Self-supervised representation learning using multimodal Transformer for emotion recognition

Theresa Götz **??**
Fraunhofer Institute for integrated Circuits IIS 91054 Erlangen, Germany; CIML Group, Biophysics, University of Regensburg, 93040 Regensburg, Germany; Clinic of Rheumatology, University Hospital Erlangen 91054 Erlangen, Germany; Department of Industrial Engineering and Health, Technical University of Applied Sciences Amberg-Weiden, Weiden, Germany
theresa.goetz@iis.fraunhofer.de

Pulkit Arora
Fraunhofer Institute for integrated Circuits IIS 91054 Erlangen, Germany
pulkit.arora@iis.fraunhofer.de

F. X. Erick
Fraunhofer Institute for integrated Circuits IIS 91054 Erlangen, Germany
franciskus.xaverius.erick@iis.fraunhofer.de

Nina Holzer
Fraunhofer Institute for integrated Circuits IIS 91054 Erlangen, Germany
shrutika.sawant@iis.fraunhofer.de

Shrutika Sawant
Fraunhofer Institute for integrated Circuits IIS 91054 Erlangen, Germany
shrutika.sawant@iis.fraunhofer.de

ABSTRACT

In this paper, we present a Modality-Agnostic Transformer based Self-Supervised Learning (MATS^{2}L) for emotion recognition using physiological signals. The proposed approach consists of two stages: a) Pretext stage, where the transformer model is pre-trained with unlabeled physiological signal data using masked signal prediction as pre-training task and form contextualized signal representations. b) Downstream stage, where self-supervised learning (SSL) representations extracted from a pre-trained model are utilized for emotion recognition tasks. Modality-agnostic approach allows the transformer model to focus on exploring mutual features among different physiological signals and learning more meaningful embeddings to estimate emotions effectively. We conduct several experiments on a public dataset WESAD and perform comparisons with fully supervised and other competitive SSL approaches. Experimental results showed that the proposed approach is capable of learning meaningful features and superior to other competitive SSL approaches. Moreover, a transformer model trained on SSL features outperforms fully supervised transformer model. We also present detailed ablation studies to prove the robustness of our approach.

iWOAR 2023, September 21, 22, 2023, Lübeck, Germany

ACM ISBN 979-8-4007-0816-9/23/09.
https://doi.org/10.1145/3615834.3615837

CCS CONCEPTS

• **Computing methodologies**; • **Machine Learning**; • **Learning paradigms; Self-supervised learning**; • **Deep learning networks**; • **Transformer**;

KEYWORDS

Emotion recognition, Physiological signals, Modality-agnostic, Self-supervised learning, Transformer

ACM Reference Format:
Theresa Götz **??**, Pulkit Arora, F. X. Erick, Nina Holzer, and Shrutika Sawant. 2023. Self-supervised representation learning using multimodal Transformer for emotion recognition. In *8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence (iWOAR 2023), September 21, 22, 2023, Lübeck, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3615834.3615837

1 INTRODUCTION

Humans express their emotions in several ways, including facial expressions, speech, body language and changes in physiological signals. Understanding human emotions is a complex task, since emotions are considered as physiological and psychological expressions that depend on an individual's mood and personalities. Emotion recognition has become an emerging research field not only within the application of so-called 'affective computing', but also in other research domains such as healthcare, psychology, robotics, e-gaming, or cognitive studies [1-3]. Most of the research on emotion recognition so far has focused on the analysis of the modalities, such as facial expressions, speech or text. Recently, physiological signals such as the electrocardiogram (ECG), Electrodermal activity (EDA) or galvanic skin response (GSR), skin temperature, electroencephalogram (EEG), and respiration have shown significant performance in understanding different emotional states of

humans. In the past decades several machine learning approaches have been studied to recognize emotions by exploiting facial expressions from images and videos [4-5], by extracting behavioral or emotional signs from audio signals or by analyzing different physiological signals [1, 6-8]. Lately, deep learning based approaches have been widely studied for effective emotion recognition and achieved promising results [9-11]. This has attributed to the power of deep neural networks and their variants to explore meaningful emotion related features. However, most of the existing machine learning and deep learning based emotion recognition models need abundant labeled data to work effectively. It is complex and time consuming process to provide labels to physiological data (or get human annotated data). Moreover, the annotation process needs an expert's knowledge. These requirements have motivated researchers to utilize the plethora of unlabeled data available in the dataset and learn their representations. Therefore, to overcome the issue of labeled data scarcity, pre-training models in a self-supervised manner has become a prominent solution in computer vision (CV) and natural language processing (NLP) domains.

