CogniFuse and Multimodal Deformers: A Unified Approach for Benchmarking and Modeling Biosignal Fusion

Anthony Richardson[1], Michael Beetz[2], Tanja Schultz[1] and Felix Putze[1]

***Abstract*— Continuously monitored physiological signals carry rich information about the human body and the biological processes happening within. Extracting this information from casually collected biosignal data in activities of daily living holds great potential for real-time monitoring of physical and mental states, but comes with increased difficulty due to the influence of noise and artifacts. Thus, we create CogniFuse, the first publicly available multi-task benchmark for multimodal biosignal fusion in such challenging environments. For many biosignals, especially electrophysiological signals, the information contained in different frequency bands plays a significant role in analyzing the physiological states of the body. Therefore, we introduce a group of novel fusion models, called Multimodal Deformers, that capture multi-level power features as well as long- and short-term temporal dependencies in multimodal biosignal data. In particular, our proposed Multi-Channel Deformer achieves the highest average benchmark score, outperforming all models of comparison. To assure full transparency and reproducibility, and to support future research on multimodal biosignal fusion, all code and data is made publicly available.**

***Clinical relevance*— Human biosignals provide important insight into various medical conditions. By advancing the field of multimodal biosignal fusion in activities of daily living, this work contributes to the early detection of complex diseases and impairments outside of highly controlled clinical conditions.**

I. Introduction

Due to their ability to provide continuous monitoring of physiological signals, modern wearable sensors have been increasingly integrated into daily living [1]. In combination with previous advances in multimodal neural architectures [2], [3], this integration has great potential to revolutionize real-time monitoring of physical and mental well-being outside of highly controlled clinical conditions. However, this potential comes with a number of difficulties. Manually defining features is not only a time-consuming process that does not scale well for multiple tasks and biosignals, but can also limit a models capability of analyzing the data optimally. Moreover, the importance of each biosignal, and thus the impact it should have on the prediction, can vary between tasks. As this complexity increases with the number of different physical and mental states, architectures are desired that are capable of not only automatically extracting relevant features from raw data, but also self-regulating the impact of each biosignal. This is further complicated by the influence of noise and artifacts caused, for example, by movement [4]. On top of that, research progress in this field is slowed down by the absence of suitable public benchmarks that could facilitate reproducibility and a fair comparison of competing approaches. To the best of our knowledge, there exists no public multi-task multimodal benchmark for biosignal fusion during activities of daily living. Therefore, we provide a novel benchmark based on a comprehensive dataset of people performing different cognitive tasks while driving a car in a simulated environment. Even though this benchmark is not centered around human well-being, it shares the underlying challenge of processing multimodal biosignal data to predict multiple different user states during an activity of daily living. Moreover, we combine the predictive power of Transformers with insights from biosignal processing to create novel models for multimodal biosignal fusion.

II. Related Work

Multimodal Biosignal Benchmarks: There has been a rise in public multimodal biosignal datasets. Koelstra et al. [5] presented DEAP, a dataset for analyzing affective states based on electroencephalogram (EEG) and peripheral biosignals. Miranda-Correa et al. [6] provided AMIGOS, a dataset for affect, personality, and mood prediction based on EEG, electrocardiogram (ECG), and electrodermal activity (EDA). Mason et al. [7] presented EASE-TSD, a dataset for human activity recognition. The recorded biosignals include, motion capture, eye-tracking, speech, as well as muscle and brain activity. However, none of these datasets are accompanied by a benchmarking system for biosignal fusion, which restricts their potential to accelerate research progress in this field. The same was true for MIMIC-III [8], a dataset of patients in intensive care units, until Harutyunyan et al. [9] published a corresponding benchmarking system. In contrast to the MIMIC-III data, which was collected under highly controlled clinical conditions, we provide a more casually collected dataset of multimodal biosignals, including EEG, EDA, photoplethysmography (PPG), and respiration (RESP) recordings, where participants solved various cognitive tasks while driving a car in a simulated environment. This activity of daily living presents increased difficulty, as the recorded biosignals are expected to contain more artifacts and noise due to influences such as movement [4]. To accelerate progress on biosignal fusion in such challenging environments, we publish a benchmarking system that reduces

*This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 459360854 as part of the Research Unit "Lifespan AI: From Longitudinal Data to Lifespan Inference in Health" (DFG FOR 5347), University of Bremen (http://lifespanai.de)

[1] These authors are with the Cognitive Systems Lab, University of Bremen, Germany. Corresponding author: Anthony Richardson. Contact information is provided at: https://orcid.org/0009-0002-0524-8294.

[2] Michael Beetz is with the Institute for Artificial Intelligence, University of Bremen, Germany

the initial effort required by researchers and in doing so, create the first public multi-task multimodal benchmark for biosignal fusion during activities of daily living.

