Multi-modal Fusion for Continuous Emotion Recognition by Using Auto-Encoders

Salam Hamieh, Vincent Heiries, Hussein Al Osman, Christelle Godin

Multi-modal Fusion for Continuous Emotion Recognition by Using Auto-Encoders

Salam Hamieh†
CEA
Grenoble, France
Salam.hamieh@cea.fr

Vincent Heiries
CEA
Grenoble, France
Vincent.heiries@cea.fr

Hussein Al Osman
University of Ottawa
Ottawa, Canada
Hussein.alosman@uottawa.ca

Christelle Godin
CEA
Grenoble, France
Christelle.Godin@cea.fr

ABSTRACT

Human stress detection is of great importance for monitoring mental health. The Multimodal Sentiment Analysis Challenge (MuSe) 2021 focuses on emotion, physiological-emotion, and stress recognition as well as sentiment classification by exploiting several modalities. In this paper, we present our solution for the Muse-Stress sub-challenge. The target of this sub-challenge is continuous prediction of arousal and valence for people under stressful conditions where text transcripts, audio and video recordings are provided. To this end, we utilize bidirectional Long Short-Term Memory (LSTM) and Gated Recurrent Unit networks (GRU) to explore high-level and low-level features from different modalities. We employ Concordance Correlation Coefficient (CCC) as a loss function and evaluation metric for our model. To improve the unimodal predictions, we add difficulty indicators of the data obtained by using Auto-Encoders. Finally, we perform late fusion on our unimodal predictions in addition to the difficulty indicators to obtain our final predictions. With this approach, we achieve CCC of 0.4278 and 0.5951 for arousal and valence respectively on the test set, our submission to MuSe 2021 ranks in the top three for arousal, fourth for valence, and in top three for combined results.

CCS CONCEPTS

• Information systems~Information retrieval~Retrieval tasks and goals~Sentiment analysis •Computing methodologies~Artificial intelligence

MuSe '21, October 24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8678-4/21/10…$15.00
https://doi.org/10.1145/3475957.3484455

KEYWORDS

Affective Computing; Emotion Estimation; Long Short-Term Memory; Gated Recurrent Unit; Auto-encoders; Multi-modality; Audio; Video; Stress

ACM Reference format:

Salam Hamieh, Vincent Heiries, Hussein Al Osman, Christelle Godin. 2021. Multi-modal Fusion for Continuous Emotion Recognition by Using Auto-Encoders. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, Oct. 24, 2021, Virtual Event, China.* ACM, New York, NY, USA. 7 pages. https://doi.org/10.1145/3475957.3484455

1 INTRODUCTION

The field of affective computing is concerned with the design of computer systems capable of analyzing, recognizing, and simulating human emotions. Given the modernization of the world and the integration of computers in our daily life, the need for automatic human emotion recognition is increasingly gaining importance. Affective computing approaches are proving to be valuable for educational systems [41], social robots[21], healthcare applications [19], and many other technologies that involve Human-Computer Interaction (HCI) capacities.

"An emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response" [13]. The latter two are manifested most prominently through facial expressions, speech, and physiological indicators such as electrocardiography, blood-pressure volume, etc. Humans as well as the computers recognize emotions by analyzing these signals. Therefore, significant work has been dedicated to find the best modalities and features for affective computing. The predecessor of MuSe, the Audio/Visual Emotion Challenge (AVEC), focused mainly on the exploitation of the audio and visual modalities. The MuSe challenge focuses on biological signals as well.

There are generally two approaches to emotional modeling: the categorical and the dimensional approach. In the categorical approach, we define emotions as a set of discrete classes, e.g., Ekman's basic emotions[7]. As for the dimensional approach, its

