2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871421

Multimodal Neurophysiological Transformer for Emotion Recognition

Sharath Koorathota[1], Zain Khan[1], Pawan Lapborisuth[1] and Paul Sajda[1,2]

Abstract— **Understanding neural function often requires multiple modalities of data, including electrophysiogical data, imaging techniques, and demographic surveys. In this paper, we introduce a novel neurophysiological model to tackle major challenges in modeling multimodal data. First, we avoid non-alignment issues between raw signals and extracted, frequency-domain features by addressing the issue of variable sampling rates. Second, we encode modalities through "cross-attention" with other modalities. Lastly, we utilize properties of our parent transformer architecture to model long-range dependencies between segments across modalities and assess intermediary weights to better understand how source signals affect prediction. We apply our Multimodal Neurophysiological Transformer (MNT) to predict valence and arousal in an existing open-source dataset. Experiments on non-aligned multimodal time-series show that our model performs similarly and, in some cases, outperforms existing methods in classification tasks. In addition, qualitative analysis suggests that MNT is able to model neural influences on autonomic activity in predicting arousal. Our architecture has the potential to be fine-tuned to a variety of downstream tasks, including for BCI systems.**

I. Introduction

Recent work in emotional state detection, helpful in prediction of clinical outcomes [1], [2], has led to the notion of a brain-body system with complex interactions between different components [3], [4]. How the brain generates states of psychological stress [5], [6], which may lead to physical disease [7], [8] or increased disease vulnerability [9], [10], is a question of growing interest. The real-time integration of neuroimaging and autonomic methods with subjective measurements of emotion, such as arousal or valence, is providing new opportunities to expand our knowledge on the brain-body system to design real-time monitoring systems capable of analysis of large volumes of multimodal data.

Recent research in emotion recognition has utilized a variety of sources. These include external behavioral signals such as posture [11], facial expression [12], speech [13] and environmental factors [14]. Internal neurophysiological sources such as electroencephalography (EEG), galvanic skin response (GSR), functional magnetic resonance imaging (fMRI), and photoplethysmogram (PPG) have also been used [15]–[17]. However, mostly driven by the rapid advancements in computer vision and natural language processing, multimodal research has mainly focused on the fusion of visual and audio signals [18]. Only recently have neurophysiological researchers adapted state-of-the-art architectures to understand brain function [19], [20].

EEG is the most commonly used non-invasive neural measure, yielding not only accurate predictions of emotional state but in furthering our understanding of functional relationships between brain regions associated with emotion [21]. Most existing emotion recognition and brain computer interface (BCI) studies utilizing EEG have dealt with the density of data through extracting features channel-wise from different brain regions and using these extracted features to classify emotion. While this method is apt to deal with noise and provides a level of interpretability to the results, extensive pre-processing and feature extraction removes potentially useful information from raw signals [22], [23]. Furthermore, critical information, such as decreased signal complexity in patients with emotional processing challenges, are unable to be captured with common pre-processing and modeling approaches [24]. A similar finding has been shown in cardiac signal processing, where end-to-end, raw electrocardiogram (ECG) feature extraction, in comparison with template-based feature matching, yields higher sensitivity in the detection of ectopic beats useful for emotional state recognition [25]. However, few studies have have made use of raw signal data in conjunction with extracted features for emotion recognition. Only recently have deep learning architectures allowed the efficient processing of multivariate and multimodal neurophysiological data [26], [27].

