Multimodal Fusion Networks for Workload Modeling

Shengnan Hu
Department of Computer Science
Central China Normal University
Wuhan, China
Email: shengnanhu@ccnu.edu.cn

Gita Sukthankar
Department of Computer Science
University of Central Florida
Orlando, FL USA
Email: gita.sukthankar@ucf.edu

***Abstract*—The advent of low cost sensors for measuring gaze, heart rate, EEG, and galvanic skin response have made it feasible to cheaply collect physiological data from human operators. However, leveraging this data for machine learning problems requires a good multimodal fusion architecture. When dealing with multimodal features, uncovering the correlations between different modalities is as crucial as identifying effective unimodal features. This paper proposes a hybrid multimodal tensor fusion network that is effective at learning both unimodal and bimodal dynamics for cognitive workload modeling. Our architecture comprises two parts: (1) intra-modality for learning high-level representations of each signal modality (2) inter-modality for modeling bimodal interactions using a tensor fusion layer created from the Cartesian product of modality embeddings. We compare this architecture to the usage of a cross-modal transformer fusion module that learns an inter-modality embedding. Experimental results conducted on the HP Omnicept Cognitive Load Database (HPO-CLD) show that both techniques outperform the most commonly used techniques used for multimodal fusion of physiological data and that the cross-modal transformer fusion module is especially effective.**

***Index Terms*—fusion architectures, workload modeling, multimodal learning, physiological data**

I. Introduction

A key desideratum for effective human-machine teaming is the ability to identify cognitive workload, the amount of mental effort exerted by the human operator [1]–[3]. Rather than relying on subjective, post-task questionnaires such as the NASA Task Load Index [4], physiological sensors have made it feasible to monitor human workload in real-time [5], [6]. Unfortunately, single modality sensing solutions are more vulnerable to external confounds such as muscle movement, session length, and temperature [7], [8]. Multimodal machine learning systems are robust to single modality sensor failures but are sensitive to the multicollinearity arising from correlated features from different modalities [9]. This paper introduces a hybrid multimodal tensor fusion network that accurately estimates workload in real-time from physiological data from gaze trackers and cardiovascular monitors. Our architecture learns a high level representation for each signal modality and explicitly models bimodal interactions using a tensor fusion layer created from the Cartesian product of modality embeddings. Our ablative study demonstrates that there is a clear benefit to learning separate pipelines for unimodal and multimodal interactions, despite the redundancies in representation. We demonstrate that further performance improvements can be achieved by enforcing the learning of an inter-modality embedding using a cross-modal transformer.

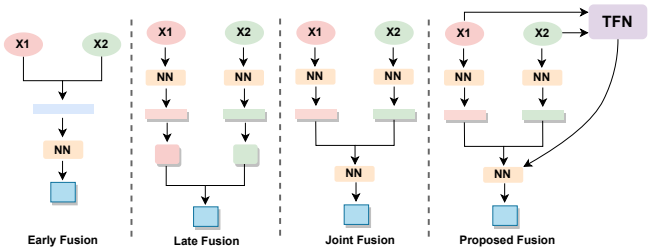

Fig. 1. Multimodal fusion approaches.

Cognitive workload can be estimated with both behavioral and physiological signals [1]–[3]; however physiological data can be non-invasively acquired and possesses greater domain adaptation potential, due to lower task dependency [10]. Although brain-computing sensors such as EEG, fNIRS, MEG, and fMRI can be used to detect cognitive workload directly [7], several studies [2], [8], [11] have shown that gaze, cardiovascular measures, and galvanic skin response remain important data sources for workload monitoring even when more direct neurophysiological data is available. Unlike multimodal models developed for tasks such as audio-visual speech recognition [12] and sentiment analysis [13] that incorporate computer vision and natural language data, multimodal physiological models can not rely on pre-trained models and large data sets due to the difficulties of generalizing biosensor data across differently calibrated setups [8].

