M&M: Multimodal-Multitask Model Integrating Audiovisual Cues in Cognitive Load Assessment

Long Nguyen-Phuoc[1,2] [a], Renald Gaboriau[2] [b], Dimitri Delacroix[2] [c] and Laurent Navarro[1] [d]

[1]*Mines Saint-Étienne, University of Lyon, University Jean Monnet, Inserm, U 1059 Sainbiose, Centre CIS, 42023 Saint-Étienne, France*

[2]*MJ Lab, MJ INNOV, 42000 Saint-Etienne, France*

Keywords: Cognitive Load Assessment, Multimodal-Multitask Learning, Multihead Attention.

Abstract: This paper introduces the M&M model, a novel multimodal-multitask learning framework, applied to the AVCAffe dataset for cognitive load assessment (CLA). M&M uniquely integrates audiovisual cues through a dual-pathway architecture, featuring specialized streams for audio and video inputs. A key innovation lies in its cross-modality multihead attention mechanism, fusing the different modalities for synchronized multitasking. Another notable feature is the model's three specialized branches, each tailored to a specific cognitive load label, enabling nuanced, task-specific analysis. While it shows modest performance compared to the AVCAffe's single-task baseline, M&M demonstrates a promising framework for integrated multimodal processing. This work paves the way for future enhancements in multimodal-multitask learning systems, emphasizing the fusion of diverse data types for complex task handling.

1 INTRODUCTION

In the dynamic field of cognitive load assessment (CLA), understanding complex mental states is crucial in diverse areas like education, user interface design, and mental health. Cognitive load refers to the effort used to process information or to perform a task and can vary depending on the complexity of the task and the individual's ability to handle information (Block et al., 2010). In educational psychology, cognitive load, essential for instructional design, reflects the mental demands of tasks on learners (Paas and van Merriënboer, 2020). It suggests avoiding working memory overload for effective learning (Young et al., 2014). In user interface design, it involves minimizing user's mental effort for efficient interaction (Group, 2013). In mental health, it relates to cognitive tasks' mental workload, significantly impacting those with mental health disorders or dementia (Beecham et al., 2017).

CLA encompasses diverse traditional methods. Dual-Task Methodology in multimedia learning and Subjective Rating Scales in education evaluate cognitive load, requiring validation in complex contexts (Thees et al., 2021). Wearable devices and Passive Brain-Computer Interfaces offer real-time and continuous monitoring (Jaiswal et al., 2021; Gerjets et al., 2014). Hemodynamic Response Analysis and linguistic behavioral analysis provide accurate assessment in specific tasks (Ghosh et al., 2019; Khawaja et al., 2014). Mobile EEG and physiological data analysis further contribute to multimodal measurement strategies (Kutafina et al., 2021; Vanneste et al., 2020).

Recent progress in assessing cognitive load through machine learning and deep learning has been notable. Neural networks, particularly under deep learning frameworks, have achieved accuracies up to 99%, with artificial neural networks and support vector machines being key techniques (Elkin et al., 2017). Deep learning, especially models like stacked denoising autoencoders and multilayer perceptrons, have outperformed traditional classifiers in estimating cognitive load (Saha et al., 2018). Enhanced methods have shown impressive results in classifying mental load using EEG data, comparing favorably with CNNs (Jiao et al., 2018; Kuanar et al., 2018) and RNN (Kuanar et al., 2018). Additionally, machine learning has been effective in detecting cognitive load

[a] https://orcid.org/0000-0003-1294-8360
[b] https://orcid.org/0000-0001-5565-5088
[c] https://orcid.org/0009-0002-2143-7343
[d] https://orcid.org/0000-0002-8788-8027

869

states using signals like ECG and EMG (Oschlies-Strobel et al., 2017) or PPG (Zhang et al., 2019). Although these study highlight the growing role of machine learning in accurately assessing cognitive load, they often fall short in capturing the multifaceted nature of cognitive load, necessitating multimodal learning.