Advancements in multimodal sequence modeling have come from the area of computer vision and language processing. Specifically, transformer networks introduced by Vaswani et al [28] leveraging self-attention have not only successfully been applied for improving accuracy in language translation, video captioning, and learning sentence representations, but have found success in multimodal settings. Attention is a critical component of transformers, where multiple, independent "heads" gather relevant information from the input vectors. In self-attention, for each query representation $\boldsymbol{q}(\boldsymbol{x}_i)$ of input vector $\boldsymbol{x}_i$ at timepoint i, we compute an output vector $\boldsymbol{y}_i \in \mathbb{R}^d$ using a sequence of input vectors $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^d$ as:

$$\boldsymbol{y}_i = \sum_{j=1}^{n} \alpha_{i,j} \boldsymbol{v}(\boldsymbol{x}_j) \tag{1}$$

$$\alpha_{i,j} := \mathrm{softmax}_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}} \left(\frac{\boldsymbol{q}(\boldsymbol{x}_i) \boldsymbol{k}(\boldsymbol{x}_j)^\top}{\sqrt{d'}} \right) \in \mathbb{R} \tag{2}$$

After key, $\boldsymbol{k}$, lookups of $\boldsymbol{x}_j$, attention gathers value vectors $\boldsymbol{v}(\boldsymbol{x}_j)$ using weights $\alpha_{i,j}$ prior to applying additional transformations on $\boldsymbol{y}_i$.

Past work related to interpreting transformers has primarily focused on qualitative interpretation of attention weights $\alpha_{i,j}$. A major critique of this approach is that analysis of

[1]Department of Biomedical Engineering, Columbia University
[2]Department of Electrical Engineering, Columbia University

weights relies on researchers selecting inputs and layers, making interpretation results hard to replicate. In our work, we study differences in intermediary network activations $\alpha_{i,j}$ with underlying emotional state to understand the relationship between the autonomic and neural systems.

While the original transformer architecture has been used for unidirectional translation from source to target texts, Tsai et al [29] were the first to show how human multimodal language time-series can also be well-represented using similar architectures through their MulT architecture. These data sources, similar to physiological data, have sequences of each modality with different frequencies. For a given list of modalities, $M = \{\alpha^1, ..., \alpha^n\}$, consider the input sequences of data from α^p as $M_{\alpha^p} \in \mathbb{R}^{T_{\alpha^p} \times d_{\alpha^p}}$, where T_{α^p} denotes the sequence length and d_{α^p} denotes the feature dimension for modality α^p. The sequence undergoes a one-dimensional convolution and is projected onto a common, predefined dimension. In this way, dimensionality across all modalities can be standardized, prior to attention modeling. Following this, a positional embedding is augmented to allow the sequences to capture temporal feature information. The sequences then enter a crossmodal transformer which consists of crossmodal attention blocks μ_M. A crossmodal attention block is a simple re-structuring of attention specified in Equation 2, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are derived from different modalities:

$$\mu_{i,j}^{p \to r} := \text{softmax}_{\boldsymbol{x}_i^r, \boldsymbol{x}_j^p \in \hat{M}} \left(\frac{\boldsymbol{q}^r \left(\boldsymbol{x}_i^r \right) \boldsymbol{k}^p \left(\boldsymbol{x}_j^p \right)^\top}{\sqrt{d'}} \right) \in \mathbb{R} \quad (3)$$

where the adaptation from modality α^p to α^r at respective timepoints j and i requires the transformation of vector $\boldsymbol{x}_i^r$ using a r-specific query function $\boldsymbol{q}^r(\boldsymbol{x}_i)$ and p-specific key function $\boldsymbol{k}^p(\boldsymbol{x}_j)$. In this way, $\boldsymbol{y}_i^r$ represents the latent, adapted vector resulting from crossmodal attention of modality-specific sequences i and j. The output of the crossmodal transformer is concatenated if it shares the same target modality and passes through a self-attention model prior to classification, such as for binary classification of valence or arousal emotional states.

In this work, we investigate corollary information in multimodal neurophysiological data and evaluate the performance of a transformer architecture for emotion recognition in a public dataset. We propose a multimodal neurophysiological transformer adapted for sequential modeling of EEG, PPG and GSR - for both pre-processed time series and extracted features. Our framework is advantageous in that sequences are processed in one-go rather than in-order. We demonstrate competitive performance for multimodal emotion recognition, while taking into account information from all modalities. We present a novel interpretation method for attention-based models, through assessing state-driven activations, and demonstrate how latent, cardiac and autonomic data is transformed by neural data prior to classification.