coordinates in the Euclidean space characterize the emotion. Russel's model [29] is one of the most widely adopted dimensional representations in affective computing. It describes an emotion using three dimensions: valence (positivity of the emotion), arousal (intensity), and dominance (degree of control).
The MuSe-stress sub-challenge targets predicting the arousal and valence for people in a stress-induced situation. It provides several features sets covering the audio, visual, textual, and physiological modalities. The participants are encouraged to explore and combine these signals and obtain an optimal continuous predictor.
The first step in building a robust model is feature selection. In earlier work, mostly hand-crafted features were employed [30–32]. However, deep data representations by neural networks proved to be effective as well [43]. Therefore, in our approach, we explore both low-level and high-level features for the audio, visual, and textual modalities. In the second step, we build our model using a BiGRU network to ensure capturing contextual information from the input signals. To enhance our model's performance, we feed our model an additional input "difficulty indicator" on unimodal prediction and fusion stage. The difficulty indicator is obtained using an Auto-Encoder (AE) for each modality.
The remainder of this article is organized as follows. In Section 2, we briefly present related work. In Section 3, we give a detailed explanation of our approach and describe the structure of our model. Then, in Section 4, we provide a description of the dataset, features, and model training settings. In section 5, we present and discuss the results of our experiments. Finally, we present our conclusions in Section 6.

2 RELATED WORK

2.1 Features

In earlier work, handcrafted features were used in speech-based emotion recognition systems (SER)[30–32], e.g. Mel-frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP), etc. Recently, with the advent of deep learning, more effort has been expended on creating models with as little human intervention as possible. Trigeorgis et al. [37] demonstrated that their end-to-end SER system that takes a raw wavelet as input outperforms traditional ones. Moreover, Zhao et al. [43] showed that deep representations extracted using the VGGish model [12] surpass expert-based knowledge features for the task of arousal and valence prediction in AVEC 2018. As for the visual modality, Face Action Units (FAU) [28] are the most used features for emotion recognition. In the early AVEC series, expert-knowledge features such as Local Binary Patterns (LBP) [20] and Gabor[10] were used as video baseline features for emotion prediction. Latterly, deep representations [25,35] extracted through deep convolutional neural networks (e.g. VGG-16 [33]) have been used.

2.2 Models

Traditional machine learning approaches such as Support Vector Machine Regression (SVR) [26,27,39,40] were predominantly used in earlier work on automatic emotion recognition. However, given that these models are not capable of capturing temporal dependencies, which is crucial for continuous emotion prediction tasks, they have been largely replaced by Recurrent Neural Network (RNN) approaches [3]. Today, Long Short Term Memory (LSTM) networks are used as the baseline for emotion recognition challenges [25,34,35]. Sun et al. [36], the winners of MuSe2020, proved that adding a self-attention mechanism to LSTM boosts its performance. Moreover, with the increased popularity of end-to-end learning, Convolutional Neural Networks (CNN) followed by LSTM-RNN are often used [38] [17]. The CNNs extract the features most relevant for the desired task and LSTM ensures capturing dynamic temporal relationships. To further improve performance, some researchers are incorporating data difficulty information into their models. Han et al. [11] showed that adding the uncertainty as an output along the prediction of emotions improves the performance of the model for the task of emotion recognition. Zhang et al. [42] exploited difficulty indicators of the data to update the weights of the model in a manner where more effort is provided with high difficulty data; their work outperforms state-of-the-art.

2.3 Multimodal Fusion

Since emotions can be recognized through several modalities, the fusion of multimodal information plays an important role. There are three established approaches to combine different modalities: early fusion, late fusion, and model fusion. Early fusion consists of feeding the features of all modalities to a single model at once, whereas late fusion combines the predictions of several models. Model-level fusion is a compromise between the latter two where the fusion happens between the intermediate representations of the multimodal features. In AVEC 2018, Huang et al. [14] showed that late fusion is better at predicting valence and arousal than early fusion. Sun et al. [36] chose late fusion with an LSTM network that captures dynamic information as a fusion model. Moreover, Chen et al. [2] proposed the joint use of early and late fusion using Bidirectional Deep Long Short-Term Memory networks (BDLSTM). The results showed that early and late information may be complementary [2].

3 METHOD

In this section, we explain our approach in detail. Our method focuses on exploiting several modalities for continuous emotion prediction and optimizing the fusion of the signals. We train a unimodal predictor for each modality. We use a fusion model to combine the results of the unimodal predictors. To enhance the prediction performance, we add an extra feature called "difficulty indicator". We obtain the difficulty indicator of the modality by training an AE on the corresponding features set and calculating its Reconstruction Error (RE). As shown in Fig. 1, we feed the RE error of the unimodal features to the unimodal predictor. Then, we feed the outputs of each unimodal predictor along with the respective REs to the fusion model. Our approach is built with

three training stages: Auto-Encoder (3.1), Unimodal Predictions (3.2) and Fusion Model (3.3) training.