Self-supervised learning (SSL) allows models to learn the general and robust representations from unlabeled data, reducing the need for labeled data. In the field of affective computing, many SSL based methods have been introduced in the last few years for emotion recognition. However, most of them are focusing on modalities such as audio, video or text data and little research has been done for learning representations from physiological signals using SSL. Inspired by the success of SSL in affective computing, in this contribution we propose a Modality-Agnostic Transformer based Self-Supervised Learning (MATS^{2}L) for classifying various emotional states based on physiological signals. Transformer models have been seen to achieve quite a success in NLP for interpreting sequence of words as well as handling multiple modalities such as, audio, text or videos in CV due to its excellent attention mechanism. The use of SSL feature representations provide a new paradigm to address the label scarcity problem in physiological signal based emotion recognition.

Over the past few years, physiological signals based deep learning approaches have demonstrated excellent performances in classifying emotions. However, which physiological signals are to be selected and how to combine them for capturing the most discriminative emotion-related features remain challenging. Among the different physiological signals, specifically ECG and EDA are seen to be most helpful in extracting emotion-related features [11-13]. Hence, in this work, we explore MATS^{2}L to learn shared representations from ECG and EDA signals for estimating emotions effectively. We evaluate our approach on the publicly available WESAD dataset. Experimental analysis shows that our proposed approach is more effective for multimodal emotion recognition and is furthermore more advantageous than fully supervised methods. The main contributions of our work can be briefly summarized as follows:

1. We present MATS2L, a novel SSL approach to learn multimodal representations by capturing meaningful features across different physiological signals, especially, ECG and EDA.
2. We use SSL feature representations extracted from a pre-trained model that can be effectively adapted for the downstream emotion recognition task.
3. We conduct a series of experiments on a publicly available dataset and demonstrate that the proposed transformer-based SSL model performs effectively better than the same transformer model when trained in a fully supervised manner.

2 RELATED WORKS

Since SSL is an active field of research with diverse research directions, for concision, we only discuss applications of SSL in the field of affective computing. There exist very few works in the field of affective computing that have utilized SSL methods to physiological signals. Sarkar and Etemad [14] explored the idea of SSL for learning ECG representations during pre-training tasks and used them for a downstream emotion recognition task. Six different signal transformations were applied on ECG to train the 1DCNN model using unlabeled ECG data. The weights of the pre-trained model are then transferred to downstream tasks to carry out emotion recognition tasks in a fully supervised manner. Similarly, Quispe et.al [15] discussed the use of SSL for learning ECG representations and subsequently employing them for emotion recognition. In [15], authors used the same pre-training framework as mentioned in [14], which is transferred and fine-tuned in the downstream stage, whereas the pre-trained network is kept frozen in the downstream stage of [14]. Dissanayake et.al [16] proposed a contrastive learning based SSL approach for learning representations from various physiological signals and concatenating the learned feature embeddings to estimate emotions. The generative adversarial network based SSL using EEG data is introduced in [17] for emotion recognition. The presented architecture uses an adversarial augmentation network to generate masked parts of EEG signal and synthesize the EEG signals. A multi-factor training network then uses stimulated EEG signals for training the emotion recognition model. Shen et al. [29] have used contrastive learning methods to learn data representations from EEG signals and then applied the pre-trained model to downstream emotion recognition tasks.

While transformer based SSL architectures have seen to be dominant across NLP tasks (e.g. BERT, GPT3), they also have revealed promising results when trained with different modalities (text, audio, and video) in multimodal emotion recognition [18-23]. In addition, a few studies have utilized transformers in supervised manner to learn spatio-temporal representations from physiological signals [30-31]. BERT inspired transformer based SSL approach was recently introduced in [24] from the perspective of suitability of attention mechanism for processing time-series data. Authors used masked ECG signals for training a transformer model in a pretext task and transferring weights to the downstream stage for emotion recognition. The same approach has been extended in [25] to combine ECG and EEG embeddings for estimating various emotional states, where SSL features of ECG and EEG embeddings are combined in a late fusion manner by simple concatenation method. In contrast with the earlier mentioned fusion approach, we present an approach for combining ECG and EDA during SSL pre-training and leverage the complimentary information across them.