II. Methods

Our transformer model (MNT, Figure 1) is adapted from the MulT architecture to allow modality-specific convolutions and class-specific activations.

A. Adaptable convolutions

We distinguish not only between modalities (e.g. EEG, PPG) but also modality types (i.e. time series, extracted features). This abstraction allows our model to consider each modality type differently in the initial convolution using:

$$\hat{M}\left(\alpha^p, \mathcal{M}\right) = \text{Conv1D}\left(M_{\alpha^p}, k_{\mathcal{M}}, s_{\mathcal{M}}\right) \in \mathbb{R}^{T_{\mathcal{M}} \times d} \quad (4)$$

where $\hat{M}$ is the convolved or down-sampled input sequence used in the transformer. $\hat{M}$ is a function of the modality α^p and modality type $\mathcal{M}$. $\mathcal{M} \in \{\text{time series}, \text{features}\}$ is used to determine the convolution hyperparameters. In this way, we control the amount of data loss to be minimal for extracted features (which may not be sequentially ordered) and possibly high for time series data with large sample rate (such as from an EEG device) which faces computational memory limits during training. While multimodal computer vision or language processing sequence lengths are typically determined by the number of words or video frames sampled in short clips, finding meaningful neurophysiological signatures requires analysis of data at higher sample rates and effective use of features extracted from time series data, which our parametric $\hat{M}$ allows.

B. Assessing interactions (SAAD)

The increasing interest in using transformers for prediction is partly driven by qualitative interpretations of attention weights. Figure 1f shows example attention weights or activations from a single, training sample at one crossmodal block. Convolved data from the first EEG window (j) is used as the query and is looked up against keys from PPG windows (i) for scoring. The intensity of the attention map (f) represents the attention that the EEG window pays to the PPG window at $(i = 4, j = 1)$. In most analyses, researchers have used individual examples to highlight how the network attends to words, images or video similarly to humans. We take a novel approach to assessing interactions between neural and autonomic data, through (1) computing class-level differences in activations, sample-by-sample. Specifically, we compute the latent contribution of modality p to r using:

$$\phi_{i,j}^{p \to r} := \frac{1}{H} \sum_{h=1}^{H} \mu_{i,j}^{p \to r} \quad \forall_s \in S \quad (5)$$

$$\boldsymbol{\delta}_{i,j}^{p \to r} = \frac{1}{L} \sum_{l=1}^{L} \left| \frac{1}{S_+} \sum_{s=1}^{S_+} \phi_{i,j}^{p \to r} \right| - \left| \frac{1}{S_\cdot} \sum_{s=1}^{S_\cdot} \phi_{i,j}^{p \to r} \right| \quad \forall_{p,r} \in M \quad (6)$$

where, given a trained model, we average attention weights across heads H to compute the sample-specific, within-layer attention weight $\phi_{i,j}^{p \to r}$. The sum of absolute activation differences (SAAD) metric is then computed for each modality