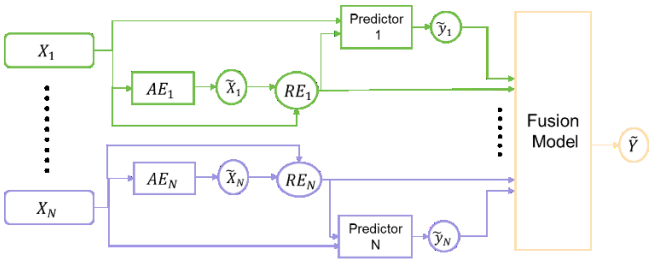

Figure 1: Diagram of the proposed solution. X_i refers to the i^{th} unimodal features set. $\widetilde{X_i}$ refers to their reconstruction using the auto-encoder. RE refers to the averaged reconstructed error. $\tilde{y}_i$ refers to the unimodal prediction using the ith features set and $\tilde{Y}$ refers to the multi-modal end prediction.

3.1 Auto-Encoder:

Auto-encoders (AE) are neural networks that encode the inputs into a new dimensional space and then reconstruct the input from the encoded values. In the case of a basic auto-encoder with one hidden layer, an input example $x \in \mathbb{R}^d$ will pass through the hidden layer $h(x) \in \mathbb{R}^p$ where:

$$h(x) = g(W_1 x + b_1) \quad \#(1)$$

and $g(z)$ is a non-linear activation function. Then, the model will decode the hidden representation $h(x)$ to produce a reconstruction of the input $\tilde{x} \in \mathbb{R}^d$:

$$\tilde{x} = g(W_2 h(x) + b_2) \#(2)$$

The training of the AE consists in finding the parameters $\{W_1, W_2, b_1, b_2\}$ that minimizes the Reconstruction Error (RE) which is described in the following loss function:

$$RE = \mathcal{L}(W_1, W_2, b_1, b_2) = \sum_{x \in \mathbb{R}^d} \left\| x - \tilde{x} \right\|^2 \#(3)$$

Typically, an AE is used to solve the high dimensionality problem since an increase in dimensions raises the required complexity of the model and demand in data and computation capacities. However, more recently, AEs are being used to detect anomalies/novelties [18] e.g. unexpected or unusual events. In this case, the RE determines whether the input to the AE is a novelty. This is achieved by comparing the RE to a set threshold, the sample is considered a novelty if its RE surpasses the threshold. AE is also used for classification [23]. Hence, an AE is trained for each class and the RE is used as a class membership indicator. More recently, Zhang et al.[42] used RE in dynamic difficulty awareness training (DDAT). DDAT relies on the assumption that a model will perform better if it is provided with the learning difficulty of the data. They train a model that reconstructs the input and predicts emotions in a multi-task learning framework. They calculate the RE of the inputs and use it as a difficulty indicator to update the model. The RE is re-injected into the model to update its weights accordingly. Similar to [42], we train an AE for each feature set. Given that the RE represents the difficulty of the task for the regressor (the unimodal predictor in our case), the AE and the regressor are designed with the same architecture. This architecture is detailed in section 3.2. We define the difficulty indicator as the averaged mean squared error between the input and its reconstruction .We use this indicator in two stages: 1. Unimodal predictions stage: we feed the features and their corresponding RE to the prediction model, and 2. the fusion stage: we feed the predictions + RE to the fusion model. These stages are further explained in sections 3.2 and 3.3.

3.2 Unimodal prediction

For each modality, we train a separate regressor for arousal and valence continuous prediction. As shown in Fig. 1, we feed each regressor the feature set along with the RE calculated from the AE of the respective feature set. We assume the RE represents the difficulty of the data, and that this information will help the model perform better. Temporal information is of great importance for emotion recognition. Therefore, we utilize a 4-layered Recurrent Neural Networks (RNN) followed by a feed forward layer for the architecture of our model. We evaluate two types of RNNs: a LSTM and a GRU networks.