Multimodal Methods for Sequential Data: Transformers have been a popular choice for integrating different modalities of sequential data. Camgoz et al. [2] developed Multi-Channel Transformers, outperforming the state-of-the-art in multi-articulatory sign language translation. Similarly, Chang et al. [3] leverage Transformers for multi-channel speech recognition. Instead of using transformer-based solutions, Lopez et al. [10] designed PHemoNet, a hypercomplex network for physiological signals, claiming to outperform the state-of-the-art in multimodal emotion recognition. However, when processing physiological signals, power bands carry rich temporal information. Yet, none of the previously mentioned models have mechanisms aimed at learning such contents. In contrast, Ding et al. [11] introduced the EEG-Deformer, a unimodal convolutional Transformer capable of learning multi-level power features from raw EEG signals. The resulting model outperformed other state-of-the-art approaches in attention, fatigue, and cognitive workload classification. In subsequent work, Zhang et al. [12] predict the direction of attended speakers by incorporating audio spatial information into the EEG-Deformer. Nonetheless, proficient ways of creating multimodal Deformers and scaling them to more than two modalities are largely unexplored. In this work, we combine multi-channel as well as early-, intermediate-, and late-fusion approaches with concepts from the Deformer architecture. The resulting models can capture multi-level power features as well as long- and short-term temporal dependencies in an arbitrary number of biosignals.

Beyond the state of the art, the contributions of this paper can be summarized as follows:

1) Establishing a standard for benchmarking systems that supports future research by assuring comparability, robustness, reproducibility, and accessibility.
2) Providing the first multi-task multimodal benchmarking system for biosignal fusion during activities of daily living, following our proposed standard.
3) Introducing fusion models that capture multi-level power features as well as long- and short-term temporal dependencies in multimodal biosignal data.

III. Data

The data used in this work was collected as part of the CognitFit project[3] at Karlsruhe Institute of Technology (KIT) and the study was audited by the ethics committee of the State Chamber of Physicians of Baden-Württemberg (Germany). The data in question for this paper is limited exclusively to anonymous data collected by the Cognitive Systems Lab and is made publicly available[4]. As part of the study, 152 participants were presented a variety of cognitive tests. Jarvis et al. [4] used these tests to design four binary discrimination tasks, which are shown in Table I. The tests, included in the four tasks, are defined as follows:

TABLE I
BINARY DISCRIMINATION TASKS BETWEEN LOW AND HIGH WORKLOAD.

Task Name	Low	High
Switching vs. Relax	Relax phase	Switching paradigm
Auditive vs. Relax	Relax phase	Lane change with additional auditive switch-back-2 test
Auditive vs. LC	Lane change	Lane change with additional auditive switch-back-2 test
Visual vs. LC	Lane change	Lane change with additional visual search test

Switching paradigm: Participants were shown a number from the set $[1, ..., 4, 6, ..., 9]$, displayed within either a solid or dashed box. The line style of the box indicated the required action: solid lines prompted an even/odd judgment, while dashed lines required a decision on whether the number is greater or smaller than five.

Lane change (LC): Participants operated a realistic driving simulator built from a real car corpus, where they had to change lanes in response to road signs. Biosignals recorded during this test are expected to contain an increased amount of movement artifacts.

Auditive switch-back-2: Participants listened to a sequence of pre-recorded numbers from the set $[1, ..., 9]$ and were tasked to identify any repeating numbers occurring exactly two positions apart.

Visual search: Participants were presented a screen containing a set of graphical symbols and asked to indicate the position of a specific target symbol.

As part of the study, four biosignals were recorded. Peripheral blood volume pressure was monitored using a PPG attached to the participant's left index finger. EDA was measured through a skin conductivity sensor placed on the palm of the left hand. RESP was tracked using a respiration belt positioned below the thorax and worn over all layers of clothing. Finally, brain activity was recorded using an active EEG with 16 electrodes (Brain Products' ActiCap) [4].

IV. Preprocessing

To enhance the data quality, we performed multiple preprocessing steps for all biosignals. First, butterworth bandpass filters were applied in forward and backward direction to remove unwanted frequencies without causing a phase shift. We chose this filter due to its flat passband response and steep stopband response, which enables it to effectively filter out undesired noise while preserving information on the useful bandwidth [13]. Secondly, all signals were temporally aligned and used to create shorter samples. Every two seconds, a multimodal sample is formed by combining simultaneously starting chunks of varying duration, one for each biosignal. In the case of missing biosignals, the respective sample was discarded. Thirdly, down sampling was performed to reduce the size of the data without losing relevant information. The exact setting for preprocessing are

[3]We owe special thanks to Dominic Heger formerly of KIT, Janina Krell-Rösch and Klaus Bös of KIT, as well as Alexander Stahn of Charité Berlin for collaborating in participant recruitment and organization of the study.

[4]Instructions for obtaining access to the data can be found at https://github.com/anthony-richardson/CogniFuse.

TABLE II

PREPROCESSING SETTINGS OF THE FOUR RECORDED BIOSIGNALS. THE FILTER ORDER REFERS TO THE BUTTERWORTH FILTER.

Biosignal	Filter Order	Low Cutoff Frequency	High Cutoff Frequency	Recording Frequency	Down Sampling Frequency	Chunk Size
EEG	4	0.1 Hz	40 Hz	256 Hz	128 Hz	4 s
PPG	2	0.01 Hz	10 Hz	256 Hz	128 Hz	6 s
EDA	2	0.01 Hz	1 Hz	256 Hz	64 Hz	4 s
RESP	2	0.05 Hz	1 Hz	512 Hz	32 Hz	10 s

shown in Table II. The filter order, cutoff frequencies, and down sampling frequency are chosen individually for each biosignal to ensure minimal impact on potentially relevant frequency ranges. The chunk size for the EEG data is chosen based on work from Ding et al. [11]. For all other chunk sizes we follow Jarvis et al. [4]. After preprocessing, we are left with 119.435 multimodal samples from 134 participants. To create person independent data splits, we first randomly extract 10% of the participants that form the test set. The remaining 90% of the participants are used for 10-fold cross-validation, where in each fold 10% of the remaining participants are used for validation and the rest for training. This results in a train, validation, and test ratio of 81:9:10.