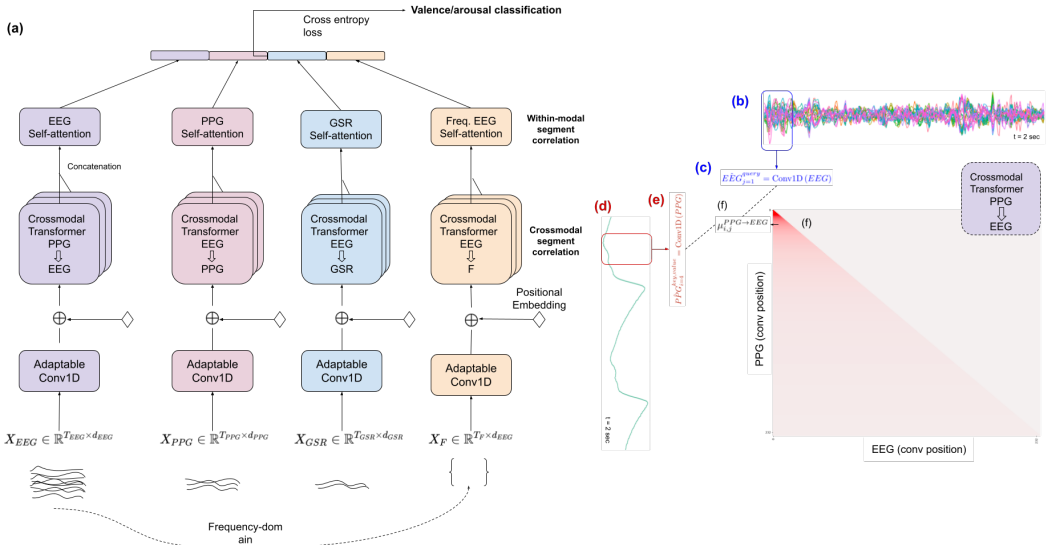

Fig. 1: (a) Multimodal Neurophysiological Transformer (MNT) architecture. (b) Attention weights or activations (c) from a single, training sample at the crossmodal block $PPG \rightarrow EEG$. Each sample uses data from 2 seconds of EEG (b) and PPG (d), with the convolution parameters determining the size of the activation map. The adaptable convolution (c, e) standardizes the dimensionality of both modalities, and the score-lookup (f) is a standard implementation of softmax-attention.

pair $p \rightarrow r$ using the average differences in activations $\boldsymbol{\delta}$ between the positive and negative classes, across all layers L, sequence indices i, j and normalized for all $\{p, r\} \in M$.

C. Dataset

The DEAP dataset was first introduced in 2018 [13]. EEG signals and peripheral physiological signals of 32 participants were recorded as they watched music videos selected to induce emotional response. We used 32-channel EEG signals capturing neural response and single-channel PPG and GSR signals capturing autonomic response. Each participant (n=32) completed 40 trials with each trial (i.e. music video) spanning 63 seconds. Subjects were asked to self-assess intensity of five different emotional states - valence, arousal, dominance, liking, and familiarity - using a range from 1 to 9.

In order to best compare the performance of our proposed method with previous results, we only used valence and arousal states in our analyses and split each trial into windows of overlapping samples prior to analysis. We divided the samples into two different classes using a threshold of 5 to allow binary classification of high or low valence, and binary classification of high or low arousal. 70% samples were used as training data, 15% as validation data and the rest were used as test data.

In addition to pre-processed time-series, we were interested in using global EEG features characterizing signal variance, complexity and frequency components. We computed a total of 24 features for each of the 32 channels: θ-, α-, β-, γ- band powers, band-specific Hjorth activity, mobility and complexity, band-specific HFD, and band-specific sample entropy - markers previously shown to improve classification of arousal and valence [30].

Method	Valence (%)	Arousal (%)	Sample Windows
Xing et al [31]	81.1	74.38	60
Rozgic et al [32]	76.9	69.1	60
Li et al [33]	58.4	64.3	60
Features MNT	77.4	76.2	60
Features MNT	70.5	71.8	27
EEG MNT	67.3	68.9	27
MNT	58.0	69.4	27

TABLE I: Model performances (binary classification accuracy) with ablation. MNT achieves similar accuracy with fewer samples. MNT accuracy is assessed using similar data augmentation (sample windows) as other methods.

III. Results

A. Emotional state classification

While several studies have evaluated the DEAP dataset, we note that many windowed samples into shorter intervals (e.g. 0.5-second windows or 1-second windows with 0.5 second overlap) or used features extracted from more modalities than our study, such as respiration and eye movement. We compare MNT to studies using EEG data and find that binary classification accuracy matches or exceeds existing results (Table I).

Our results show that using features extracted from EEG yielded the greatest accuracy. For subsequent analysis of emotion-dependant activations, we use MNT with 27 windows (referred to generally as MNT) since it allows the greatest amount of study in crossmodal adaptations of latent representations and the number of training samples fit within reasonable memory limitations (128 GB RAM, K80 GPU, 10 epochs early stopping criteria) for subsequent analyses.