LSTM is a type of recurrent neural network that allows us to capture temporal information. It can process sequential inputs by using its internal state (memory). In contrast to conventional RNN, LSTM has a cell variable c_t and three gates: input gate i_t, output gate o_t, and forget gate f_t. These gates help the LSTM overcome the vanishing gradient problem that the RNN suffers from. Moreover, it allows it to better handle long input sequences. The equations of the forward pass of an LSTM are the following:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \#(4)$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \#(5)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \#(6)$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes \sigma_h(W_c x_t + U_c h_{t-1} + b_c) \#(7)$$
$$h_t = o_t \otimes \sigma_h(c_t) \#(8)$$

Where x_t is the current passed input, h_t the current hidden state, σ_g and σ_h are the sigmoid and hyperbolic tangent functions and $\otimes$ denotes element-wise multiplication. W, U and b are the weight matrices and biases.

The GRU is a simplified version of the LSTM. It has only two gates: the update gate z_t and the reset gate r_t[4]. It has less parameters than the LSTM. It typically has a comparable performance to the LSTM [5,24].

A forward pass of a sample x_t through the GRU is described in the following equations:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \#(9)$$
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \#(10)$$
$$\hat{h}_t = \sigma_h(W_h x_t + U_h(r_t \otimes h_{t-1}) + b_h) \#(11)$$
$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \hat{h}_t \#\#(12)$$

Where x_t is the current passed input, h_t the hidden state at time t, σ_g and σ_h are the sigmoid and hyperbolic tangent functions and $\otimes$ denotes element-wise multiplication. W, U and b are the weight matrices and biases.
We also explore using bidirectional LSTM and GRU, where the inputs are processed in both directions. After passing through recurrent layers, we further encode the outputs through a feed-forward layer.

3.3 Fusion Model

In our solution, we adopt late fusion strategy. We feed our fusion model the predictions from each modality and the RE of the respective modality, as shown in Fig. 1. We assume that the RE not only indicates the difficulty of the data but may also denote the level of reliance on each unimodal feature set. Our fusion model consists of a four-layered bidirectional RNN.

4 EXPERIMENTS

4.1 Corpus Description

The dataset used for the MuSe-Stress sub-challenge is the Ulm-TSST database[34]. It consists of 69 German-speaking participants, aged between 18 and 39 years, in a stress inducing setup following
the Trier Social Stress Test procedure [15]. Each participant is asked to present orally for approximately five minutes in front of two interviewers. Video and physiological recordings are taken during the presentation. The total duration of the database is 5h: 47min: 27s. Arousal and valence are annotated by three raters and the fusion of the annotations is done using the RAAW method[34]. The given modalities are audio, video, transcripts in addition to the physiological signals Electrodermal Activity (EDA), Electrocardiogram (ECG), respiration and heart rate (BPM).

4.2 Features

The MuSe2021 provides a range of extracted acoustic, visual, and textual features for the participants to use. In our approach, we used the following parameter sets.

4.2.1 Acoustic Features

eGeMAPS features: We explore using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [8] which can be extracted using the free openSMILE toolkit[9]. It consists of low-level acoustic descriptors including frequency, energy, and spectral parameters. To extract these features, the audio signal is divided in overlapping 6 seconds windows. We normalize the eGeMAPS features.
DeepSpectrum: We also experiment with DeepSpectrum features[1]. DeepSpectrum is a deep CNN pre-trained with the spectrograms of audio signals. 4096 features were extracted using this model.

4.2.2 Visual Features

VGGFace: VGGFace[22] is the output result from a deep CNN used for face recognition. The VGGFace features correspond to the 512-output vector after detaching the last layer of the model.
FAU: Using the Multi Cascaded Convolutional Neural Networks (MTCNN), 17 facial action unit intensities are obtained from the center and left sides of the face.

4.2.3 Textual Features

Bert : For the textual modality, Bidirectional Encoder Representations from Transformers (BERT) [6] features are provided. The high-level contextual embedding proved to deliver state of the art results for several Natural Language Processing (NLP) tasks. Since Ulm-TSST is a German database, BERT is pre-trained on German texts.