Figure 1 illustrates the different options for multimodal fusion. In early fusion models, decision making is deferred until after different modalities have been aggregated [14]. In late fusion models, classification is performed on each modality separately, and the final decision is reached by aggregating classifier outputs using techniques such as weighting, averaging, or voting [12]. Late fusion models are less vulnerable to multicollinearity, but lose the ability to learn multimodal feature representations. Our proposed model (shown to the far

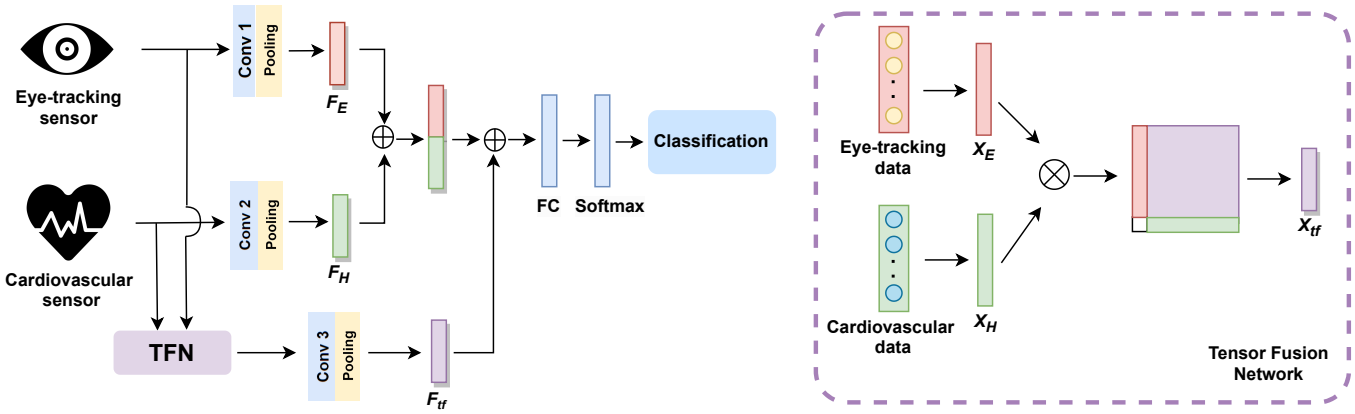

Fig. 2. Architecture of our proposed hybrid tensor fusion model. Two convolutional networks are used to learn separate unimodal embeddings for the gaze and cardiovascular features. The bimodal embedding is learned directly from the original features using a tensor fusion network. These three embeddings are jointly fused using fully connected and softmax layers to predict the cognitive workload.

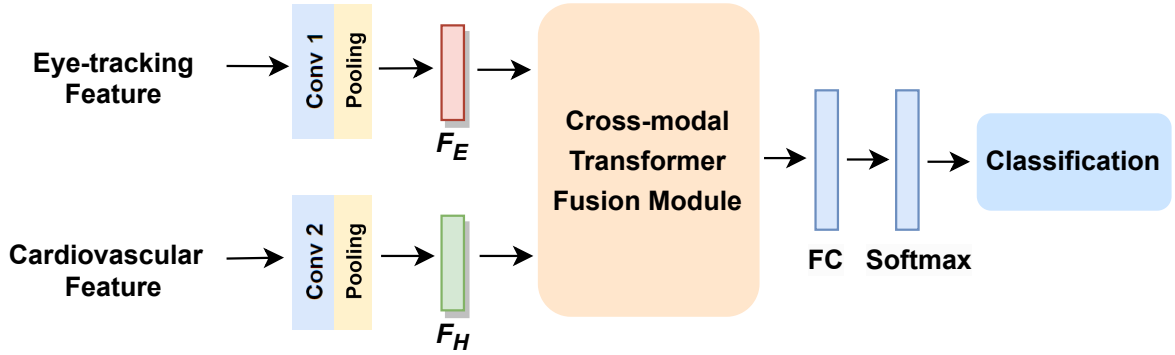

Fig. 3. Our cross-modal transformer architecture. Two convolutional networks are used to learn separate unimodal embeddings for the gaze and cardiovascular features. A cross-modal transformer fusion module then enforces the learning of an inter-modality embedding, and the final cognitive workload prediction is performed using fully connected and softmax layers.

right of Fig. 1) is an example of a joint fusion model that leverages both unimodal and bimodal representation learning.

II. Method

A. Problem Statement

The framework of the proposed algorithm is shown in Fig. 2, and our code is publicly available at https://github.com/shengnanh20/Hybrid_TFN_for_Workload_Modeling. There are three main phases in our algorithm. First, the intra-modality embedding subnetworks take unimodal features, including eye-tracking and cardiovascular features, as input and produce rich unimodal embeddings. Second, the inter-modality tensor fusion module models interactions between bimodal inputs by utilizing a Cartesian product derived from the modality embeddings. Finally, the cognitive load inference subnetwork ingests both intra-modality embedding features and inter-modality fused features as input and predicts the subject's cognitive load.