Our M&M model is designed to accurately assess cognitive load across multiple contexts, effectively leveraging multimodal inputs and multitask learning. Recognizing the intricate nature of cognitive load, we integrate audiovisual data to capture a comprehensive picture of cognitive state. The primary contributions of our work are:

- Incorporation of multimodal data fusion: By leveraging the complementary of audio and video data that mirrors human sensory observation, M&M ensures a comprehensive and non-intrusive capture of cognitive load.

- Implementation of multitask learning: This approach not only simplifies the training process by jointly learning various aspects of cognitive load but also improves the overall accuracy and robustness of the model.

2 RELATED WORKS

2.1 Multimodal Learning for CLA

The field of CLA has witnessed significant advancements through the creation of multimodal datasets and dedicated research in multimodal learning.

Dataset creation initiatives (Mijić et al., 2019; Gjoreski et al., 2020; Oppelt et al., 2022; Sarkar et al., 2022) prioritize the development of robust, annotated datasets that capture a range of modalities, including physiological signals, facial expressions, and environmental context. These datasets serve as critical benchmarks for evaluating and training CLA models, ensuring that researchers have access to high-quality, diverse data sources for algorithm development.

In contrast, studies focused on multimodal learning in CLA leverage existing datasets to investigate and refine techniques for integrating and interpreting data from multiple sources. (Chen, 2020) estimates task load levels and types from eye activity, speech, and head movement data in various tasks using event intensity, duration-based features, and coordination-based event features. (Cardone et al., 2022) evaluates drivers' mental workload levels using machine learning methods based on ECG and infrared thermal signals, advancing traffic accident prevention. (Daza

et al., 2023) introduces a multimodal system using Convolutional Neural Networks to estimate attention levels in e-learning sessions by analyzing facial gestures and user actions.

2.2 Multitask Learning for CLA

Recent advancements in multitask learning (MTL) neural networks could have significantly impacted CLA. These advancements include: improved methods (Ruder, 2017), personalized techniques (Taylor et al., 2020), data efficiency and regularization (Søgaard and Bingel, 2017).

Moreover, cutting-edge developments, such as the introduction of adversarial MTL neural networks, have further propelled this field. These networks are designed to autonomously learn task relation coefficients along with neural network parameters, a feature highlighted in (Zhou et al., 2020). However, despite these advancements, (Gjoreski et al., 2020) remains one of the few studies demonstrating the superiority of MTL networks over single-task networks in the realm of CLA. This indicates the potential yet unexplored in fully harnessing the capabilities of MTL in complex and nuanced areas like cognitive load assessment.

2.3 Multimodal-Multitask Learning for CLA

In our exploration of the literature, we found no specific studies on Multimodal-Multitask Neural Networks tailored for cognitive load assessment. This gap presents an opportunity for pioneering research. Therefore, in this section, we broaden our focus to the wider domain of cognitive assessment.