4.3 Model Training

We implement our solution using the Pytorch toolkit [44] . For each features set, our unimodal predictors, AEs, and late fusion model share the same architecture: four-layered bi-directional RNN with 64 hidden neurons followed by a feedforward layer. The choice between BiLSTM and biGRU as recurrent layer is determined based on the results on the validation set. We utilize the Adam optimizer and varied learning rates (0.001, 0.005, 0.0005). As a form of regularization, we apply dropout and evaluate with several rates (0.1, 0.2, or 0.5). We train the model for 100 epochs at most and apply early stopping if the validation performance does not improve after 15 epochs. For the loss function, we use the Concordance Correlation Coefficient (CCC) loss [16] which is defined as:

$$\mathcal{L} = 1 - CCC \#(13)$$

$$CCC = \frac{2\rho\sigma_{\hat{Y}}\sigma_y}{{\sigma_{\hat{Y}}}^2 + {\sigma_y}^2 + (\mu_{\hat{Y}} - \mu_Y)^2} \#(14)$$

where $\mu_{\hat{Y}}$ and μ_Y are the mean of the prediction $\hat{Y}$ and the label Y,and $\sigma_{\hat{Y}}$ and σ_y are the corresponding standard deviations. ρ is the Pearson Correlation Coefficient (PCC) between $\hat{Y}$ and Y.

5 RESULTS

5.1 Ablation study

We conduct an ablation study to determine the best type of recurrent layer for our model. We explore four types of models: LSTM, GRU, biGRU, and biLSTM using the standard approach where only unimodal features are inputted to the model. All of the four models are 4-layered networks with 64 hidden neurons. We present the results of the performance of each model in Table 1 and Table 2 for arousal and valence respectively. Generally, the BiGRU and BiLSTM models outperform both LSTM and GRU models. These results show that both past and future information is relevant for emotion prediction. Since BiGRU and BiLSTM have close performances and BiGRU have less parameters, we choose to continue with BiGRU model for better generalization on the testing set. We also experiment with 3 loss functions on the BiGRU model: "CCC", "MSE", and "L1" loss. Generally, CCC gives better performance as shown in Table 3 and Table 4. The results agree with the findings in [36,37].

Table 1: CCC performance comparison between recurrent models for unimodal predictions on arousal dimension on the validation set.

Features	LSTM	BiLSTM	GRU	BiGRU
eGeMAPS	0.4714	**0.5466**	0.4739	0.5322
VGGface	0.1809	**0.3561**	0.1283	0.2293
FAU	0.3260	0.3637	0.3641	**0.3688**
Bert	0.2250	**0.3166**	0.2349	0.2681
DeepSpectrum	**0.3339**	0.2617	0.2538	0.2185

Table 2: CCC performance comparison between recurrent models for unimodal predictions on valence dimension on the validation set.

Features	LSTM	BiLSTM	GRU	BiGRU
eGeMAPS	**0.5926**	0.5597	0.5671	0.5646
VGGface	0.5650	0.5671	0.5414	**0.6481**
FAU	0.5480	**0.5952**	0.5531	0.5143
Bert	0.3025	0.2538	0.2828	**0.4473**
DeepSpectrum	0.5548	**0.5678**	0.5532	0.5630

5.2 Reconstruction Error and Unimodal Predictions

In Table 5 and Table 6, we present the performances achieved on the validation set for the arousal and valence prediction where we feed the model unimodal features only compared to unimodal features and RE. The results show improvement for almost all the modalities for both tasks. This indicates that the difficulty information helps the model perform better. Our results are consistent with those of [42].

Table 3: CCC performance on the arousal obtained by using different loss functions on the validation set.

Features	MSE	L1	CCC
eGeMAPS	0.3038	0.3306	**0.5322**
VGGFace	0.0772	0.0159	**0.2293**
FAU	0.3083	**0.4354**	0.3688
Bert	**0.3277**	0.3160	0.2681
DeepSpectrum	0.0666	0.1201	**0.2185**

Table 4: CCC performance on the valence obtained by using different loss functions on the validation set.