V. BENCHMARK CONSIDERATIONS

When designing this benchmark, multiple considerations were made to support future research on biosignal fusion:

Comparability: When training machine learning models, the choice of training hyperparameters, such as the type of optimizer or the learning rate, have a strong impact on model performance [14], [15]. Consequently, when comparing models trained with different training hyperparameters, a difference in performance cannot be reliably attributed to the underlying architectures. As such a loss of comparability would severely limit the usefulness of a benchmark, we use the same training hyperparameters for all models. For the same reason, the data splits and evaluation approach are consistent throughout the whole benchmark. As different models may reach their highest performance at different training steps, checkpoints are periodically stored and the best one is selected based on validation performance.

Robustness: The ideal choice of training hyperparameters may vary depending on the model [15]. However, comparability restricts them to be the same across all models. Therefore, training hyperparameters used in a benchmark should not be optimized for any particular model. Instead, they should be selected based on robustness, i.e. their ability to perform well across different models and tasks, without requiring fine-tuning for each specific case. Adam [16] and its improved version AdamW [17] have been widely adopted by the machine learning community due to their demonstrated robustness and performance in a variety of tasks [18]. Thus, we choose AdamW as the optimizer for this benchmark. In early experiments, we used a learning rate of 1e-03 and found that some of the learning curves, produced by the models, showed poor convergence. According to Wilson et al. [19], high learning rates can lead to poor convergence but lower learning rates tend to require more training time. Therefore, we decrease the learning rate to 1e-05 and increase the training duration to account for the potentially larger training time requirements. When revisiting the learning curves of the trained models, we observe improved convergence.

Reproducibility: To improve the credibility and trustworthiness of research findings, all results should, if possible, be completely reproducible. Therefore, we make all code, data, and hyperparameters publicly available. Moreover, we use fixed random seeds and configure our benchmarking system to replace all non-deterministic algorithms in neural networks, such as zeroing out units during dropout, with deterministic alternatives. Consequently, repeated training runs result in the exact same benchmark scores.

Accessibility: While public availability of all data and code is crucial for making a benchmark accessible, further steps can be taken. If the aim of a benchmarking system is to allow different users to fairly compare their model performances, we argue that this process should be simplified as much as possible. Therefore, we provide a benchmarking system, named *CogniFuse*, that only requires the user to provide their own model class, inheriting from a provided base class. All other aspects, including parameter loading, optimizer setup, data loading, as well as model training and evaluation, are done automatically, reproducibly, and in compliance with the already existing benchmark results. The source code and a detailed usage guide are provided at https://github.com/anthony-richardson/CogniFuse.

VI. BACKGROUND

The Multimodal Deformers, which will be introduced in Section VII, incorporate multiple neural architectures.

EEG-Deformer: To enhance the perception of temporal dynamics in EEG data, Ding et al. [11] introduced the EEG-Deformer, a convolutional Transformer built to capture both fined-grained (short-term) and coarse-grained (long-term) temporal dependencies. Its network structure is shown in Figures 1 and 2. After an initial CNN-based shallow feature encoder comprising collaborative temporal and spatial convolutional layers, the EEG-Deformer employs

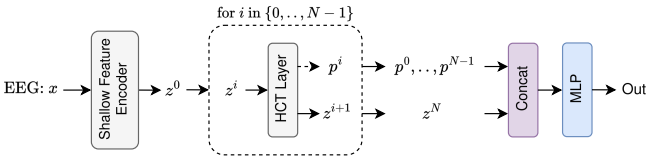

Fig. 1. Overview of the EEG-Deformer.

Hierarchical Coarse-to-Fine Transformer (HCT) blocks that integrate a Fine-grained Temporal Learning (FTL) branch into Transformers. The FTL branch employs a 1-D CNN to capture short-period temporal dynamics of the EEG signal, generating fine-grained temporal representations. These representations are then adaptively fused with the coarse-grained temporal representations encoded by the Transformers, thereby providing more discriminative long- and short-term temporal information. Moreover, Ding et al. [11] aim to utilize the rich multi-level temporal information encoded within the intermediate layers of the network. Therefore, they proposed Information Purification Units (IP-Units), that transmit multi-level representations from the HCT layers to the final representation. When designing the IP-units, Ding et al. [11] drew inspiration from neural engineering, where power features of EEG signals in different frequency bands are used for brain activity analysis. The IP-units calculate the power $\boldsymbol{p}^i$ of a learned 1-D hidden representation $\boldsymbol{f}^i$ by

$$\begin{aligned}\boldsymbol{p}^i &= \mathrm{IP}_{power}(\boldsymbol{f}^i) \\ &= \{\log\left(\frac{1}{l}\sum(\boldsymbol{f}^{i,j})^2\right) : \boldsymbol{f}^{i,j} \in \boldsymbol{f}^i\},\end{aligned}$$

where $\log(\cdot)$ is the logarithm and $\boldsymbol{f}^{i,j} \in \mathbb{R}^l$ is a learned representation from the j-th CNN kernel in a Fine-grained Temporal Learning (FTL) branch at depth i. While Ding et al. [11] designed this concept for processing EEG signals, we argue that the importance of different power bands is not limited to neural activity, but can find application in many types of biosignals. For instance, Jarvis et al. [4] use different power bands of PPG and RESP signals to predict cognitive workload. With regards to EDA signals, Posada-quintero et al. [20] and Shimomura et al. [21] discover significant increases of power in frequency bands when participants are subjected to cognitive stress. Therefore, we generalize the EEG-Deformer and apply it to all biosignals included in this work. More importantly, we use its components to create a group of novel fusion models, called Multimodal Deformers.