(Tan et al., 2021) developed a bioinspired multisensory neural network capable of sensing, processing, and memorizing multimodal information. This network facilitates crossmodal integration, recognition, and imagination, offering a novel approach to cognitive assessment by leveraging multiple sensory inputs and outputs. (El-Sappagh et al., 2020) presented a multimodal multitask deep learning model specifically designed for detecting the progression of Alzheimer's disease using time series data. This model combines stacked convolutional neural networks and BiLSTM networks, showcasing the effectiveness of multimodal multitask approaches in tracking complex mental health conditions. (Qureshi et al., 2019) demonstrated a network that improves performance in estimating depression levels through multitask representation learning. By fusing all modalities, this network outperforms single-task networks,
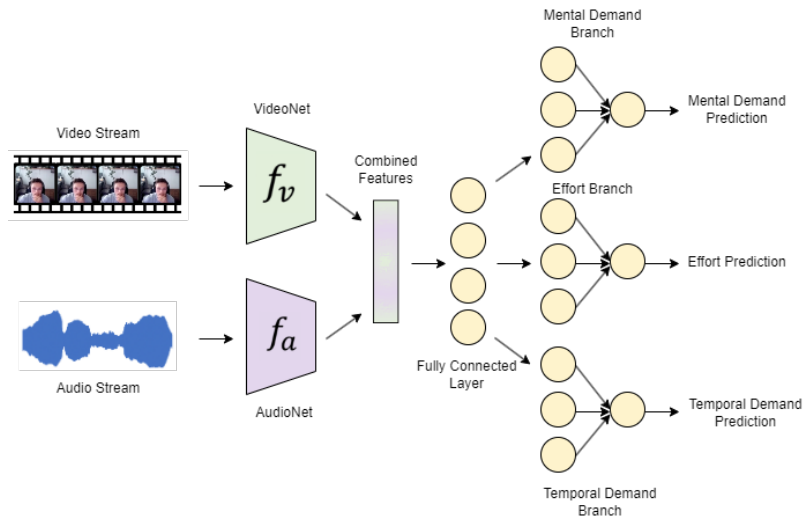
Figure 1: The M&M model's architecture.

highlighting the benefits of integrating multiple data sources for accurate cognitive assessment.

3 METHODS

Our motivation for developing the M&M model stems from several key objectives. Firstly, it aims to offer a compact and efficient AI solution, easing training demands in settings with limited computational power. In the context of human-robot interaction, M&M simplifies deployment by unifying multiple labels and inputs into one model, reducing the infrastructure needed for effective operation. Lastly, it confirms the interest of multimodality for all studies related to the analysis of information provided by human beings (Kress, 2009).

3.1 Signal Processing

3.1.1 Audio Processing

The preprocessing stage consist first in downsampling the audio stream at 16 kHz. The waveform is then zero-padded or truncated to match the same length. Afterward, the mel-spectrogram (Arias-Vergara et al., 2021) of the segments are computed using $n_{\mathrm{mels}} = 80$ mel filters with a Fourier transform window size of 1024 points. The hop length, which determines the stride for the sliding window, is set to 1% of the target sample rate, effectively creating a 10 ms hop size between windows. The MelSpectrogram transformation is particularly suitable for audio processing because it mirrors human auditory perception more closely than a standard spectrogram.

3.1.2 Video Processing

The visual stream is downsampled at 5 frames per second and resized to a spatial resolution of of 168×224 pixels, reducing computational load for faster training. Zero padding and truncating are also needed to obtain same length. After this step, all frames are transformed in the following pipeline. Initially, video frames are randomly flipped horizontally half the time, imitating real-world variations. A central portion of each frame is then cropped, sharpening the focus on key visual elements. Converting frames to grayscale emphasizes structural details over color, reducing the model's processing load and bias due to color temperatures of different videos. Finally, normalization, with means and standard deviations both set to 0.5, ensures the pixel values to be in the comparable range of [0, 1] which stabilizes training and reduces skewness.

3.2 Model Architecture

The M&M's architecture, illustrated in Figure 1, can be segmented into four principal components, each contributing uniquely to the overall functionality of the model.

3.2.1 AudioNet

AudioNet, a convolutional neural network architecture (Wang et al., 2023) designed for processing audio data, accepts as input a Mel spectrogram, represented by the tuple (n_{mels}, seq_a). Here, n_{mels} denotes the number of mel filters, and seq_a is the length of the audio tensor obtained after the transformation of raw audio input as described in Section 3.1.1. Pre-

	Participant A			Participant B		
	Open discussion	Montclair map	Multi-task	Open discussion	Montclair map	Multi-task
Effort	8	17	16	2	16	12
Mental demand	3	12	13	4	12	15
Temporal demand	0	20	11	3	16	6

Figure 2: Some examples of the AVCAffe dataset. The self-reported cognitive score shown here at the NASA-TLX scale.

cisely, AudioNet consists of four convolutional layers, each followed by batch normalization. The convolutional layers progressively increase the number of filters from 16 to 128, extracting hierarchical features from the audio input. A max pooling layer follows the convolutional blocks to reduce the spatial dimensions of the feature maps. The network includes a dropout layer set at 0.5 to prevent overfitting. The flattened output from the convolutional layers is then fed into a fully connected layer with 128 units. The AudioNet uses ReLU activation functions and is capable of end-to-end training, transforming raw audio input into a compact representation.