Features	MSE	L1	CCC
eGeMAPS	0.4465	0.4414	**0.5646**
VGGFace	0.5354	0.3980	**0.6481**
FAU	0.4011	0.2885	**0.5143**
Bert	0.3658	0.3095	**0.4473**
DeepSpectrum	0.5262	0.4984	**0.5630**

Table 5: CCC performance comparison for unimodal predictions on arousal dimension on the validation set

Inputs	Model	
	Unimodal Features	Unimodal Features + RE
eGeMAPS	0.5322	**0.5829**
VGGFace	0.2293	**0.3926**
FAU	**0.3688**	0.3668
Bert	0.2681	**0.3457**
DeepSpectrum	0.2185	**0.2498**

Table 6: CCC performance comparison for unimodal predictions on valence dimension on the validation set

Modalities	Model	
	Unimodal Features	Unimodal Features + RE
eGeMAPS	0.5646	**0.6353**
VGGFace	0.6481	**0.6798**
FAU	0.5143	**0.5307**
Bert	0.4473	**0.4655**
DeepSpectrum	0.5630	**0.5676**

5.3 Multimodal predictions

First, we conduct fusion on the unimodal predictions only. We try several combinations for fusion, as shown in Table 7. We observe that fusing several modalities boosts the performance significantly (eGeMAPS+ VGGFace, eGeMAPS+ VGGFace + FAU). We can also find that for the visual modality, using low-level (FAU) and high-level features (VGGface) results in better performance. We notice that adding the textual modality causes a drop in the performance; this can be explained by the fact that textual information in interviews may not reflect true emotions. We also observe that adding DeepSpectrum features does not improve the performance. This could be the result of the high dimensionality of this features set (4096 features). Second, we apply our proposed approach by fusing the best combination of the unimodal predictions (eGeMAPS + VGGFace + FAU) and RE's of the corresponding features sets. As shown in Table 8, we achieve better results by adding the RE's of features sets as extra features at fusion stage. We hypothesize that the addition of this information at the fusion level leads the model to rely more on the unimodal predictions that are most reliable.

Table 7: CCC performance of multi-modal features on the arousal and valence dimension on the validation set.

Modalities	Arousal	Valence
eGeMAPS + VGGFace	0.6205	0.7024
eGeMAPS + VGGFace+ Bert	0.6031	0.6811
eGeMAPS + VGGFace+ DeepSpectrum	0.6199	0.7320
eGeMAPS + VGGFace+ FAU	**0.6469**	**0.7653**

Table 8: CCC performance on arousal and valence using as fusion model inputs, unimodal predictions only (first row) and unimodal prediction along with RE (second row) on the validation set

Inputs	Arousal	Valence
Fused Unimodal predictions	0.6469	0.7653
Fused Unimodal predictions + RE	**0.6554**	**0.8036**

Table 9: Best submission results of our approach on the validation and test sets.

Partition	Emotion	Baseline	Proposed
Val	Arousal	0.5043	**0.6554**
	Valence	0.6966	**0.8036**
	Combined	0.6005	**0.7295**
Test	Arousal	**0.4562**	0.4278
	Valence	0.5614	**0.5951**
	Combined	0.5088	**0.5115**

Our final submissions are shown in Table 9. Our proposed approach significantly outperforms the baseline on the validation set for arousal and valence predictions. On the test set, our method outperforms the baseline system on the valence with CCC 0.5951 vs 0.5614. As for arousal, we achieve CCC of 0.4278 vs 0.4562 for the baseline. The difference between our performance on the validation and test set indicates that there may be an overfitting on the validation set. It can also imply that the two sets may have different distributions. Our submissions rank fourth for the valence prediction, in the top three for arousal prediction, and in the top three for combined performance in the Muse-Stress 2021.

6 CONCLUSION

In this paper, we presented our solution for the sub-challenge MuSe-Stress of MuSe 2021. We explored the performance of high-level and low-level features for continuous emotion prediction. We compared the performance of uni and bi-directional GRU and LSTM and chose BiGRU as our main model to extract temporal dynamic information from the data. To further enhance our model's performance, we train an AE for each feature set and calculate its corresponding RE to represent the data's difficulty. Our results showed that adding the RE as input for the uni-modal prediction as well as for the fusion improves the model's performance. These findings confirm that the RE reflects relevant information.