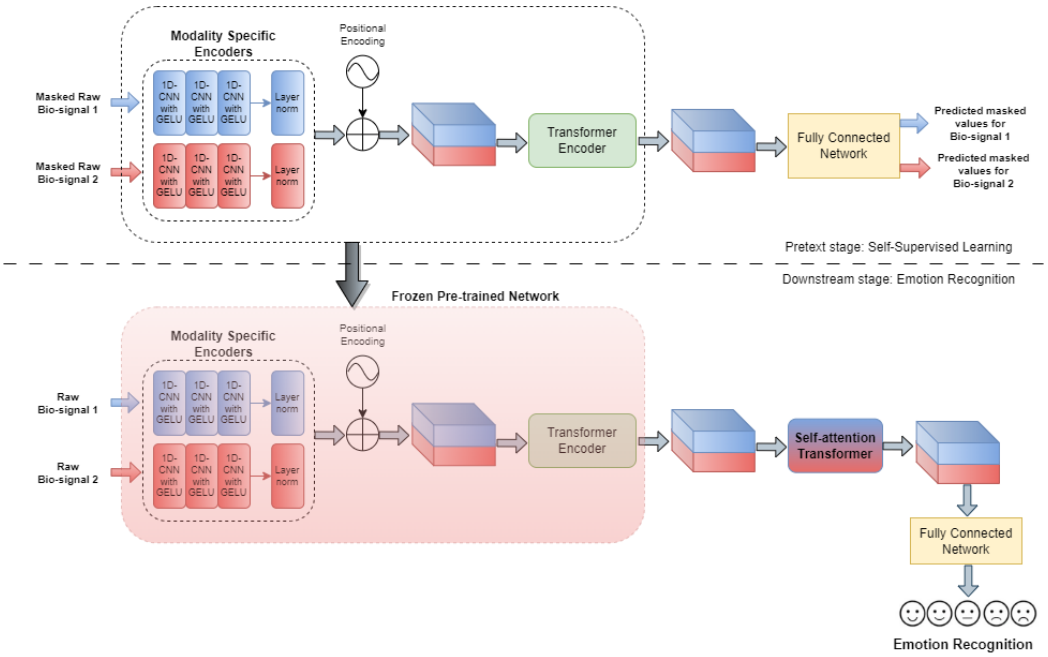

Figure 1: An overview of our proposed MATS2L framework for self-supervised physiological signals representation learning.

3 THE PROPOSED APPROACH

This paper presents a transformer based SSL model for learning multimodal representations from different physiological signals, especially ECG and EDA to estimate different emotions in downstream tasks. The overall framework of the proposed MATS^{2}L is presented in Fig.1. The proposed approach consists of two stages: 1) Pretext stage – Self-supervised learning, where SSL feature representations are obtained through masked signal prediction tasks and 2) Downstream stage – Supervised emotion recognition, where a pre-trained SSL model is used to perform emotion recognition task.

3.1 Pretext Stage: Self-supervised Learning

During the pretext stage, we pre-train the transformer model in a self-supervised manner with the pretext goal of masked signal value prediction. As shown in Figure 1, we first encode physiological signals using modality-specific encoders to accommodate the heterogeneous behavior of different physiological signals.

Each physiological signal is encoded using three layers of 1D-CNN model with GELU activation function [23] independently. Similar to BERT model [23], the obtained feature embeddings are prepended with a special token named as CLS and then concatenated representations are fed into a transformer module with two layers of encoders. In this study, we aim to use modality-agnostic transformer modules to explore commonalities across different physiological signals and lessen the heterogeneity gap between them. The modality-agnostic transformer allows each modality to attend other modality, so ECG signal may be modified strongly by EDA signal or vice versa. As a result, the output of the transformer provides features with strong agreement between ECG and EDA signals. We used a transformer based SSL architecture inspired from [24], as shown in Figure 2. Each transformer encoder consists of a multi-head self-attention module followed by a fully connected feed-forward network with batch normalization, GELU [23] and residual connections. Since physiological signals can be seen in the form of time-series signals, the use of batch normalization can alleviate the effect of outlier values in time-series signals. The proposed model is pre-trained using unlabeled ECG and EDA signals with a similar SSL approach mentioned in [24]. More specifically, we train the transformer model using a masked modelling objective, where some parts of both ECG and EDA signals are masked and must be predicted. We use a masking scheme introduced in [26] instead of random masking, where each masked segment follows a geometric distribution. Random masking method is suitable for short sequences, whereas the geometric masking scheme makes it easier to predict long masked sequences with good approximation. The findings from [26] encourage us to use geometric masking since it helps the model to learn masked time-series segments effectively. To predict the masked part of the signal, a FCN is employed on top of the transformer module. We use mean square error as loss function during the pretext stage.