B. Intra-modality Embedding Subnetworks

Given feature vectors from two modalities, including the eye-tracking features X_E and cardiovascular features X_H, a 1-dimensional convolutional network is used to capture the intrinsic dynamics within each modality. Next, we incorporate a max-pooling layer followed by a Rectified Linear Unit (ReLU) activation layer to further enhance the unimodal representations. Note that these models could easily be replaced by specialized modality-specific models if available.

C. Inter-modality Embedding Subnetworks

Inspired by the multimodal tensor fusion proposed by Zadeh et al. [13] for multimodal sentiment analysis, we incorporate

Algorithm 1 Forward Pass of the Proposed Model

function FORWARD(X_E, X_H, X_{tf})
  $X_E \leftarrow \text{ReLU}(\text{Conv1}(X_E))$
  $F_E \leftarrow \text{MaxPool}(X_E)$
  $X_H \leftarrow \text{ReLU}(\text{Conv2}(X_H))$
  $F_H \leftarrow \text{MaxPool}(X_H)$
  $X_{\text{tf}} \leftarrow \text{ReLU}(\text{Conv3}(X_{\text{tf}}))$
  $F_{\text{tf}} \leftarrow \text{MaxPool}(X_{\text{tf}})$
  $\text{Feat} \leftarrow \text{Concatenate}(F_E, F_H, F_{\text{tf}}, \text{axis} = 2)$
  $\text{Feat} \leftarrow \text{FullyConnectedLayer}(\text{Feat})$
  $\text{Out} \leftarrow \text{Softmax}(\text{Feat})$
  return Out
end function

a Tensor Fusion Network into our architecture to enforce the learning of an inter-modality embedding.

Given each input pair (X_E, X_H), an extra constant dimension with value 1 is introduced such that unimodal dynamics can be represented within the tensor model as $[X_E, 1]^T$ and $[X_H, 1]^T$. Then, to capture the bimodal interactions, a differentiable outer product between these two embeddings is conducted:

$$Tensor(X_E, X_H) = \begin{bmatrix} X_E \\ 1 \end{bmatrix} \otimes \begin{bmatrix} X_H \\ 1 \end{bmatrix}, \tag{1}$$

where $Tensor(X_E, X_H)$ can be used for bimodal representation and $\otimes$ implies the outer product between vectors. The fused tensor is then flattened and fed into a 1-dimensional convolutional network to learn a high-level representation of the inter-modality embedding. Although in this paper, we only demonstrate the use of a bimodal tensor; the tensor fusion network can be generalized to an arbitrary number of dimensions assuming that there is sufficient data to fit the parameters.

D. Cross-modal Transformer Fusion

In cases where the goal is to learn bimodal dynamics, a single cross-attention transformer block (Figure 3) is sufficiently powerful to learn the correlation between between pairs of modalities. Embedding features extracted from each modality are fed into the transformer fusion block via attention. Here we adopt the decoder structure of transformer [15] as our fusion module to amalgamate these features. Thus, our cross-attention block can capture the correlations between each pair of gaze and cardiovascular feature vectors. Given extracted gaze features F_E and cardiovascular features F_H, our cross-attention fusion is defined as:

$$\begin{aligned} Attention(Q, K, V) &= Attention(F_E, F_H, F_H) \\ &= \text{softmax}(\frac{F_E F_H^T}{\sqrt{d}}) F_H, \end{aligned}$$

where d indicates the dimension of F_H. This allows the correlation between the eye-tracking features and cardiovascular features to be constructed using attention computation.

E. Classification and Loss Function

After performing both intra-modality and inter-modality learning, we concatenate all three feature vectors into a unified vector representation. Subsequently, we introduce a fully connected layer, followed by a softmax layer, to facilitate the classification process. The complete training process is illustrated in Algorithm 1. A cross-entropy loss function is applied throughout the training to guide the learning process effectively:

$$CE(y^{out}, y) = -\sum_{i=0}^{N-1} y_i log(y_i^{out}), \tag{2}$$

where y_i represents the ground truth for class i, y_i^{out} implies the predicted probability of class i obtained by the softmax function, and N is the total number of classes.