However, as a limitation in this work, our model likely overfit the validation set. The gap for the CCC performance between the validation and test sets show that our model required additional data to generalize. This may be addressed using data augmentation methods. Moreover, because RE helps to focus on relevant feature sets, we think that using RE could bring more robustness to the fusion model toward corrupted modalities. In the future, we would like to test this hypothesis by adding noise to feature sets or directly to raw data.

REFERENCES

[1] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features. In *Interspeech 2017*, ISCA, 3512–3516. DOI:https://doi.org/10.21437/Interspeech.2017-434

[2] Haifeng Chen, Yifan Deng, Shiwen Cheng, Yixuan Wang, Dongmei Jiang, and Hichem Sahli. 2019. Efficient Spatial Temporal Convolutional Features for Audiovisual Continuous Affect Recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop - AVEC '19*, ACM Press, Nice, France, 19–26. DOI:https://doi.org/10.1145/3347320.3357690

[3] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ACM, Mountain View California USA, 19–26. DOI:https://doi.org/10.1145/3133944.3133949

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv14061078 Cs Stat* (September 2014). Retrieved August 6, 2021 from http://arxiv.org/abs/1406.1078

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv14123555 Cs* (December 2014). Retrieved August 6, 2021 from http://arxiv.org/abs/1412.3555

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs* (May 2019). Retrieved July 21, 2021 from http://arxiv.org/abs/1810.04805

[7] Paul Ekman. 1999. Basic emotions. In *Handbook of cognition and emotion*. John Wiley & Sons Ltd, New York, NY, US, 45–60. DOI:https://doi.org/10.1002/0470013494.ch3

[8] Florian Eyben, Klaus Scherer, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.*, 14.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia - MM '10*, ACM Press, Firenze, Italy, 1459. DOI:https://doi.org/10.1145/1873951.1874246

[10] D. Gabor. 1946. Theory of communication. Part 1: The analysis of information. *J. Inst. Electr. Eng. - Part III Radio Commun. Eng.* 93, 26 (November 1946), 429–441. DOI:https://doi.org/10.1049/ji-3-2.1946.0074

[11] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, ACM, Mountain View California USA, 890–897. DOI:https://doi.org/10.1145/3123266.3123383

[12] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, New Orleans, LA, 131–135. DOI:https://doi.org/10.1109/ICASSP.2017.7952132

[13] Don H. Hockenbury and Sandra E. Hockenbury. 2007. *Discovering psychology, 4th ed.* Worth Publishers, New York, NY, US.

[14] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. 2018. Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ACM, Seoul Republic of Korea, 57–64. DOI:https://doi.org/10.1145/3266302.3266304

[15] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. 1993. The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress

Responses in a Laboratory Setting. *Neuropsychobiology* 28, 1–2 (1993), 76–81. DOI:https://doi.org/10.1159/000119004

[16] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268. DOI:https://doi.org/10.2307/2532051

[17] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ACM, Amsterdam The Netherlands, 35–42. DOI:https://doi.org/10.1145/2988257.2988267

[18] Erik Marchi, Fabio Vesperini, Stefano Squartini, and Björn Schuller. 2017. Deep Recurrent Neural Network-Based Autoencoders for Acoustic Novelty Detection. *Comput. Intell. Neurosci.* 2017, (2017), 1–14. DOI:https://doi.org/10.1155/2017/4694860

[19] Mingyue Niu, Jianhua Tao, Bin Liu, and Cunhang Fan. 2019. Automatic Depression Level Detection via ℓp-Norm Pooling. In *Interspeech 2019*, ISCA, 4559–4563. DOI:https://doi.org/10.21437/Interspeech.2019-1617

[20] T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 7 (July 2002), 971–987. DOI:https://doi.org/10.1109/TPAMI.2002.1017623

[21] Anastasia K. Ostrowski, Daniella DiPaola, Erin Partridge, Hae Won Park, and Cynthia Breazeal. 2019. Older Adults Living With Social Robots: Promoting Social Connectedness in Long-Term Communities. *IEEE Robot. Autom. Mag.* 26, 2 (June 2019), 59–70. DOI:https://doi.org/10.1109/MRA.2019.2905234

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep face recognition. (2015). Retrieved July 21, 2021 from https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1