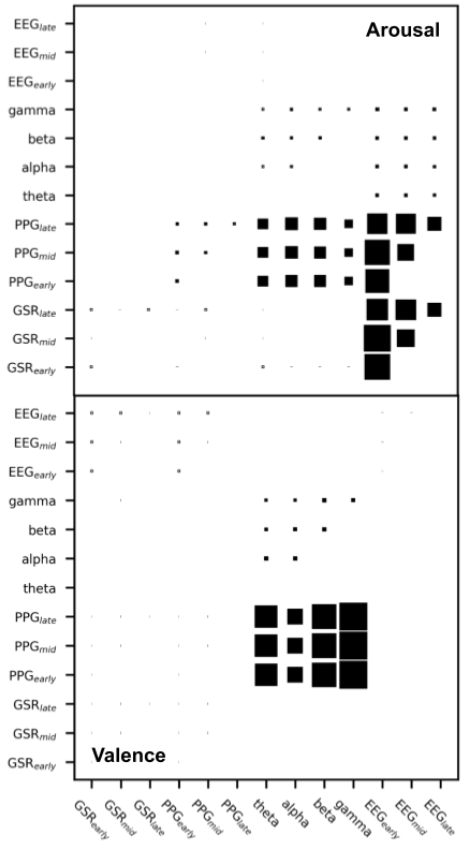

Fig. 2: Most discriminative interactions, as assessed through activation differences (SAAD) in the trained arousal (top) and valence (bottom) models. Early refers to the first third of the sample duration (2 seconds), mid and late refer to the second and last third respectively. For example, $EEG_{early} \rightarrow PPG_{early}$ indicates the aggregated, differences (between high and low arousal samples) in crossmodal attention by pre-processed PPG at the start of the sample (query) to pre-processed EEG at the start of a sample (key, value). A larger surface area for a measure captures the relatively greater contribution of the interaction towards discriminating between high and low arousal or valence.

B. Class-level activation differences

Our primary aim was to use attention in MNT to study interactions between neural and autonomic signals. Although researchers have previously qualitatively reported attention weights across a multi-layer transformers [34]–[36], our approach formalized in Equation 6 aggregates meaningful differences in class-level activations (e.g. high vs. low arousal) across the entire training dataset to understand whether EEG latently adapts PPG and GSR signals. Through assessing class-level differences in activations $\boldsymbol{\delta}_{i,j}^{p \rightarrow r}$, we found several interactions with non-zero differences.

Our results suggest that the largest contributor to model performance is the relationship between features extracted from EEG and PPG signal. In other words, MNT leverages both neural and autonomic signals, through latent adaptation of the autonomic signal, to classify emotional state. The directionality of this adaptation suggests that global, extracted, features are more easily able to adapt time series modalities that may face a relatively greater number of deviations due to noise or movement. While we expected to see the largest interaction between neural and autonomic systems, we were surprised to see the generally large reliance on features extracted from EEG rather than the source signal itself. We expected that modeling EEG time series through a transformer network would extract some characteristics of signal variance, entropy and possibly frequency-domain qualities. However, our finding agrees with recent results showcasing weaknesses of transformers in capturing global, frequency-domain characteristics of source signals without additional processing [37].

While we did not find existing literature on the interaction between EEG and PPG over time, we were able to verify that MNT, through class-level activation differences in the self-attention of EEG features, affirms commonly reported relationships between EEG activity, autonomic activity, arousal and valence. For example, increase in alpha and beta EEG activity were accompanied by increases in muscle artifacts, and heart rate acceleration [38]. We show normalized class-level activation differences through SAAD in Figure 2. The classification power for beta power for prediction of continuous valence and arousal measures have been previously reported [30] and found in our results. We found more consistent literature on the effect of images and video on the modulation of gamma power for valence measures than arousal, which our findings validate [39], [40].