Multi-Channel Attention: To model sequence-to-sequence learning problems where the source information is embedded across several asynchronous channels, Camgoz et al. [2] introduced Multi-Channel Transformers. They calculate multi-channel attention by using a primary channel as query and concatenating the remaining channels on the time axis to obtain keys and values. We reformulate this mechanism and create a novel Multi-Channel Encoder (MCE) layer, specifically designed for convolutional Transformers. The structure of this layer is depicted in Figure 3. Initially, the queries Q_s, keys K_s, and values V_s are obtained independently for each channel s. Given a primary channel a and three secondary channels b, c, and d, the query for the multi-head attention is given by Q_a, while $K_{[b,c,d]}$ and $V_{[b,c,d]}$ are created by concatenating the respective keys and values of the secondary channels. Instead of concatenating over the time axis as done by Camgoz et al. [2], we perform the concatenation over the kernel dimension of our convolutional Transformer. This makes it possible to process the different channels with

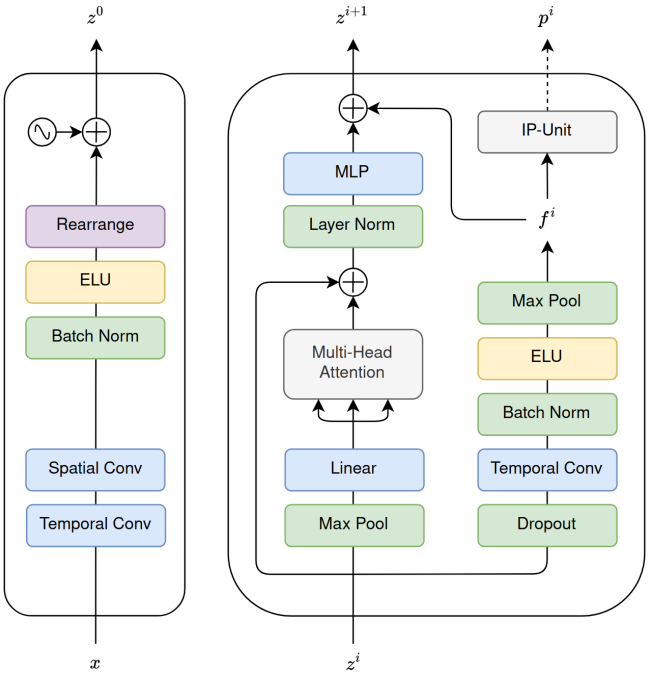

Fig. 2. Network structures of the shallow feature encoder (left) and HCT layer (right). Both components are part of the EEG-Deformer.

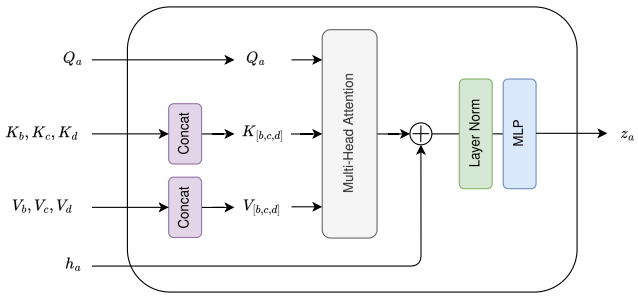

Fig. 3. Network structure of our MCE layer for four modalities, where a is the primary modality and b, c, and d are the three secondary modalities.

different amounts of kernels, or in other words, to allocate model resources according to the complexities of the different biosignals. After adding a residual connection for h_a, i.e. the input used to generate Q_a, we apply layer normalization and pass the output to a Multi-Layer Perceptron (MLP) to obtain the final output of the MCE layer.

VII. Multimodal Deformers

To capture multi-level power features as well as long- and short-term temporal dependencies in multimodal biosignal data, we introduce Multimodal Deformers, a group of novel fusion models that incorporate mechanisms from the Deformer architecture into various fusion approaches.

Multi-Channel Deformer: Combining concepts from the Deformer architecture with our MCE layer from Figure 3, we introduce Multi-Channel Deformers, shown in Figure 4. We will refer to the different input channels as biosignals. First, all biosignals are encoded by independent shallow feature encoders. Each resulting encoding is then processed by a sequence of alternating HCT and MCE layers, where the HCT