3.2 Downstream Stage: Supervised Emotion Recognition

In the second stage of the proposed approach, a pre-trained model is transferred to a downstream emotion recognition task. First, SSL embeddings are extracted from a frozen pre-trained model and then fed to a self-attention transformer. Output of the self-attention transformer provides CLS tokens along with the remaining embedding

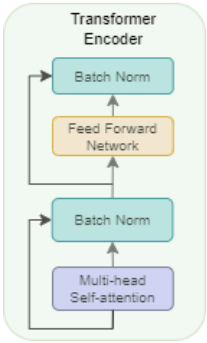

Figure 2: Transformer encoder module.

sequence. Since CLS contains aggregated information from modalities, we use only CLS tokens from a self-attention transformer and pass through a fully connected network to make emotion predictions. The self-attention transformer and fully connected network are trained in a fully supervised manner (using labelled data) to perform the emotion recognition task. We use weighted cross entropy loss for fine-tuning the framework of our downstream stage.

4 EXPERIMENTS

4.1 Dataset and Pre-processing

To assess the effectiveness of the proposed approach, we evaluate it on the public dataset WESAD [27]. This dataset was generated from 15 participants while they were shown video clips intended to elicit emotions or performed tasks such as reading, public speaking, arithmetic tasks, and meditation. WESAD dataset was recorded for studying four affective states such as neutral, amused, stressed, and meditated. In this study, we consider three affective states while ignoring the meditated state like mentioned in [14]. Various physiological signals such as EDA, temperature, BVP, ECG, and EMG were recorded at a sampling frequency of 700 Hz. For this study, we only consider EDA and ECG signals. Instead of performing various complex pre-processing operations on the raw ECG and EDA signals, we tried to keep it simple. This helps to understand the influence of the proposed approach on learning characteristics of the unobstructed raw data input. We perform subject-specific normalization using 'min-max' normalization. ECG and EDA signals are down-sampled to 256 Hz; ECG signal is filtered with a notch filter having frequency of 0.05 Hz, whereas EDA is filtered using Low-pass filter with a cutoff frequency of 3 Hz and a 4th order Butterworth filter. Both ECG and EDA signals are segmented into windows of 10-second segments.

4.2 Implementation Details

For the pre-training task, we used seven out of 15 subjects without any labels from the WESAD dataset and the remaining eight subjects were used in downstream emotion recognition tasks. We made sure to not use the same samples in pre-training and downstream stages. We use the masking scheme introduced in [26] instead of random masking, where each masked segment follows a geometric distribution with masking ratio of 0.5 and a mean length of three.

Table 1: Hyper-parameters setting

	Pre-training task	Downstream task(self-attention transformer)
Number of transformer encoders	2	1
Number of attention heads	4	4
Batch size	256	256
Model dimensions	128	128
Epochs	100	50
Dropout rate	0.1	0.1
Learning rate	5e-5	5e-5

We use weighted cross entropy loss in downstream stage to deal with the class imbalance scenario in WESAD dataset.

The transformer model in the pre-training stage contains two transformer encoders, each containing four self-attention heads with model dimension of 128 and 2 layers of FCN. We use batch normalization instead of layer normalization [26] as originally used for transformers in NLP [28]. The FCN used to predict masked physiological signals is a single layer followed by GELU. We use Adam optimizer in both pretext and downstream stages with a learning rate of 0.00005, an exponential decay rate for the first moment estimates of 0.9, an exponential decay rate for the second moment estimates of 0.999 and a batch size of 256 for both the pre-training and the downstream stage. The models are implemented within the Pytorch framework. The details of hyper-parameters used in the proposed approach are reported in Table 1. We used the same parameter settings throughout our experiments.

5 RESULTS AND DISCUSSIONS

This section discusses the results obtained from the proposed transformer based SSL approach to estimate different emotions in downstream tasks.

5.1 Emotion Recognition Performance of the Proposed Approach

Table 2 reports the performance of the proposed MATS^{2}L approach on the WESAD dataset for classification of affective states. We use mean accuracy and mean F1 score (F1 score between three affective states) over five runs as evaluation metrics. We reported the performance of MATS^{2}L approach under two modes - 1) Frozen: here weights of the pre-trained modality-agnostic transformer are kept frozen in downstream stage and remaining network was fine-tuned with labelled data for classifying affective states. 2) Fine-tuned: in this mode, weights of the pre-trained modality-agnostic transformer are used for initialization and the whole network in the downstream stage was fine-tuned with labelled data for classifying affective states. In order to show the effectiveness of the proposed approach, we reported the performance of modality specific transformers (single modality models) in classifying affective states. In single

Table 2: Emotion recognition model performance. Bold depicts best values.