[23] Stavros Petridis and Maja Pantic. 2016. Prediction-Based Audiovisual Fusion for Classification of Non-Linguistic Vocalisations. *IEEE Trans. Affect. Comput.* 7, 1 (January 2016), 45–58. DOI:https://doi.org/10.1109/TAFFC.2015.2446462

[24] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. 2018. Light Gated Recurrent Units for Speech Recognition. *IEEE Trans. Emerg. Top. Comput. Intell.* 2, 2 (April 2018), 92–102. DOI:https://doi.org/10.1109/TETCI.2017.2762739

[25] Fabien Ringeval, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, Maja Pantic, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, and Shahin Amiriparian. 2019. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop - AVEC '19*, ACM Press, Nice, France, 3–12. DOI:https://doi.org/10.1145/3347320.3357688

[26] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ACM, Mountain View California USA, 3–9. DOI:https://doi.org/10.1145/3133944.3133953

[27] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, ACM, Brisbane Australia, 3–8. DOI:https://doi.org/10.1145/2808196.2811642

[28] Erika L. Rosenberg and Paul Ekman. 2020. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.

[29] James Russell. 1980. A Circumplex Model of Affect. *J. Pers. Soc. Psychol.* 39, (December 1980), 1161–1178. DOI:https://doi.org/10.1037/h0077714

[30] Enrique Sánchez-Lozano, Paula Lopez-Otero, Laura Docio-Fernandez, Enrique Argones-Rúa, and José Luis Alba-Castro. 2013. Audiovisual three-level fusion for continuous estimation of Russell's emotion circumplex. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, ACM, Barcelona Spain, 31–40. DOI:https://doi.org/10.1145/2512530.2512534

[31] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. 2011. AVEC 2011–The First International Audio/Visual Emotion Challenge. In *Affective Computing and Intelligent Interaction*, Sidney D'Mello, Arthur Graesser, Björn Schuller and Jean-Claude Martin (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 415–424. DOI:https://doi.org/10.1007/978-3-642-24571-8_53

[32] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12*, ACM Press, Santa Monica, California, USA, 449. DOI:https://doi.org/10.1145/2388676.2388776

[33] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* (April 2015). Retrieved July 18, 2021 from http://arxiv.org/abs/1409.1556

[34] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W Schuller. 2021. The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress. (2021), 10.

[35] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, ACM, Seattle WA USA, 35–44. DOI:https://doi.org/10.1145/3423327.3423673

[36] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multi-modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, ACM, Seattle WA USA, 27–34. DOI:https://doi.org/10.1145/3423327.3423672

[37] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Bjorn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Shanghai, 5200–5204. DOI:https://doi.org/10.1109/ICASSP.2016.7472669

[38] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Bjorn W. Schuller, and Stefanos Zafeiriou. 2017. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* 11, 8 (December 2017), 1301–1309. DOI:https://doi.org/10.1109/JSTSP.2017.2764438

[39] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ACM, Amsterdam The Netherlands, 3–10. DOI:https://doi.org/10.1145/2988257.2988258

[40] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*, ACM Press, Orlando, Florida, USA, 3–10. DOI:https://doi.org/10.1145/2661806.2661807

[41] Elaheh Yadegaridehkordi, Nurul Fazmidar Binti Mohd Noor, Mohamad Nizam Bin Ayub, Hannyzzura Binti Affal, and Nornazlita Binti Hussin. 2019. Affective computing in education: A systematic review and future research. *Comput. Educ.* 142, (December 2019), 103649. DOI:https://doi.org/10.1016/j.compedu.2019.103649

[42] Zixing Zhang, Jing Han, Eduardo Coutinho, and Björn Schuller. 2019. Dynamic Difficulty Awareness Training for Continuous Emotion Prediction. *IEEE Trans. Multimed.* 21, 5 (May 2019), 1289–1301. DOI:https://doi.org/10.1109/TMM.2018.2871949

[43] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. 2018. Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ACM, Seoul Republic of Korea, 65–72. DOI:https://doi.org/10.1145/3266302.3266313

[44] PyTorch: An Imperative Style, High-Performance Deep Learning Library. Retrieved August 7, 2021 from https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html