language and Social Psychology*, vol. 35, no. 3, pp. 317–326, 2016.

[4] N.Seneviratne and C.Espy-Wilson, "Multimodal Depression Severity Score Prediction Using Articulatory Coordination Features and Hierarchical Attention Based Text Embeddings," in *Proc. Interspeech 2022*, 2022, pp. 3353–3357.

[5] M. R.Morales and R.Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *2016 IEEE SLT*, 2016, pp. 136–143.

[6] S.Guohou, Z.Lina, and Z.Dongsong, "What reveals about depression level? the role of multimodal features at the level of interview questions," *Information & Management*, vol. 57, no. 7, pp. 103349, 2020.

[7] F.Tao, X.Ge, W.Ma, A.Esposito, and A.Vinciarelli, "Multi-local attention for speech-based depression detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] Z.Zhang, W.Lin, M.Liu, and M.Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *IEEE FG 2020*, pp. 344–350.

[9] M.Niu, K.Chen, Q.Chen, and L.Yang, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *ICASSP 2021*, 2021, pp. 4235–4239.

[10] G.Lam, H.Dongyan, and W.Lin, "Context-aware deep learning for multi-modal depression detection," in *ICASSP*, 2019.

[11] J.Ye, Y.Yu, Q.Wang, W.Li, H.Liang, Y.Zheng, and G.Fu, "Multi-modal depression detection based on emotional audio and evaluation text," *Journal of Affective Disorders*, vol. 295, 2021.

[12] E.Villatoro-Tello, S. P.Dubagunta, J.Fritsch, G. R.de-la Rosa, P.Motlicek, and M.Magimai-Doss, "Late Fusion of the Available Lexicon and Raw Waveform-Based Acoustic Modeling for Depression and Dementia Recognition," in *Interspeech 2021*, 2021.

[13] F.Ceccarelli and M.Mahmoud, "Multimodal temporal machine learning for bipolar disorder and depression recognition," *Pattern Anal. Appl.*, vol. 25, no. 3, pp. 493–504, aug 2022.

[14] A.Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.

[15] Z.Yu, J.Yu, C.Xiang, J.Fan, and D.Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE TNNLS*, vol. 29, no. 12, pp. 5947–5959, 2018.

[16] M.Rodrigues Makiuchi, T.Warnita, K.Uto, and K.Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, New York, NY, USA, 2019, AVEC '19, p. 55–63, ACM.

[17] T.Al Hanai, M.Ghassemi, and J.Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," in *Proc. Interspeech 2018*, 2018, pp. 1716–1720.

[18] E.Toto, M.Tlachac, and E. A.Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *CIKM '21*. 2021, p. 4145–4154, ACM.

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

[20] P.-C.Wei, K.Peng, A.Roitberg, K.Yang, J.Zhang, and R.Stiefelhagen, "Multi-modal depression estimation based on sub-attentional fusion," in *Computer Vision – ECCV 2022 Workshops*, L.Karlinsky, T.Michaeli, and K.Nishino, Eds., Cham, 2023, pp. 623–639, Springer Nature Switzerland.

[21] M.Rohanian, J.Hough, and M.Purver, "Detecting Depression with Word-Level Multimodal Fusion," in *Proc. Interspeech 2019*, 2019, pp. 1443–1447.

[22] F.Tao, A.Esposito, and A.Vinciarelli, "The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection," in *Proc. INTERSPEECH 2023*, 2023, pp. 4149–4153.

[23] A.Radford, J. W.Kim, T.Xu, G.Brockman, C.McLeavey, and I.Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. 2023, JMLR.org.

[24] B.McFee, C.Raffel, D.Liang, D. P.Ellis, M.McVicar, E.Battenberg, and O.Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8, pp. 18–25.

[25] T.Sachan, N.Pinnaparaju, M.Gupta, and V.Varma, "Scate: shared cross attention transformer encoders for multimodal fake news detection," in *ASONAM '21*, New York, NY, USA, 2022, ACM.

[26] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A. N.Gomez, L.Kaiser, and I.Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, p. 6000–6010, Curran Associates Inc.

[27] J.Arevalo, T.Solorio, M.Montes-y Gomez, and F. A.González, "Gated multimodal networks," *Neural Computing and Applications*, pp. 1–20, 2020.

[28] H.Ben-younes, R.Cadene, M.Cord, and N.Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *ICCV*, 2017, pp. 2631–2639.

[29] J.-H.Kim, K.-W.On, W.Lim, J.Kim, J.-W.Ha, and B.-T.Zhang, "Hadamard product for low-rank bilinear pooling," in *International Conference on Learning Representations*, 2017.

[30] H.Ben-younes, R.Cadene, N.Thome, and M.Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," *AAAI*, vol. 33, no. 01, Jul. 2019.

[31] L.De Lathauwer, "Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness," *SIMAX*, vol. 30, no. 3, pp. 1033–1066, 2008.

[32] J. D.Carroll and J.-J.Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[33] L.Zhao, G.Han, Y.Zhao, Y.Jin, T.Ge, W.Yang, R.Cui, S.Xu, and B.Li, "Gender differences in depression: Evidence from genetics," *Frontiers in Genetics*, vol. 11, 2020.

[34] B.Patria, "The longitudinal effects of education on depression: Finding from the indonesian national survey," *Frontiers in Public Health*, vol. 10, 2022.

[35] F.Eyben, M.Wöllmer, and B.Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, p. 1459–1462, ACM.

[36] E.Del Barrio, J. A.Cuesta-Albertos, and C.Matrán, "An optimal transportation approach for assessing almost stochastic order," in *The Mathematics of the Uncertain*, pp. 33–44. Springer, 2018.

[37] R.Dror, S.Shlomov, and R.Reichart, "Deep dominance - how to properly compare deep neural models," in *ACL 2019*.

[38] D.Ulmer, C.Hardmeier, and J.Frellsen, "deep-significance-easy and meaningful statistical significance testing in the age of neural networks," *arXiv preprint arXiv:2204.06815*, 2022.

[39] N.Reimers and I.Gurevych, "Why comparing single performance scores does not allow to draw conclusions about machine learning approaches," *arXiv preprint arXiv:1803.09578*, 2018.