III. Experiments

A. Dataset and Settings

We conducted our analysis on the HP Omnicept Cognitive Load Database (HPO-CLD) [16]. HPO-CLD includes data from 100 participants who performed a series of tasks that were explicitly designed to require different levels of mental effort (low, medium, high), to complete. Gaze tracking data was gathered using the HTC Vive Pro-eye head-mounted display equipment, which includes eye tracking and pupillometry capabilities. In addition, a BITalino (r)evolution wired pulse plethysmography (PPG) sensor was utilized to measure cardiac activities non-invasively during task execution. We selected this dataset since it uses off the shelf sensor equipment that can be cheaply integrated into many human-robot interaction setups. Although the dataset only contains two sensor modalities, it does not rely on subjective workload assessments and has three times as many subjects as most cognitive workload assessment datasets (see [3] for a literature review on data availability).

a) Eye tracking features: Following [16], we extract gaze features in 12.5-second windows on the data collected through the eye tracking API, which records pupil position, pupil diameter, gaze position, and gaze direction while participants were engaged in cognitive load (CL) tasks. After data buffering and normalization procedures, we extracted a total of eleven variables to represent gaze movements. These variables included features related to pupil diameter, blink behavior, and saccadic eye movements.

b) Cardiovascular features: Heart activities are measured with a lightweight PPG sensor, which detects alterations in blood flow at the specific skin location where the sensor is positioned. Following a sequence of data filtering, decomposition, and normalization, we extract a feature set that comprises nine distinct variables closely associated with heart rate as a comprehensive representation of the cardiac activity being monitored.

The dataset was randomly divided into training and testing sets at a ratio of $4:1$, with no overlap. Results are presented on the testing set only. The workload modeling problem is

Fusion Method	Acc	Recall	F1	AUC-ROC
Early fusion	67.5	67.7	66.7	83.9
Late fusion	63.9	64.1	62.9	81.4
Joint fusion	67.4	67.5	66.7	83.9
Proposed	**69.9**	**70.1**	**69.4**	**84.0**

TABLE I
COMPARISON BETWEEN DIFFERENT FUSION METHODS.

Model	Multimodal input			Eye-tracking feature		
	Acc	Recall	F1	Acc	Recall	F1
Logistic Regression [17]	44.4	44.4	41.8	59.3	59.4	58.7
Naive Bayes (Gaussian) [18]	46.6	46.7	43.9	42.4	42.4	39.6
KNN [19]	45.9	45.9	45.8	59.9	60.1	59.0
Random Forest [20]	66.8	67.1	65.2	64.7	64.9	63.9
Proposed (Tensor Fusion Network)	**69.9**	**70.1**	**69.4**	**65.3**	**65.5**	**64.6**
Proposed (Cross-Attention)	**71.2**	**70.4**	**71.4**	-	-	-

TABLE II
COMPARISON WITH OTHER BASELINE MODELS

treated as a three class classification problem in which the aim is to predict the task type as requiring low, medium, or high mental effort. We trained our model for 500 iterations, employing an initial learning rate of 0.00005 and utilizing the Adam optimizer for the training process. Throughout the training, a batch size of 32 is employed. All experiments were conducted using the PyTorch framework. The source code for our system is available upon request.

B. Evaluation Metrics

Our models were evaluated using the following commonly used metrics: (1) Accuracy (Acc); (2) Recall; (3) F1-Score ; (4) AUC-ROC. Specifically, accuracy represents the percentage of correct predictions. Recall calculates the ratio of true positive predictions to the total number of actual positive instances, which quantifies a model's ability to correctly identify all positive instances. F1 score is a harmonic mean of the precision and recall. AUC-ROC is the area under the Receiver Operating Characteristic curve, which represents the degree or measure of separability. The best value of these three metrics is 100 (%), and 0 is the worst.