Pre-training approach	Approach	Modality used	Accuracy	F1
Modality Specific	Single modality	ECG	0.8632 ±0.021	0.8156±0.043
	Single modality	EDA	0.7737 ±0.151	0.6573±0.521
	Concatenation	ECG, EDA	0.9407 ±0.001	0.9301±0.005
Modality - agnostic	MATS^{2}L (Frozen)	ECG, EDA	**0.9644 ±0.510**	**0.9553±0.005**
	MATS^{2}L (Fine-tuned)	ECG, EDA	0.9431 ±0.550	0.9308±0.006

modality experiments, each physiological signal is pre-trained independently. The CLS token extracted from the pre-trained model was fed to the self-attention transformer in the downstream stage. ECG based model shows 86.32% accuracy in recognizing affective states, whereas EDA bio-signal based model achieves 77.37% accuracy in recognizing affective states. Furthermore, we also explored the effect of combining SSL features extracted from modality-specific transformers based SSL. For this, we extracted SSL representations from each pre-trained SSL model, used average pooling across all representations and concatenated them to form a single representation as input to the self-attention transformer in the downstream stage. Concatenation of SSL features obtains an accuracy of 94.07%. As shown in Table 2, under both training modes, our MATS^{2}L approach shows superior performance in classifying three affective states.

In addition, it is observed that the proposed approach achieves highest accuracy of 96.44% under frozen mode over the performance of 94.31% under fine-tuned mode. This was expected due to the fact that, above experiments are performed under low-labeled data regimes. The substantial gap between the performance of concatenation based approach (modality specific) and the proposed MATS^{2}L approach emphasizes the benefits of modality-agnostic pre-training approach. These observations evidently show that our MATS^{2}L approach is able to capture diverse and meaningful features across ECG and EDA.

5.2 Comparison with different emotion recognition approaches

In order to have a fair comparison, we considered three SSL based [14][16][25] and one supervised [12] emotion recognition approaches on the WESAD dataset. We reported results of different approaches in Table 3. For work [12], we reported results directly from respective reference. Although this approach used slightly different experimental settings, we reported it in order to understand the performance of SSL on the WESAD dataset. The work in [12] explored cross-modal attentions among ECG-EDA modalities in supervised fashion. On the other hand, we re-implemented and tested the approach proposed in [16] and [25], with the same modalities and experimental settings we used. For SSL work [14], we used the code provided by authors and experimented with settings we used.

Table 3: Comparison of different approaches on WESAD dataset. Bold depicts best values.

Approach	Modality used	Accuracy	F1
Attx type III [12]	ECG, EDA	0.9208	0.9111
SigRep [16]	ACC, BVP, EDA and ECG	0. 8583	0.8243
SSL work [14]	ECG	0.9457 ±0.130	0.9432 ±0.030
Late fusion using Concatenation [25]	ECG, EDA	0.9547 ±0.010	0.9523 ±0.130
MATS^{2}L (Our approach)	ECG, EDA	**0.9644 ±0.510**	**0.9553 ±0.005**

As shown in Table 3, our MATS^{2}L approach outperforms the other SSL models by a fair margin. Although SigRep [16] uses SSL mechanism and Attx type III [12] focuses on cross-attention mechanism, their approach did not use transformers, which allows us to highlight the effectiveness of the transformer based SSL features. SSL work presented in [14] performs emotion recognition by learning only ECG representations. Furthermore, the approach introduced in [25] employs modality specific transformer approach for pre-training and concatenates the two embeddings in a late fusion manner for estimating the emotions. In contrast, our modality-agnostic transformer focuses on extracting commonalities across both ECG and EDA modalities and provides stronger and more effective ECG-EDA representations. Our approach achieves an accuracy of 96.44% for estimating three affective states of the WESAD dataset, which is ~1% more than that of the concatenation based fusion approach [25]. This indicates that use of a modality-agnostic approach contribute effectively to improving emotion recognition performance.

5.3 Effectiveness of pre-trained SSL features for emotion recognition

To prove the superiority of our approach in learning contextualized features, we tested our framework by training the downstream stage from scratch (random initialization), instead of using pre-trained

Table 4: Fully supervised vs self-supervised. Bold depicts best values.