3.2.2 VideoNet

VideoNet in the M&M model, based on the I3D backbone (Carreira and Zisserman, 2018), is a sophisticated neural network tailored for visual data processing. It commences with sequential 3D convolutional layers, succeeded by max-pooling layers, instrumental in feature extraction and dimensionality reduction. The network's essence lies in its Inception modules, adept at capturing multi-scale features. These modules amalgamate distinct convolutional and pooling branches, facilitating a comprehensive feature analysis. Post-feature extraction, an adaptive average pooling layer condenses the data, culminating in a fully connected layer with 128 units after dropout and linear transformation. The network takes as input video data represented by the tuple (d, h, w), where h and w reflecting the post-processed video spatial dimensions and d the depth which is the product of clip length and target frame rate as explained in Section 3.1.2.

3.2.3 Crossmodal Attention

Cross-modality attention is a concept derived from the attention mechanism (Phuong and Hutter, 2022). It has been adapted for scenarios where the model must attend to and integrate information from multiple different modalities, such as audio, text, and images (Tsai et al., 2019). Consider two modalities,

Audio and Video, the cross-modality attention can be represented as:

$$\text{Attention}(Q_v, K_a, V_a) = \text{softmax}\left(\frac{Q_v K_a^T}{\sqrt{d_k}}\right) V_a \quad (1)$$

In this structure, the attention mechanism takes video features as the query Q_v and audio features as both key K_a and value V_a. The attention module inputs features from both audio and video modalities, each with a dimension of 128. It outputs a combined feature set, effectively integrating relevant features from both sources for enhanced processing.

3.2.4 Multitask Separated Branches

After processed by a common fully connected layer, the combined features $\hat{AV}$ is distributed to three distinct branches, each tailored to a specific cognitive load task. These branches employ individual sigmoid activation functions in their output layers to predict binary classifications for each task, facilitating a probabilistic interpretation of the model's predictions.

The flexibility to adjust weights or modify the architecture asymmetrically for each branch permits the model to be customized based on the complexity or priority of tasks, hence optimizing performance and ensuring tailored learning (Nguyen et al., 2020).

However, in a multi-branch neural network with shared layers, the learning dynamics in one branch can significantly influence those in others. This interdependence enhances the model's overall learning capability, as advancements in one branch may depend on simultaneous updates in others. Such interconnection not only highlights the collaborative nature of the learning process across different branches but also fosters a richer information exchange (Fukuda et al., 2018). This synergy can lead to more robust and comprehensive learning outcomes, leveraging shared insights for improved performance in each task-specific branch.

3.3 Implementation Details

The central M&M architecture leverages gradients from three specialized branches—each corresponding to a unique cognitive load task—to train shared neural network layers. This holistic training approach is facilitated by the Adam optimizer for efficient stochastic gradient descent. The model built with PyTorch's fundamental components, undergoes an end-to-end training process, allowing simultaneous learning across the diverse modalities of audio and visual data, ensuring distinct tasks of CLA.

The Binary Cross-Entropy (BCE) Loss function is utilized individually for each branch to cater to the binary classification nature of our tasks. The BCE loss for a single task is expressed as:

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1-y_i)\cdot\log(1-p_i)] \quad (2)$$

The global loss is a weighted sum, allowing us to prioritize tasks asymmetrically based on their complexity or importance:

$$L_{global} = \sum_{k=1}^{K} w_k \cdot L_{bce}^k \quad (3)$$

In these equations, N is the number of samples, y_i is the true label, p_i is the predicted probability for the ith sample, K is the total number of tasks, w_k is the weight for the kth task, and L_{bce}^k is the BCE loss for the kth task. The weights w_k are adjustable to focus the model's learning on specific tasks based on their difficulty or relevance.