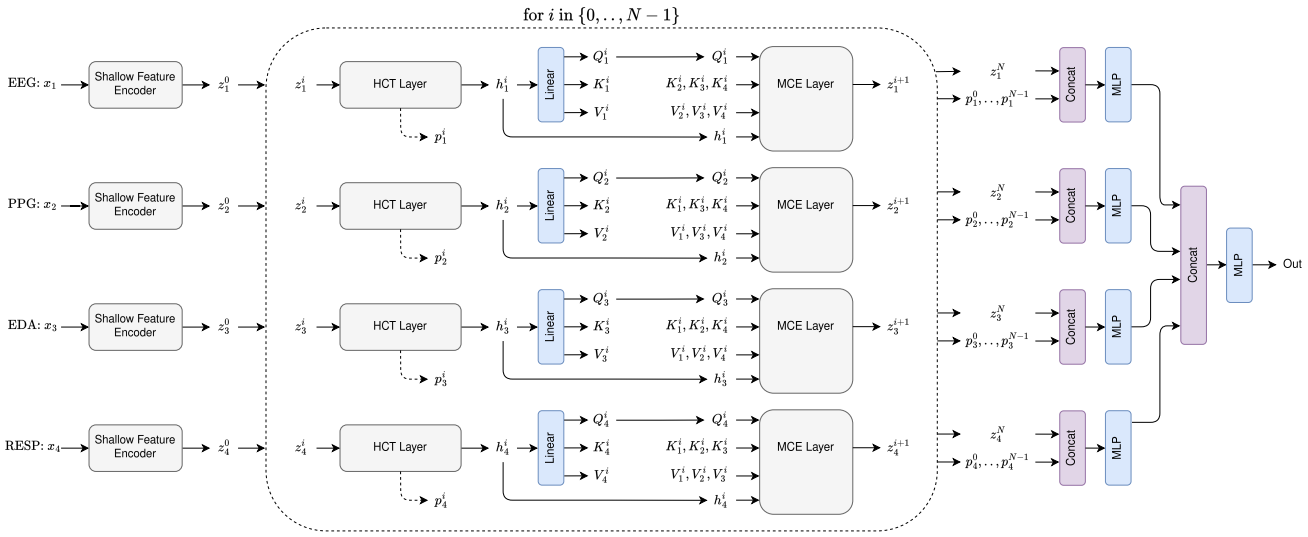

Fig. 4. Overview of our proposed Multi-Channel Deformer.

layers capture short- and long-term temporal dependencies, while the MCE layers learn the contextual relationship between the different biosignals. At the end of each HCT layer, a pooling operation reduces the temporal dimension of the learned representation, such that the outputs of the IP-units represent power features at different temporal resolutions. The pooling operations also help to reduce the model size. After N alternations of HCT and MCE layers follow two information aggregation stages. In the first aggregation stage, the multi-level power features and the output of the last MCE layer are concatenated and passed to a MLP, independently for each biosignal, resulting in multiple biosignal specific representations. In the second stage, these representations are concatenated and passed to another MLP, producing the final output of the model. Not only is the resulting architecture capable of processing an arbitrary number of biosignals with different durations and sampling frequencies, but it also enables an unequal distribution of model resources to its channels. This means that the available model parameters can be distributed according to the complexity of the biosignals.

Late-Fusion Deformer: Independent Deformers are trained for each biosignal. The final prediction is formed by calculating the weighted average of the four predictions, where the weights are determined by the validation performance of the individual models.

Intermediate-Fusion Deformer: Initially, each biosignal is processed by an independent Deformer. However, instead of obtaining a prediction from each Deformer, the final representations of the biosignals are concatenated and then passed to a MLP that makes a single prediction.

Early-Fusion Deformer: Initially, each biosignal is encoded by an individual shallow feature encoder. The encoded biosignals are then concatenated on the time axis and positional encoding is added. From there on, the result is passed to a single sequence of HCT layers and processed the same way as in the original EEG-Deformer.

VIII. Evaluation

Tables III and IV show the performance of different models on the validation and test data. Scores from Jarvis et al. [4], while based on manually defined features as well as different preprocessing and data splits, are included as rough reference points despite limited comparability. To ensure fair comparisons, we linearly scale the models to match the Multi-Channel Deformer's size within $\pm$ 0.1%. Excluded from this procedure are LDA-based models, since they are not scalable, as well as the Late-Fusion Deformer, for which we only use the unscaled unimodal Deformers as fusion components. Due to the existence of multiple benchmark tasks and the interest in models that perform well in all of them, a key evaluation metric will be their average scores. These are reported in the last columns of Tables III and IV.

First, it can be observed that on average all unimodal Deformers, except the unscaled EDA-Deformer, outperform their respective LDA counterpart on the test data, demonstrating that unimodal Deformers are capable of capturing relevant temporal dependencies in all of the four biosignals.

When comparing all results from the benchmark, our proposed Multi-Channel Deformer achieves the highest average score, both on the validation and test data. In particular, all of our Multimodal Deformers on average outperform all fusion models from related work, which do not have mechanisms aimed at learning power features from physiological signals. Nonetheless, on certain tasks, unimodal models achieve the highest scores, reflecting the challenge, introduced in Section I, of using multiple biosignals for multiple different tasks that vary with regards to the importance of specific biosignals. For example, unimodal Deformer scores suggest that EDA and PPG signals are most relevant for the Auditive vs. Relax task, while EEG and RESP signals are more relevant for the Switching vs. Relax task. To perform well in the benchmark, a fusion approach must be capable of utilizing biosignals that provide insight for each respective task, without suffer-

ing from the influence of other, less insightful biosignals. Whether a model is capable of doing so becomes apparent when comparing its performance to the best unimodal models for each respective task and when inspecting its average score across all tasks. Late-Fusion Deformers, for instance, are outperformed by EEG-Deformers in all tasks except Auditive vs. Relax, indicating that the late-fusion approach impairs the predictive potential contained in the EEG data. In contrast, Multi-Channel Deformers are either close to or outperform all unimodal Deformers at every benchmark task.