Pre-train	Accuracy	F1
No (Fully supervised)	0.9373±0.086	0.9240±0.001
Yes (Proposed MATS^{2}L)	**0.9644±0.510**	**0.9553±0.005**

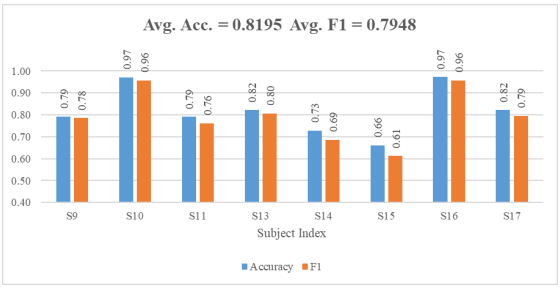

Figure 3: Subject independent test results showing the classification accuracy of each subject for affective states.

weights of a modality-agnostic transformer. Results reported in Table 4 highlight the benefits of pre-trained SSL features in improving the emotion recognition performance. Table 4 indicates that a model without pre-training (fully supervised model) achieves 93.73% accuracy in classifying different emotions, whereas model with pre-training (self-supervised) reaches almost 96.44% of accuracy. This clearly depicts the use of pre-trained SSL features in improving the emotion recognition performance by fine margin. Furthermore, we also observed that the model without pre-training tends to overfit quickly. These observations demonstrate that our transformer based SSL model is highly efficient for emotion recognition in comparison to the same model when trained in a fully supervised manner.

5.4 Subject Independent Classification

To demonstrate the influence of interpersonal variations in the proposed MATS^{2}L approach, subject independent experiments were conducted for classification of affective states. The subject independent setting is very similar to using new subject's data as test data in a real world scenario. For this purpose, Leave-one-subject-out (LOSO) cross validation scheme was implemented in the downstream stage, where the physiological data of one subject are used for testing and the physiological data of remaining subjects are used for fine-tuning. Note that we made sure to not use the same subjects during pretext and downstream stages. We used data from eight subjects out of 15 subjects of the WESAD dataset during the downstream stage. These experiments were performed to further challenge the proposed MATS^{2}L approach in the LOSO setting. Figure 3 shows a plot of classification accuracies for each subject tested using the LOSO scheme. It is observed that the MATS^{2}L approach performs satisfactorily in subject independent evaluation.

Table 5: Ablation study on segment length. Bold depicts best values.

Segment length	Accuracy	F1
10	**0.9644±0.510**	**0.9553±0.005**
30	0.8050±0.684	0.7226±0.035
40	0.7029±0.536	0.5678±0.050

Table 6: Performance of different model variations of MATS^{2}L by varying embedding dimensions. Bold depicts best values.

Embedding dimension	Accuracy	F1
64	0.9193±0.051	0.9000±0.004
128	**0.9644±0.510**	**0.9553±0.005**
256	0.9434±0.021	0.9361±0.049

Table 7: Performance of different model variations of MATS^{2}L by varying encoder blocks. Bold depicts best values.

Encoder blocks	Accuracy	F1
1	0.9054±1.011	0.8676±0.015
2	0.9360±0.600	0.9177±0.007
3	**0.9644±0.510**	**0.9553±0.005**

5.5 Ablation Study

We performed an ablation study to understand the influence of different components on the proposed MATS^{2}L approach.

5.5.1 Segment Length. This ablation study was conducted to identify the most suitable segment length to divide physiological signal into input sequences. We compared three different segment lengths {10, 30 and 40} second segment. As reported in Table 5, shorter segment provides highest accuracy in recognizing affective states. Shorter segments enhance the emotion recognition performance by an extensive margin. We suspect that longer segments need more complex transformer models, which are difficult to train with limited labeled data of WESAD.

5.5.2 Ablation study on different model variations of MATS^{2}L by varying embedding dimensions. This ablation study was conducted to identify the most suitable embedding dimension to achieve the optimal performance with the proposed model. We tested our model with varying embedding dimensions of {64, 128 and 256} and reported the results in Table 6. As shown in Table 6, the model with embedding dimensions of 128 achieved best performance.

5.5.3 Ablation study on different model variations of MATS^{2}L by varying blocks in 1DCNN encoder. This ablation study was conducted to identify the most suitable number of blocks for 1DCNN encoder to achieve best performance with the proposed model. We tested our model with varying encoder blocks of {1, 2 and 3} and reported the results in Table 7.

As shown in Table 7, we observed that our model achieves highest performance with three 1DCNN encoders. This analysis helps us to understand that the embeddings obtained with deeper encoder architecture are generalizable and enhance the performance of downstream emotion recognition tasks.