Furthermore, the cross-modal attention mechanism is implemented as explicitly described in Algorithm 1, ensuring a cohesive integration of audio and video data streams.

4 EXPERIMENT

In this section, extensive experimentation illustrates the compact and novel character of M&M. We first introduce the AVCAffe dataset Section 4.1. Then, the experiment setup, including the criterion metrics and the hyper-parameters setting situation, is presented in Section 4.2. In Section 4.3, we compare with the dataset's authors baseline model.

4.1 AVCAffe

The AVCAffe dataset (Sarkar et al., 2022) offers rich audio-visual data to study cognitive load in remote

Data: Audio features A, Video features V
Result: Combined features set $\hat{AV}$
Hyperparameters: H, number of attention heads
Parameters: W^Q, W^K, W^V, W^O weight matrices for query, key, value, and output projections.
for $h \in \{1\dots H\}$ **do**
   $Q_h = AW_h^Q$;
   $K_h = VW_h^K$;
   $V_h = VW_h^V$;
   Compute scaled dot-product attention
   head$_h$ = Attention(Q_h, K_h, V_h);
   $\hat{AV}$ = Concat(head$_1,\dots,$head$_H)W^O$
end

Algorithm 1: Cross-Modal Multihead Attention Mechanism.

work, with 108 hours of footage from a globally diverse participant group. It explores the cognitive impact of remote collaboration through task-oriented video conferencing, employing NASA-TLX for multidimensional cognitive load measurement. Challenges in detailed classification led to the adoption of a binary approach. For model validation and to avoid data leakage, we ensured representative splits and attempted to emulate the authors' data augmentation methods as detailed in Table 1, despite the lack of public implementations.

Table 1: The parameter details of audio-visual augmentations.

Audio Augmentation	
Volume Jitter	range=±0.2
Time Mask	max size=50, num=2
Frequency Mask	max size=50, num=2
Random Crop	range=[0.6,1.5], crop scale=[1.0,1.5]
Visual Augmentation	
Multi-scale Crop	min area=0.2
Horizontal Flip	p=0.5
Color Jitter	b=1.0, c=1.0, s=1.0, h=0.5
Gray Scale	p=0.2
Cutout	max size=50, num=1

4.2 Setup

4.2.1 Criterion Metrics

Following the metric construction in the (Sarkar et al., 2022) that publushes the AVCAffe dataset, we employ weighted F1-measures for evaluation to counterbalance imbalanced class distribution, a common is-

Table 2: CLA results (F1-Score) on AVCAffe validation set (Binary Classification).

Modalities	Architecture	Global F1 score	Mental Demand	Effort	Temporal Demand
(Sarkar et al., 2022)					
Audio	ResNet+MLP	–	0.61	0.55	0.55
Video	R(2+1)D+MLP	–	0.59	0.55	0.55
Audio-Video	Multimodal-Single Task	–	0.62	0.61	0.59
Ours					
Audio	CNN+MLP+Multitask	0.58	0.56	0.58	0.56
Video	I3D+MLP+Multitask	0.48	0.44	0.49	0.48
Audio-Video	Multimodal-Multitask (M&M)	0.59	0.60	0.61	0.54

sue in cognitive datasets (Kossaifi et al., 2019; Busso et al., 2008). We also calculates a global micro-averaged F1 score, combining all labels ⟨*Mental demand, Effort, Temporal demand*⟩ to provide an overall assessment of the model's performance across all tasks, considering both precision and recall in a unified metric. The formula of the F1-Scores is shown below:

$$\text{Weighted F1-score} = \frac{\sum_{i=1}^{N} w_i \times \frac{2 \times \text{TP}_i}{2 \times \text{TP}_i + \text{FP}_i + \text{FN}_i}}{\sum_{i=1}^{N} w_i} \quad (4)$$

$$\text{Global Micro-Averaged F1} =$$

$$2 \times \frac{\sum_{\text{labels}} \text{TP}_{\text{labels}}}{\sum_{\text{labels}} \text{TP}_{\text{labels}} + \frac{1}{2}\left(\sum_{\text{labels}} \text{FP}_{\text{labels}} + \sum_{\text{labels}} \text{FN}_{\text{labels}}\right)} \quad (5)$$

In the given equations, N represents the total number of samples, with i indexing each sample, and w_i denoting the weight associated with the ith sample. TP_i, FP_i, and FN_i correspond to the True Positives, False Positives, and False Negatives for the ith sample, respectively. The term 'Labels' is defined as the set of all labels under consideration, specifically ⟨*Mental demand, Effort, Temporal demand*⟩. TP_{labels}, FP_{labels}, and FN_{labels} represent the sums of True Positives, False Positives, and False Negatives across these labels.

4.2.2 Hyperparameters

In this study, several hyperparameters are meticulously chosen to ensure an optimal balance between training efficiency and model accuracy. The learning rate is set at 0.001, utilizing the Adam optimizer for gradual and precise model updates. A step learning rate scheduler adjusts the learning rate every 10 epochs by a factor of 0.1, assisting more effective convergence and overcoming local minima. Training spans 30 epochs on 1 GPU NVIDIA T4 16GB, providing sufficient iterations for multimodal-multitask learning. Additionally, early stopping is implemented

with a patience of 10 epochs to halt training when no improvement in validation loss is observed, ensuring training efficiency and preventing overfitting.

For the crossmodal attention using PyTorch multi-head attention component, the number of head is fixed at 4.

Moreover, due to limited computational resources, we utilized randomly only 25 clips of 6 seconds each per participant for each task, aligning with the dataset's average clip length. Thus, the final input dimension becomes $(n_{mels} = 80, seq_a = 601)$ for AudioNet and $(d = 30, h = 148, w = 144)$ for VideoNet. This resulted in dataset partitions containing 15,213 clips for training, 3,804 for validation, and 4,474 for testing. These clips are subsequently grouped into batches of 256 for model training and evaluation.

4.3 Experimental Results and Comparison

We report the experimental results in Table 2 which outlines a comparative analysis of different architectures applied to the AVCAffe dataset for binary classification of cognitive load, indicated by F1-Score metrics across three task domains: Mental Demand, Effort, and Temporal Demand. It compares results from (Sarkar et al., 2022) with those obtained from our proposed M&M model.

Overall, the F1 scores obtained range from 0.44 to 0.62 reflecting the CLA complexity. Low scores for temporal demand suggest that the task is more challenging or the data is less qualitative than the others.

(Sarkar et al., 2022) separate models used to process audio, video or both modalities. The highest score achieved is for Mental Demand (0.62) using a audio-visual late fusion approach in a single-task framework. However, no global F1 score was reported, suggesting that the authors chose to focus on task-specific performance rather than an aggregated metric across all tasks.

The M&M model shows competitive results, particularly in the Mental Demand and Effort categories,

where it outperforms the single-modality approaches and is on par with or slightly below the (Sarkar et al., 2022) multimodal single-task model. However, it appears to have a lower score in the Temporal Demand, because even if it achieves improved overall task-average performance, they may still yield degraded performance on Temporal Demand individual. Such behavior conforms the finding of the literature (Nguyen et al., 2020).

These results could indicate that while our M&M model has a balanced performance across tasks and benefits from the integration of audio-visual data, there might be room for optimization, especially in an asymmetric temporal demand branch. This also suggests that M&M can provide a more compact and computationally efficient solution without significantly sacrificing performance, which can be particularly advantageous in resource-constrained environments or for real-time applications.