Last but not least, the results often show significant performance drops in the test scores compared to the validation scores. We suspect the cause for this to be a combination of fluctuation in performance during training and choosing the best models based on the validation data. As a result, only the test scores are effected by this fluctuation, thus widening the gap in performance between validation and testing.

IX. Conclusion

We presented a novel multi-task multimodal biosignal benchmark with a strong relation to everyday life, where data collection is often uncontrolled and the importance of specific biosignals can vary between tasks. Under careful consideration of comparability, robustness, reproducibility, and accessibility, we designed a benchmarking system that not only enabled a thorough evaluation of our proposed neural architectures, but also minimizes the required effort of future researchers when adding new models. Anticipating that the presented challenge is closely related to the detection of different impairments and diseases in people's everyday life, we aim to bridge the gap between the methods presented in this work and their medical application. Therefore, future work will consist of collecting multimodal biosignal data of impairments and diseases during activities of daily living.

After transferring the EEG-Deformer to other unimodal biosignals, we introduced Multimodal Deformers, a group of novel fusion models capable of capturing multi-level power features as well as long- and short-term temporal dependencies in multimodal biosignal data. Here, our proposed Multi-Channel Deformer achieved the highest average benchmark score, demonstrating its ability to effectively utilize multiple biosignals that provide insight for each respective task, without suffering from performance loss due to other, less insightful biosignals. In addition, all of our Multimodal Deformers on average outperformed all models from related work that lack mechanisms for learning power features, highlighting the importance of such features. However, for most of the trained models, a gap between the validation and test scores was observed and attributed to performance fluctuations during training. Future work could explore the use of learning rate schedulers that improve training stability.

Training individual models for each task comes with a high demand in compute resources. In the future, we plan to use unsupervised pre-training as well as fine-tuning methods, to first learn a general knowledge representation of human physiology and then adapt that representation to different downstream tasks.

References

[1] Xu, Z., Hao, Y., Luo, A., and Jiang, Y. (2024). Technologies and applications in wireless biosensors for real-time health monitoring. Med-X, 2(1), 24.

[2] Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020). Multi-channel transformers for multi-articulatory sign language translation. In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16 (pp. 301-319). Springer International Publishing.

[3] Chang, F. J., Radfar, M., Mouchtaris, A., King, B., and Kunzmann, S. (2021, June). End-to-end multi-channel transformer for speech recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5884-5888). IEEE.

[4] Jarvis, J., Putze, F., Heger, D., and Schultz, T. (2011). Multimodal person independent recognition of workload related biosignal patterns. In Proceedings of the 13th international conference on multimodal interfaces.

[5] Koelstra, S., Muhl, C., Soleymani, M., Lee, ..., and Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing, 3(1), 18-31.

[6] Miranda-Correa, J. A., Abadi, M. K., Sebe, N., and Patras, I. (2018). Amigos: A dataset for affect, personality and mood research on individuals and groups. IEEE transactions on affective computing, 12(2), 479-493.

[7] Mason, C., Gadzicki, K., Meier, M., Ahrens, F., Kluss, T., Maldonado, J., ... and Schultz, T. (2020, October). From human to robot everyday activity. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 8997-9004). IEEE.

[8] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ..., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific data, 3(1), 1-9.

[9] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific data, 6(1), 96.

[10] Lopez, E., Uncini, A., and Comminiello, D. (2024, September). PHemoNet: A Multimodal Network for Physiological Signals. In 2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI) (pp. 260-264). IEEE.

[11] Ding, Y., Li, Y., Sun, H., Liu, R., Tong, C., Liu, C., ..., and Guan, C. (2024). EEG-Deformer: A dense convolutional transformer for brain-computer interfaces. IEEE Journal of Biomedical and Health Informatics.

[12] Zhang, Y., Lu, J., Lin, Z., Chen, F., Du, H., and Gao, X. (2024). Electroencephalogram-based Multi-class Decoding of Attended Speakers' Direction with Audio Spatial Spectrum. arXiv preprint arXiv:2411.06928.

[13] Wang, B. (2024). Signal Processing Based on Butterworth Filter: Properties, Design, and Applications. Highlights in Science, Engineering and Technology.

[14] Ren, P., Xiao, Y., Chang, X., Huang, P., Li, Z., Chen, X., and Wang, X. (2021). A Comprehensive Survey of Neural Architecture Search: Challenges and Solutions. Association for Computing Machinery.

[15] Liao, L., Li, H., Shang, W., and Ma, L. (2022). An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks. Association for Computing Machinery.

[16] Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization.

[17] Loshchilov, I., and Hutter, F. (2017). Fixing weight decay regularization in adam.

[18] Zhuang, Z., Liu, M., Cutkosky, A., and Orabona, F. (2022). Understanding adamw through proximal methods and scale-freeness. Transactions on machine learning research.

[19] Wilson, D. R., and Martinez, T. R. (2001). The need for small learning rates on large problems. In IJCNN'01. International Joint Conference on Neural Networks.