C. Limitations

A major limitation of our approach in using activation differences at the attention heads and across layers is the lack of spatial resolution for EEG analysis. Because our initial convolution block reduces the channels of EEG, along with other input signals, we are unable to investigate commonly-reported measures of EEG activity modulated by image and video stimuli such as hemispheric asymmetry after the initial convolution step. However, future work can use the invertible nature of the convolution step to explore reconstruction of input signals, which may allow class-specific reconstruction of EEG activity to allow spatial understanding. While our normalized measures are a step towards understanding the total contribution of a feature towards classification power, the relative importance is more difficult to judge without extensive analysis on the interaction between features.

IV. Conclusion

We proposed a multimodal neurophysiological transformer adapted for sequential modeling of EEG, PPG and GSR - for both raw time series and extracted features. We demonstrate competitive, interpretable performance for multimodal emotion recognition, while taking into account information from all modalities and demonstrate how latent, cardiac and autonomic data is transformed by neural data prior to classification.

ACKNOWLEDGEMENT

This work was supported by grants from the National Science Foundation (OIA-1934968), the Army Research Laboratory (W911NF-21-2-0125) and a Vannevar Bush Faculty Fellowship from the US Department of Defense (N00014-20-1-2027).

REFERENCES

[1] T. Tattan and N. Tarrier, "The expressed emotion of case managers of the seriously mentally ill: The influence of expressed emotion on clinical outcomes," *Psychological medicine*, vol. 30, no. 1, pp. 195–204, 2000.

[2] A. Oldershaw, T. Lavender, and U. Schmidt, "Are socio-emotional and neurocognitive functioning predictors of therapeutic outcomes for adults with anorexia nervosa?" *European Eating Disorders Review*, vol. 26, no. 4, pp. 346–359, 2018.

[3] A. Babayan, M. Erbey, D. Kumral, J. D. Reinelt, A. M. Reiter, J. Röbbig, H. L. Schaare, M. Uhlig, A. Anwander, P.-L. Bazin *et al.*, "A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults," *Scientific data*, vol. 6, no. 1, pp. 1–21, 2019.

[4] J. LeDoux, "The emotional brain, fear, and the amygdala," *Cellular and molecular neurobiology*, vol. 23, no. 4, pp. 727–738, 2003.

[5] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state." *Psychological review*, vol. 69, no. 5, p. 379, 1962.

[6] A. J. Carson, S. MacHale, K. Allen, S. M. Lawrie, M. Dennis, A. House, and M. Sharpe, "Depression after stroke and lesion location: a systematic review," *The Lancet*, vol. 356, no. 9224, pp. 122–126, 2000.

[7] B. S. Jonas and M. E. Mussolino, "Symptoms of depression as a prospective risk factor for stroke," *Psychosomatic medicine*, vol. 62, no. 4, pp. 463–471, 2000.

[8] L. D. Kubzansky and I. Kawachi, "Going to the heart of the matter: do negative emotions cause coronary heart disease?" *Journal of psychosomatic research*, vol. 48, no. 4-5, pp. 323–337, 2000.

[9] A. D. Ong, "Pathways linking positive emotion and health in later life," *Current Directions in Psychological Science*, vol. 19, no. 6, pp. 358–362, 2010.

[10] S. Cohen, D. Janicki-Deverts, and G. E. Miller, "Psychological stress and disease," *Jama*, vol. 298, no. 14, pp. 1685–1687, 2007.

[11] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[12] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018. [Online]. Available: https://doi.org/10.1016/j.patrec.2018.04.010

[13] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2010.09.020

[14] R. Zatarain Cabada, M. L. Barrón Estrada, J. M. Ríos Félix, and G. Alor Hernández, "A virtual environment for learning computer coding using gamification and emotion recognition," *Interactive Learning Environments*, vol. 28, no. 8, pp. 1048–1063, 2020. [Online]. Available: https://doi.org/10.1080/10494820.2018.1558256

[15] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical Signal Processing and Control*, vol. 55, p. 101646, 2020. [Online]. Available: https://doi.org/10.1016/j.bspc.2019.101646

[16] D. P.-o. Bos and D. O. Bos, "EEG-based emotion recognition."