C. Comparison of Multimodal Fusion Methods

In this experiment, we conduct a comparison of classification performance between our proposed model and standard multimodal fusion approaches.

a) Early fusion: In the early fusion experiment, we first concatenate the eye-tracking feature and heart rate feature into a single unified feature representation. This combined feature then is passed through a two-layer convolutional network which performs the classification task.

b) Late fusion: In the late fusion experiment, we employ two separate two-layer convolutional networks to independently learn the features of the eye-tracking and heart rate modalities. Following the feature extraction process for each modality, we execute a weighted averaging procedure to combine the individual decisions derived from these two independent modality features to perform the final classification.

c) Joint fusion: In the joint fusion experiment, we utilize two separate two-layer convolutional networks to independently extract feature representations from the eye-tracking and heart rate modalities. However, different from late fusion, we integrate the two modalities at the feature level by concatenating their respective features. We then apply a fully connected layer followed by a softmax layer to derive the final classification outcomes.

As shown in Table I, the joint fusion model surpasses both early fusion and late fusion. This is reasonable since joint fusion leverages the strengths of both early and late fusion strategies. Furthermore, our proposed hybrid fusion model outperforms the joint fusion model by 2.5%. This illustrates the capability of our proposed model to represent both single modality features as well as synergistic relationships between different modalities, thus enhancing the discriminative power of the extracted features for workload classification. The dimensionality reduction properties of our architecture (e.g. max pooling) allow it to avoid performance degradation resulting from feature multicollinearity.

D. Comparison with Other Baseline Models

In this experiment, we compare the performance of the proposed model to the following baseline models: (1) Logistic Regression [17]; (2) Naive Bayes (Gaussian) [18]; (3) K-nearest Neighbor (KNN) [19]; (4) Random Forest [20]. We include results on relatively simple models such as logistic regression and naive Bayes since they are commonly used in the workload monitoring community due to their low number of parameters, high explainability, and fast inference speed [5], [6]. A thorough comparison is conducted with two inputs: a multimodal joint feature and a single eye-tracking feature. As shown in Table II, the random forest model outperforms the other baselines. This is unsurprising since random forest is known to be more robust to multicollinearity than regression models or naive Bayes. Our proposed hybrid tensor fusion network model surpasses the random forest by 3.1% and the cross-attention network outperforms hybrid tensor fusion by 1.3%. This demonstrates that our proposed cross-attention network is the best option for fusing bimodal data streams.

E. Ablation study

1) Fusion module: In this section, we conduct an ablation study to investigate the contribution of each fusion component in the proposed model. The results are presented in Table III. It can be observed that the proposed model obtains the best performance, signifying its effectiveness in leveraging both intra-modality and inter-modality learning. When we exclusively apply intra-modality-only or inter-modality-only learning, a decrease in performance is observed. This establishes the contribution of both intra-modality and inter-modality modules within the proposed model.

2) Input feature: In order to assess the impact of each modality on the overall performance, we conduct experiments with different inputs. From Table IV we can observe that eye-tracking features play a more important role in the cognitive recognition task than cardiovascular features. This

Model	Acc	Recall	F1	AUC-ROC
Joint Fusion (Intra-modality only)	67.4	67.5	66.7	83.9
TFN (Inter-modality only)	67.1	67.2	66.3	83.1
Proposed (Intra-modality + Inter-modality)	**69.9**	**70.1**	**69.4**	**84.0**

TABLE III
ABLATION STUDY ON FUSION MODULE.

Input feature	Acc	Recall	F1	AUC-ROC
Cardiovascular feature	43.3	43.3	43.2	62.5
Eye-tracking feature	63.9	64.1	62.7	81.8
Cardiovascular + Eye-tracking (Early fusion)	**67.4**	**67.5**	**66.7**	**83.9**

TABLE IV
ABLATION STUDY ON INPUT FEATURES.

occurs because the cardiovascular data collected from the PPG sensor can be noisy, particularly when the sensor lacks proper contact or adherence to the skin. However, training with both modalities achieves the best performance, proving the value of both feature sets.

IV. CONCLUSION

This paper introduces two network designs for fusion multimodal data: 1) a hybrid tensor fusion network and 2) a cross-attention transformer. Even though the tensor fusion network alone can represent unimodal embeddings, maintaining separate embeddings for each modality helps the model rapidly learn valuable representations from a small number of data points. Our proposed cross-attention transformer outperforms other options for learning multimodal embeddings.

We demonstrate that our methods are accurate when using noisy off the shelf biosensors and surpasses the most commonly used multimodal fusion paradigms and classifiers. The inference speed of our models are fast enough to predict the cognitive workload of human operators teloperating robots. Although our primary interest is workload modeling, we believe that the same approach can generalize to inferring affective states such as emotional stress [21] or mental fatigue [22].