6 CONCLUSIONS AND FUTURE WORK

In this study, we presented a modality-agnostic transformer based self-supervised learning for emotion recognition using physiological signals, ECG and EDA. Experimental analysis on the WESAD dataset showed that the proposed approach could efficiently recognize different affective states. Additionally, it demonstrated that the proposed approach could compete with and outperform fully supervised learning approaches. This confirms the prospective of self-supervision to capture significant signal attributes in the absence of labeled data. Moreover, our modality-agnostic approach encourages the model to capture commonalities across different physiological signals.

Although we only focused on ECG and EDA signals in this study, it would be interesting to explore other physiological signals like, EMG or EEG signals in future work. During pre-training, we trained a modality-agnostic transformer by concatenating signal embeddings; however, it would be interesting to explore cross-attention based modality-agnostic pre-training that can capture cross-modal interactions and provide a new perspective for performing feature fusion.

7 ETHICAL STATEMENT

While the study presented in this paper provides insights about the influence of self-supervision in estimating emotions effectively, the observations from the study can be used in designing intelligent emotion recognition systems. Despite use of the public dataset in validating the proposed approach, various ethical considerations need to be considered. First, we have not considered any sensitive data that might reveal participant's identity. Since our method is based on de-identified signals, it can be applied without need of any sensitive data. Furthermore, small sized dataset might affect the generalizability of the studied model. Finally, in order to support reproducibility of our method, we have provided the details of experimental settings used in Section IV. The hyper-parameters were determined through several tests. We are aware about various issues with the reproducibility such as, dependency of the results on the type of physiological signals, size of dataset, data-loaders, training-validation procedures, choice of loss functions, and many more.

ACKNOWLEDGMENTS

This work was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the Center for Analytics – Data – Applications (ADA-Center) within the framework of "BAYERN DIGITAL II" (20-3410-2-9-8).