[20] Posada-Quintero, H. F., Florian, J. P., Orjuela-Cañón, A. D., Aljama-Corrales, T., Charleston-Villalobos, S., and Chon, K. H. (2016). Power spectral density analysis of electrodermal activity for sympathetic function assessment. Annals of biomedical engineering.

[21] Shimomura, Y., Yoda, T., Sugiura, K., Horiguchi, A., Iwanaga, K., and Katsuura, T. (2008). Use of frequency domain analysis of skin conductance for evaluation of mental workload. Journal of physiological anthropology, 27(4), 173-177.

TABLE III

AVERAGE MICRO F1-SCORES OF 10-FOLD CROSS-VALIDATIONS FOR UNIMODAL AND MULTIMODAL METHODS ON THE VALIDATION DATA OF THE FOUR TASKS. THE BASELINE CLASSIFIERS (ZEROR) ALWAYS PREDICT THE MAJORITY CLASS. THE BASELINE SCORES ALSO REFLECT THE CLASS IMBALANCE IN EACH TASK. THE RESULTS FROM JARVIS ET AL. [4] ARE BASED ON MANUALLY DEFINED FEATURES AS WELL AS DIFFERENT DATA SPLITS AND PREPROCESSING STEPS, AND ARE NOT PRODUCED BY US. THE PLUS MINUS (±) INDICATES THE STANDARD DEVIATION ACROSS FOLDS. TO APPLY THE APPROACHES FROM CAMGOZ ET AL. [2] AND CHANG ET AL. [3] TO CLASSIFICATION TASKS, WE ADD A MULTI-LAYER PERCEPTRON (MLP) AT THE END OF THEIR MULTI-CHANNEL ENCODERS AND DISCARD THE DECODERS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE NEXT BEST ARE MARKED USING UNDERLINES. UP AND DOWN ARROWS (⇅) INDICATE MODEL SCALING TO MATCH THE SIZE OF THE MULTI-CHANNEL DEFORMER WITH A MARGIN OF ± 0.1%.

Method	Parameters	Switching vs. Relax	Auditive vs. Relax	Auditive vs. LC	Visual vs. LC	Average
Baseline (ZeroR)		70.31 ± 2.02	65.73 ± 0.81	50.78 ± 0.94	51.34 ± 1.05	59.54
EEG-Bandbased-LDA [4]	65	85.54	72.14	55.69	73.82	71.80
EEG-Wavelet-LDA [4]	65	84.92	49.66	65.15	62.89	65.66
PPG-LDA [4]	6	79.66	81.27	49.04	51.03	65.25
EDA-LDA [4]	7	84.90	71.70	62.64	56.45	68.92
RESP-LDA [4]	8	81.60	72.25	42.49	49.41	61.44
EEG-Deformer [11]	1.417.218	99.14 ± 0.89	78.61 ± 1.22	**68.03** ± 2.41	**82.43** ± 1.71	82.05
EEG-Deformer [11] ⇅	4.948.041	99.69 ± 0.30	79.94 ± 1.61	67.48 ± 2.09	82.18 ± 1.51	82.32
PPG-Deformer [11]	771.074	72.22 ± 1.40	83.91 ± 1.65	53.59 ± 1.36	52.50 ± 1.41	65.56
PPG-Deformer [11] ⇅	4.948.027	72.89 ± 0.98	84.28 ± 1.75	54.18 ± 1.07	52.99 ± 1.09	66.09
EDA-Deformer [11]	257.922	71.77 ± 2.25	83.93 ± 0.95	52.21 ± 0.82	52.68 ± 1.27	65.15
EDA-Deformer [11] ⇅	4.948.156	71.78 ± 2.20	85.16 ± 1.05	53.86 ± 1.86	55.02 ± 1.29	66.46
RESP-Deformer [11]	322.066	78.59 ± 1.91	80.86 ± 1.59	55.39 ± 1.57	57.51 ± 0.63	68.09
RESP-Deformer [11] ⇅	4.949.790	80.55 ± 2.20	81.79 ± 1.43	56.69 ± 1.32	56.96 ± 0.95	69.00
Late-Fusion LDA [4]	151	93.28	83.29	60.50	66.16	75.81
Multi-Channel Encoder [2] + MLP	11.100.856	98.49 ± 1.38	89.34 ± 1.43	58.37 ± 1.99	69.34 ± 1.60	78.89
Multi-Channel Encoder [2] + MLP ⇅	4.950.098	98.43 ± 1.38	89.30 ± 0.96	57.75 ± 1.85	68.42 ± 1.59	78.48
Multi-Channel Encoder [3] + MLP	14.609.666	98.15 ± 1.22	89.54 ± 1.62	56.97 ± 1.03	65.84 ± 1.56	77.63
Multi-Channel Encoder [3] + MLP ⇅	4.950.975	98.08 ± 1.27	89.41 ± 1.25	56.20 ± 1.13	64.11 ± 1.81	76.95
PHemoNet [10]	15.235.586	97.30 ± 2.19	85.65 ± 1.19	58.48 ± 0.77	69.00 ± 1.82	77.61
PHemoNet [10] ⇅	4.956.930	97.28 ± 2.18	84.07 ± 1.64	58.94 ± 1.09	69.34 ± 1.74	77.41
Late-Fusion Deformer (ours)	2.768.280	90.90 ± 1.97	**91.36** ± 0.72	63.88 ± 2.47	80.24 ± 1.82	81.60
Intermediate-Fusion Deformer (ours) ⇅	4.948.560	99.77 ± 0.32	89.85 ± 1.38	64.81 ± 4.64	81.81 ± 1.71	84.06
Early-Fusion Deformer (ours) ⇅	4.952.876	99.58 ± 0.55	89.44 ± 0.70	62.89 ± 3.13	81.01 ± 1.66	83.23
Multi-Channel Deformer (ours)	4.952.962	**99.81** ± 0.28	89.70 ± 2.23	66.79 ± 1.78	82.29 ± 1.82	**84.65**