V. ACKNOWLEDGMENTS

This research was supported with funding from Lockheed Martin Corporation.

REFERENCES

[1] Jianlong Zhou, Kun Yu, Fang Chen, Yang Wang, and Syed Z. Arshad, *Multimodal Behavioral and Physiological Signals as Indicators of Cognitive Load*, p. 287–329, Association for Computing Machinery and Morgan & Claypool, 2018.

[2] Essam Debie, Raul Fernandez Rojas, Justin Fidock, Michael Barlow, Kathryn Kasmarik, Sreenatha Anavatti, Matt Garratt, and Hussein A. Abbass, "Multimodal fusion for objective assessment of cognitive workload: A review," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1542–1555, 2021.

[3] Wonse Jo, Ruiqi Wang, Su Sun, Revanth Krishna Senthilkumaran, Daniel Foti, and Byung-Cheol Min, "MOCAS: A multimodal dataset for objective cognitive workload assessment on simultaneous tasks," 2022.

[4] S. Hart, "NASA-task load index (NASA-TLX): 20 years later," in *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, 2006, pp. 904–908.

[5] Fabio Dell'Agnola, Una Pale, Rodrigo Marino, Adriana Arza, and David Atienza, "MBioTracker: Multimodal self-aware bio-monitoring wearable system for online workload detection," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 5, pp. 994–1007, 2021.

[6] Fabio Dell'Agnola, Ping-Keng Jao, Adriana Arza, Ricardo Chavarriaga, José del R. Millán, Dario Floreano, and David Atienza, "Machine-learning based monitoring of cognitive workload in rescue missions with drones," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4751–4762, 2022.

[7] Abhishek Tiwari, Raymundo Cassani, Jean-François Gagnon, Daniel Lafond, Sébastien Tremblay, and Tiago H. Falk, "Movement artifact-robust mental workload assessment during physical activity using multi-sensor fusion," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 3471–3477.

[8] Haihong Zhang, Yongwei Zhu, Jayachandran Maniyeri, and Cuntai Guan, "Detection of variations in cognitive workload using multi-modality physiological sensors and a large margin unbiased regression machine," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 2985–2988.

[9] Yankai Wang, Bing Chen, Hongyan Liu, and Zhiguo Hu, "Understanding flow experience in video learning by multimodal data," *International Journal of Human-Computer Interaction*, pp. 1–15, 02 2023.

[10] Niraj Hirachan, Anita Mathews, Julio Romero, and Raul Fernandez Rojas, "Measuring cognitive workload using multimodal sensors," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2022, pp. 4921–4924.

[11] David Rozado and Andreas Dünser, "Combining EEG with pupillometry to improve cognitive workload detection," *Computer*, vol. 48, no. 10, pp. 18–25, 2015.

[12] Jiangchang Cheng, Yinglong Dai, Yao Yuan, and Hongli Zhu, "A simple analysis of multimodal data fusion," in *IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2020, pp. 1472–1475.

[13] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[14] SC. Huang, A. Pareek, R. Zamanian, et al., "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection," *Scientific Reports*, vol. 10, no. 22147, 2020.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] EH Siegel, J Wei, A Gomes, M Oliviera, P Sundaramoorthy, K Smathers, M Vankipuram, S Ghosh, H Horii, J Bailenson, et al., "HP Omnicept cognitive load database (HPO-CLD)–developing a multimodal inference engine for detecting real-time mental workload in VR," Tech. Rep., Technical Report, 2021.

[17] Raymond E Wright, "Logistic regression.," 1995.

[18] Harry Zhang, "The optimality of naive Bayes," in *Proceedings of the Florida AI Research Society*, 2004.

[19] Leif E Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, pp. 1883, 2009.

[20] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[21] Behnaz Jafari, Kenneth Lai, and Svetlana Yanushkevich, "Investigating association and causal relationships between physiological signals and affective state," in *International Conference on Information and Digital Technologies (IDT)*, 2023, pp. 243–250.

[22] Anping Song, Chaoqun Niu, Xuehai Ding, Xiaokang Xu, and Ziheng Song, "Mental fatigue prediction model based on multimodal fusion," *IEEE Access*, vol. 7, pp. 177056–177062, 2019.