5 CONCLUSIONS

In summary, the M&M model is designed to extract and integrate audio and video features using separate streams, fuse these features using a crossmodal attention mechanism, and then apply this integrated representation to multiple tasks through separate branches. This structure enables the handling of complex multimodal data and supports multitasking learning objectives.

The M&M model bridges a crucial research gap by intertwining multitask learning with the fusion of audio-visual modalities, reflecting the most instinctive human observation methods. This unique approach to AI development sidesteps reliance on more invasive sensors, favoring a naturalistic interaction style.

Moving forward, our research will expand to experimenting with various affective computing datasets using the proposed M&M model. This will deepen our understanding of how different tasks interact within a multimodal-multitask framework. A key focus will be the development of a tailored weighted loss function, which will be designed based on the correlations observed between these tasks. This innovative approach aims to refine and optimize the model's performance by aligning it more closely with the nuanced relationships inherent in cognitive processing tasks.

REFERENCES

Arias-Vergara, T., Klumpp, P., Vasquez-Correa, J., et al. (2021). Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications*, 24:423–431.

Beecham, R. et al. (2017). The impact of cognitive load theory on the practice of instructional design. *Educational Psychology Review*, 29(2):239–254.

Block, R., Hancock, P., and Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, 134(3):330–343.

Busso, C. et al. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Cardone, D., Perpetuini, D., Filippini, C., Mancini, L., Nocco, S., Tritto, M., Rinella, S., Giacobbe, A., Fallica, G., Ricci, F., Gallina, S., and Merla, A. (2022). Classification of drivers' mental workload levels: Comparison of machine learning methods based on ecg and infrared thermal signals. *Sensors (Basel, Switzerland)*, 22.

Carreira, J. and Zisserman, A. (2018). Quo vadis, action recognition? a new model and the kinetics dataset.

Chen, S. (2020). Multimodal event-based task load estimation from wearables. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

Daza, R., Gomez, L. F., Morales, A., Fierrez, J., Tolosana, R., Cobos, R., and Ortega-Garcia, J. (2023). Matt: Multimodal attention level estimation for e-learning platforms. *ArXiv*, abs/2301.09174.

El-Sappagh, S., Abuhmed, T., Islam, S. M., and Kwak, K. (2020). Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data. *Neurocomputing*, 412:197–215.

Elkin, C. P., Nittala, S. K. R., and Devabhaktuni, V. (2017). Fundamental cognitive workload assessment: A machine learning comparative approach. In *2017 Conference Proceedings*, pages 275–284. Springer.

Fukuda, S., Yoshihashi, R., Kawakami, R., You, S., Iida, M., and Naemura, T. (2018). Cross-connected networks for multi-task learning of detection and segmentation. *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3636–3640.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience*, 8.

Ghosh, L., Konar, A., Rakshit, P., and Nagar, A. (2019). Hemodynamic analysis for cognitive load assessment and classification in motor learning tasks using type-2 fuzzy sets. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3:245–260.

Gjoreski, M., Kolenik, T., Knez, T., Luštrek, M., Gams, M., Gjoreski, H., and Pejović, V. (2020). Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences*.

Group, N. N. (2013). Minimize cognitive load to maximize usability. https://www.nngroup.com/articles/minimize-cognitive-load/.

Jaiswal, D., Chatterjee, D., Gavas, R., Ramakrishnan, R. K., and Pal, A. (2021). Effective assessment of cognitive load in real-world scenarios using wrist-worn sensor data. In *Proceedings of the Workshop on Body-Centric Computing Systems*.

Jiao, Z., Gao, X., Wang, Y., Li, J., and Xu, H. (2018). Deep convolutional neural networks for mental load classification based on eeg data. *Pattern Recognition*, 76:582–595.

Khawaja, M. A., Chen, F., and Marcus, N. (2014). Measuring cognitive load using linguistic features: Implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction*, 30:343–368.