REFERENCES

[1] R. Assabumrungrat *et al.* 2021. Ubiquitous Affective Computing: A Review. IEEE Sens. J., vol. 22, no. 3, pp. 1867–1881, 2021, doi: 10.1109/JSEN.2021.3138269
[2] S. Poria, E. Cambria, A. Hussain, and G. Bin Huang. 2015. Towards an intelligent framework for multimodal affective data analysis. Neural Networks, vol. 63, pp. 104–116, 2015, doi: 10.1016/j.neunet.2014.10.005.
[3] M. Spezialetti, G. Placidi, and S. Rossi. 2020. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. Front. Robot. AI, vol. 7, no. December, pp. 1–11, 2020, doi: 10.3389/frobt.2020.532279
[4] S. Lee, D. K. Han, and H. Ko. 2021. Multimodal Emotion Recognition Fusion Analysis Adapting BERT with Heterogeneous Feature Unification. IEEE Access, vol. 9, pp. 94557–94572, 2021, doi: 10.1109/ACCESS.2021.3092735
[5] M. Yeasin, B. Bullot, and R. Sharma. 2006. Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. Multimed., vol. 8, no. 3, pp. 500–507, 2006, doi: 10.1109/TMM.2006.870737.
[6] C. Torres-Valencia, M. Álvarez-López, and Á. Orozco-Gutiérrez. 2017. SVM-based feature selection methods for emotion recognition from multimodal data. J. Multimodal User Interfaces, vol. 11, no. 1, pp. 9–23, 2017, doi: 10.1007/s12193-016-0222-y
[7] M. S. Zitouni, C. Y. Park, U. Lee, L. J. Hadjileontiadis, and A. Khandoker. 2022. LSTM-Modeling of Emotion Recognition Using Peripheral Physiological Signals in Naturalistic Conversations. IEEE J. Biomed. Heal. Informatics, vol. 27, no. 2, pp. 1–14, 2022, doi: 10.1109/jbhi.2022.3225330
[8] M. Khateeb, S. M. Anwar, and M. Alnowami. 2021. Multi-Domain Feature Fusion for Emotion Classification Using DEAP Dataset. IEEE Access, vol. 9, pp. 12134–12142, 2021, doi: 10.1109/ACCESS.2021.3051281
[9] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran. 2020. Automatic Emotion Recognition Using Temporal Multimodal Deep Learning. IEEE Access, vol. 8, pp. 225463–225474, 2020, doi: 10.1109/ACCESS.2020.3027026
[10] S. Siddharth, T.-P. Jung, and T. J. Sejnowski. 2019. Utilizing Deep Learning Towards Multi-modal Bio-sensing and Vision-based Affective Computing. IEEE Trans. Affect. Comput., vol. 3045, no. c, pp. 1–1, 2019, doi: 10.1109/taffc.2019.2916015
[11] K. Ross, P. Hungler, and A. Etemad. 2021. Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data. J. Ambient Intell. Humaniz. Comput., no. 0123456789, 2021, doi: 10.1007/s12652-021-03462-9
[12] A. Bhatti, B. Behinaein, D. Rodenburg, P. Hungler, and A. Etemad. 2021. Attentive Cross-modal Connections for Deep Multimodal Wearable-based Emotion Recognition. 2021 9th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos, ACIIW 2021, 2021, doi: 10.1109/ACIIW52867.2021.9666360.
[13] A. F. Bulagang, N. G. Weng, J. Mountstephens, and J. Teo. 2020. A review of recent approaches for emotion classification using electrocardiography and electrodermography signals. Informatics Med. Unlocked, vol. 20, p. 100363, 2020, doi: 10.1016/j.imu.2020.100363.
[14] P. Sarkar and A. Etemad. 2020. Self-supervised ECG Representation Learning for Emotion Recognition. IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1541–1554, 2020, doi: 10.1109/TAFFC.2020.3014842.
[15] K. G. Montero Quispe, D. M. S. Utyiama, E. M. dos Santos, H. A. B. F. Oliveira, and E. J. P. Souto. 2022. Applying Self-Supervised Representation Learning for Emotion Recognition Using Physiological Signals. Sensors, vol. 22, no. 23, 2022.
[16] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara. 2022. SigRep: Toward Robust Wearable Emotion Recognition with Contrastive Representation Learning. IEEE Access, vol. 10, pp. 18105–18120, 2022, doi: 10.1109/ACCESS.2022.3149509.
[17] Z. Zhang, S. hua Zhong, and Y. Liu. 2022. GANSER: A Self-supervised Data Augmentation Framework for EEG-based Emotion Recognition. IEEE Trans. Affect. Comput., vol. XX, no. XX, pp. 1–17, 2022, doi: 10.1109/TAFFC.2022.3170369.
[18] H. Akbari *et al.* 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. Adv. Neural Inf. Process. Syst., vol. 29, no. NeurIPS, pp. 24206–24221, 2021.
[19] P. Xu, X. Zhu, and D. A. Clifton. 2022. Multimodal Learning with Transformers: A Survey. pp. 1–23, 2022, [Online]. Available: http://arxiv.org/abs/2206.06488.
[20] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara. 2020. Multimodal emotion recognition with transformer-based self supervised feature fusion. IEEE Access, vol. 8, pp. 176274–176285, 2020, doi: 10.1109/ACCESS.2020.3026823.
[21] J. Tang *et al.* 2022. MMT: Multi-Way Multi-Modal Transformer for Multimodal Learning. IJCAI Int. Jt. Conf. Artif. Intell., pp. 3458–3465, 2022, doi: 10.24963/ijcai.2022/480.
[22] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara. 2020. Jointly fine-tuning 'BERT-like' self supervised models to improve multimodal speech emotion recognition. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-Octob, pp. 3755–3759, 2020, doi: 10.21437/Interspeech.2020-1212.
[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.
[24] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley. 2022. Transformer-based self-supervised learning for emotion recognition. In 26th International Conference on Pattern Recognition (ICPR), pp. 2605–2612, IEEE, 2022.
[25] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley. 2022. Emotion Recognition with Pre-Trained Transformers Using Multimodal Signals. 2022 10th Int. Conf. Affect. Comput. Intell. Interact. ACII 2022, 2022, doi: 10.1109/ACII55700.2022.9953852.

[26] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff. 2021. A Transformer-based Framework for Multivariate Time Series Representation Learning. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 2114–2124, 2021, doi: 10.1145/3447548.3467401.

[27] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven. 2018. Introducing WeSAD, a multimodal dataset for wearable stress and affect detection. ICMI 2018 - Proc. 2018 Int. Conf. Multimodal Interact., pp. 400–408, 2018, doi: 10.1145/3242969.3242985

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017

[29] X. Shen, X. Liu, X. Hu, and D. Zhang, 2022. Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition. IEEE Transactions on Affective Computing

[30] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad. 2021. A Transformer Architecture for Stress Detection from ECG. In 2021 International Symposium on Wearable Computers (ISWC '21), September 21–26, 2021, Virtual, USA. ACM, New York, NY, USA

[31] Ruiqi Wang and Wonse Jo and Dezhong Zhao and Weizheng Wang and Baijian Yang and Guohua Chen and Byung-Cheol Min, 2023. Husformer: A Multi-Modal Transformer for Multi-Modal Human State Recognition, arXiv.