TABLE IV

AVERAGE MICRO F1-SCORES OF 10-FOLD CROSS-VALIDATIONS FOR UNIMODAL AND MULTIMODAL METHODS ON THE TEST DATA OF THE FOUR TASKS. THE BASELINE CLASSIFIERS (ZEROR) ALWAYS PREDICT THE MAJORITY CLASS. THE BASELINE SCORES ALSO REFLECT THE CLASS IMBALANCE IN EACH TASK. THE RESULTS FROM JARVIS ET AL. [4] ARE BASED ON MANUALLY DEFINED FEATURES AS WELL AS DIFFERENT DATA SPLITS AND PREPROCESSING STEPS, AND ARE NOT PRODUCED BY US. THE SAME NOTATION IS USED AS IN TABLE III.

Method	Parameters	Switching vs. Relax	Auditive vs. Relax	Auditive vs. LC	Visual vs. LC	Average
Baseline (ZeroR)		69.87 ± 0.00	64.58 ± 0.00	48.44 ± 0.00	50.78 ± 0.00	58.42
EEG-Bandbased-LDA [4]	65	76.16	80.54	64.35	75.93	74.25
EEG-Wavelet-LDA [4]	65	76.04	71.74	**65.31**	62.82	68.98
PPG-LDA [4]	6	73.40	86.41	49.15	43.59	63.14
EDA-LDA [4]	7	84.75	67.83	57.51	49.96	65.01
RESP-LDA [4]	8	81.24	81.66	43.65	45.17	62.93
EEG-Deformer [11]	1.417.218	96.10 ± 0.99	77.19 ± 1.53	59.31 ± 0.91	**78.12** ± 0.49	77.68
EEG-Deformer [11] ⇅	4.948.041	96.07 ± 1.12	78.44 ± 1.17	59.05 ± 0.68	77.35 ± 1.91	77.73
PPG-Deformer [11]	771.074	73.57 ± 1.99	82.30 ± 1.24	50.89 ± 0.71	49.80 ± 1.27	64.14
PPG-Deformer [11] ⇅	4.948.027	74.49 ± 1.81	83.66 ± 1.01	50.54 ± 1.29	48.48 ± 1.13	64.29
EDA-Deformer [11]	257.922	70.52 ± 0.58	84.78 ± 1.98	48.52 ± 0.42	51.25 ± 1.70	63.77
EDA-Deformer [11] ⇅	4.948.156	71.67 ± 1.30	86.44 ± 0.36	49.22 ± 0.91	52.93 ± 0.84	65.07
RESP-Deformer [11]	322.066	84.64 ± 0.52	79.72 ± 1.88	50.99 ± 1.05	48.52 ± 0.94	65.97
RESP-Deformer [11] ⇅	4.949.790	85.27 ± 1.13	80.55 ± 0.88	50.06 ± 0.90	48.89 ± 1.30	66.19
Late-Fusion LDA [4]	151	90.66	**93.58**	64.48	62.40	77.78
Multi-Channel Encoder [2] + MLP	11.100.856	97.16 ± 0.58	85.48 ± 1.10	52.40 ± 0.90	62.53 ± 1.03	74.39
Multi-Channel Encoder [2] + MLP ⇅	4.950.098	**97.24** ± 0.50	85.83 ± 0.66	52.28 ± 1.36	62.68 ± 0.92	74.51
Multi-Channel Encoder [3] + MLP	14.609.666	96.28 ± 0.30	88.88 ± 0.82	52.77 ± 0.81	60.20 ± 1.59	74.53
Multi-Channel Encoder [3] + MLP ⇅	4.950.975	96.20 ± 0.36	88.78 ± 1.06	51.71 ± 0.73	59.22 ± 1.54	73.98
PHemoNet [10]	15.235.586	96.10 ± 0.29	84.20 ± 1.05	52.66 ± 0.76	64.54 ± 0.78	74.38
PHemoNet [10] ⇅	4.956.930	95.99 ± 0.45	82.55 ± 0.92	53.32 ± 1.03	65.13 ± 1.01	74.25
Late-Fusion Deformer (ours)	2.768.280	90.11 ± 1.33	89.84 ± 1.49	57.41 ± 1.03	76.36 ± 2.09	78.43
Intermediate-Fusion Deformer (ours) ⇅	4.948.560	96.80 ± 0.52	83.49 ± 2.92	58.11 ± 1.22	77.53 ± 0.76	78.98
Early-Fusion Deformer (ours) ⇅	4.952.876	96.69 ± 0.66	86.19 ± 1.09	57.48 ± 1.98	75.30 ± 0.83	78.92
Multi-Channel Deformer (ours)	4.952.962	97.07 ± 0.87	84.54 ± 1.84	59.10 ± 1.12	76.34 ± 0.68	**79.26**