[17] P. Das, A. Khasnobish, and D. N. Tibarewala, "Emotion recognition employing ECG and GSR signals as markers of ANS," *Conference on Advances in Signal Processing, CASP 2016*, pp. 37–42, 2016.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 689–696, 2011.

[19] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," *Proceedings - International Conference on Pattern Recognition*, vol. 1, pp. 1136–1139, 2006.

[20] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2018.

[21] A. Goshvarpour, A. Abbasi, and A. Goshvarpour, "An accurate emotion recognition system using ECG and GSR signals and matching pursuit method," *Biomedical Journal*, vol. 40, no. 6, pp. 355–368, 2017. [Online]. Available: https://doi.org/10.1016/j.bj.2017.11.001

[22] R. Salazar-Varas and R. A. Vazquez, "Evaluating the effect of the cutoff frequencies during the pre-processing stage of motor imagery EEG signals classification," *Biomedical Signal Processing and Control*, vol. 54, p. 101592, 2019. [Online]. Available: https://doi.org/10.1016/j.bspc.2019.101592

[23] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K. M. Su, and K. A. Robbins, "The PREP pipeline: Standardized preprocessing for large-scale EEG analysis," *Frontiers in Neuroinformatics*, vol. 9, no. JUNE, pp. 1–19, 2015.

[24] J. Dauwels, F. Vialatte, and A. Cichocki, "Diagnosis of Alzheimers Disease from EEG Signals: Where Are We Standing?" *Current Alzheimer Research*, vol. 7, no. 6, pp. 487–505, 2010.

[25] S. S. Xu, M. W. Mak, and C. C. Cheung, "Towards End-to-End ECG Classification with Raw Signal Extraction and Deep Neural Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1574–1584, 2019.

[26] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[27] K. Samiee, P. Kovács, and M. Gabbouj, "Epileptic seizure classification of EEG time-series using rational discrete short-time fourier transform," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 541–552, 2015.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[29] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Zico Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 6558–6569, 2020.

[30] F. Galvão, S. M. Alarcão, and M. J. Fonseca, "Predicting exact valence and arousal values from EEG," *Sensors*, vol. 21, no. 10, 2021.

[31] X. Xing, Z. Li, T. Xu, L. Shu, B. Hu, and X. Xu, "SAE+LSTM: A new framework for emotion recognition from multi-channel EEG," *Frontiers in Neurorobotics*, vol. 13, no. June, pp. 1–14, 2019.

[32] V. Rozgić, S. Vitaladevuni, and R. Prasad, "Robust EEG emotion classification using segment level decision fusion," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1286–1290, 2013.

[33] M. Li and B.-l. Lu, "Emotion Classification Based on Gamma-band EEG," *31st Annual International Conference of the IEEE EMBS*, pp. 1323–1326, 2009.

[34] Y. Tay, D. Bahri, D. Metzler, Z. Zhao, and C. Zheng, "Rethinking Self-Attention for Transformer Models," *Icml*, pp. 1–13, 2021.

[35] J. Vig, "A multiscale visualization of attention in the transformer model," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pp. 37–42, 2019.

[36] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," pp. 4190–4197, 2020.

[37] B. Yu, W. Li, X. Li, J. Lu, and J. Zhou, "Frequency-Aware Spatiotemporal Transformers for Video Inpainting Detection," pp. 8188–8197.

[38] E. Sforza, C. Jouny, and V. Ibanez, "Cardiac activation during arousal in humans: Further evidence for hierarchy in the arousal response," *Clinical Neurophysiology*, vol. 111, no. 9, pp. 1611–1619, 2000.

[39] M. M. Müller, A. Keil, T. Gruber, and T. Elbert, "Processing of affective pictures modulates right-hemispheric gamma band EEG activity," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1913–1920, 1999.

[40] S. Guzel Aydin, T. Kaya, and H. Guler, "Wavelet-based study of valence–arousal model of emotions on EEG signals with LabVIEW," *Brain Informatics*, vol. 3, no. 2, pp. 109–117, 2016.