Kossaifi, J. et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *Transactions on Pattern Analysis and Machine Intelligence*.

Kress, G. (2009). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. Routledge, London ; New York.

Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., and Rao, K. (2018). Cognitive analysis of working memory load from eeg, by a deep recurrent neural network. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2576–2580.

Kutafina, E., Heiligers, A., Popovic, R., Brenner, A., Hankammer, B., Jonas, S., Mathiak, K., and Zweerings, J. (2021). Tracking of mental workload with a mobile eeg sensor. *Sensors (Basel, Switzerland)*, 21(15).

Mijić, I., Šarlija, M., and Petrinović, D. (2019). Mmodcog: A database for multimodal cognitive load classification. *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 15–20.

Nguyen, T., Jeong, H., Yang, E., and Hwang, S. J. (2020). Clinical risk prediction with temporal probabilistic asymmetric multi-task learning. *ArXiv*, abs/2006.12777.

Oppelt, M. P., Foltyn, A., Deuschel, J., Lang, N., Holzer, N., Eskofier, B. M., and Yang, S. H. (2022). Adabase: A multimodal dataset for cognitive load estimation. *Sensors (Basel, Switzerland)*.

Oschlies-Strobel, A., Gruss, S., Jerg-Bretzke, L., Walter, S., and Hazer-Rau, D. (2017). Preliminary classification of cognitive load states in a human machine interaction scenario. In *2017 International Conference on Companion Technology (ICCT)*, pages 1–5. IEEE.

Paas, F. and van Merriënboer, J. (2020). Cognitive load theory: A broader view on the role of memory in learning and education. *Educational Psychology Review*, 32:1053–1072.

Phuong, M. and Hutter, M. (2022). Formal algorithms for transformers.

Qureshi, S. A., Saha, S., Hasanuzzaman, M., Dias, G., and Cambria, E. (2019). Multitask representation learning

for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34:45–52.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098.

Saha, A., Minz, V., Bonela, S., Sreeja, S. R., Chowdhury, R., and Samanta, D. (2018). Classification of eeg signals for cognitive load estimation using deep learning architectures. In *2018 Conference Proceedings*, pages 59–68. Springer.

Sarkar, S., Chatterjee, J., Chakraborty, S., and Ganguly, N. (2022). Avcaffe: A large scale audio-visual dataset of cognitive load in remote work environments.

Søgaard, A. and Bingel, J. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 164–169.

Tan, H., Zhou, Y., Tao, Q., Rosen, J., and van Dijken, S. (2021). Bioinspired multisensory neural network with crossmodal integration and recognition. *Nature Communications*, 12.

Taylor, S., Jaques, N., Nosakhare, E., Sano, A., and Picard, R. W. (2020). Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 11:200–213.

Thees, M., Kapp, S., Altmeyer, K., Malone, S., Brünken, R., and Kuhn, J. (2021). Comparing two subjective rating scales assessing cognitive load during technology-enhanced stem laboratory courses. *Frontiers in Education*, 6.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Acker, B. V. V., Larmuseau, C., Depaepe, F., and den Noortgate, W. V. (2020). Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work*, 23:567–585.

Wang, Z., Bai, Y., Zhou, Y., and Xie, C. (2023). Can cnns be more robust than transformers?

Young, J. Q., ten Cate, O., Durning, S., and van Gog, T. (2014). Cognitive load theory: Implications for medical education: Amee guide no. 86. *Medical Teacher*, 36(5):371–384.

Zhang, X., Lyu, Y., Qu, T., Qiu, P., Luo, X., Zhang, J., Fan, S., and Shi, Y. (2019). Photoplethysmogram-based cognitive load assessment using multi-feature fusion model. *ACM Transactions on Applied Perception (TAP)*, 16:1 – 17.

Zhou, F., Shui, C., Abbasi, M., Émile Robitaille, L., Wang, B., and Gagné, C. (2020). Task similarity estimation through adversarial multitask neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 32:466–480.