Assessment of Mental Workload Using a Transformer Network and Two Prefrontal EEG Channels: An Unparameterized Approach

Matin Beiramvand, *Graduate Student Member, IEEE*, Mohammad Shahbakhti, *Graduate Student Member, IEEE*, Nina Karttunen, Reijo Koivula, Jari Turunen, and Tarmo Lipping, *Senior Member, IEEE*

***Abstract*— Despite promising results reported in the literature for mental workload assessment using electroencephalography (EEG), most of the proposed methods rely on employing multiple EEG channels, limiting their practicality. However, the advent of wearable EEG technology provides the possibility of mental workload assessment for real-life applications. Yet, a few studies that considered consumer-oriented EEG headsets for mental workload assessment only used a single database for validating the proposed methods, overlooking the potential for portability. In this research, we studied 60 recordings of participants playing a three-level n-back game, utilizing data from two EEG devices, Enobio and Muse, with distinctive characteristics such as sampling rate and channel configuration. Following the denoising of the EEG signals, we segmented the signals and applied the discrete wavelet transform (DWT) to decompose them into subbands. Then, we extracted Shannon entropy (SE) and wavelet log energy (WLE) features from all subbands. Subsequently, we fed the extracted features into five classifiers: support vector machine (SVM), *k*-nearest neighbors (kNNs), multilayer perceptron (MLP), AdaBoost, and the transformer network (TN). In comparing the results across all classifiers, the TN demonstrated superiority by achieving highest mean accuracy for Database M (88%) and Database E (85%). Given the consistent outcomes achieved with the TN classifier across both databases and utilizing a three-level n-back game, our findings indicate that the proposed method holds promise for real-life applications.**

***Index Terms*— Energy, entropy, mental workload, transformer network (TN), wearable EEG device.**

I. Introduction

MENTAL workload describes the level of mental resources utilized when performing a task. Overloading the task can lead to chronic stress [1], whereas performing below one's capabilities can cause frustration [2]. Emerging advances in wearable medical equipment and machine learning methods have facilitated monitoring cognitive states such as mental workload, level of engagement, flow state, mental stress, drowsiness, and sleep stages.

Three methods can be employed to assess the mental workload: self-reporting questionnaires, measuring task performance, and physiological biomarkers [1]. The most common self-reporting techniques are the NASA task load index (TLX) [3] and the subjective workload assessment technique (SWAT) [4]. Self-reporting techniques are easy to use and provide information about the perceived workload. However, these tests are not suitable for real-time continuous-scale measurement of mental workload; they do not discriminate between the difficulty of the task and the workload and account only for consciously perceived workload [5].

Performance measures include variables such as response time, task completion time, task accuracy, or error rate [5]. The relationship between these measures and the mental workload is difficult to establish and depends mainly on the specific task type and the subject's previous experience. Therefore, performance measures are only secondary indicators of mental workload.

Physiological and neurophysiological biomarkers can provide objective real-time information about the mental workload on a continuous scale. Longo et al. [1] categorize the mental load biomarkers according to the physiological source of the underlying variable. Certain biomarkers derived from electrocardiac variables, like heart rate or blood pressure [6], respiratory variables including respiratory rate, ocular variables such as eye blinking rate or pupil size, and skin variables like skin temperature or impedance [7], have been extensively explored in existing literature. The authors note, however, that the most significant number of studies focus on biomarkers based on neurophysiological variables such as the electroencephalogram (EEG) or functional near-infrared spectroscopy (fNIRS). This is well justified as the central nervous system is the primary target of mental workload [2].

Manuscript received 14 November 2023; revised 1 March 2024; accepted 10 April 2024. Date of publication 30 April 2024; date of current version 10 May 2024. The Associate Editor coordinating the review process was Dr. Chengyu Liu. *(Corresponding author: Matin Beiramvand.)*

Matin Beiramvand, Reijo Koivula, Jari Turunen, and Tarmo Lipping are with the Faculty of Information Technology and Communication Sciences (ITC), Tampere University, 28100 Pori, Finland (e-mail: matin.beiramvand@tuni.fi).

Mohammad Shahbakhti is with the Biomedical Engineering Institute, Kaunas University of Technology, 44249 Kaunas, Lithuania.

Nina Karttunen is with the Research Center RoboAI, Satakunta University of Applied Sciences, 28101 Pori, Finland.

Digital Object Identifier 10.1109/TIM.2024.3395312

Among the neurophysiological variables, EEG is the easiest to acquire while, at the same time, providing excellent temporal resolution [8]. Therefore, EEG-based biomarkers have been intensively studied to assess and monitor mental workload [9], [10], [11]. For example, Zarjam et al. [12], [13] using 32-channel EEG, employed an artificial neural network (ANN) to distinguish between seven difficulty levels of arithmetic tasks. Deep learning techniques have recently gained popularity in mental workload assessment [14]. Two kinds of 21-channel EEG dataset arrangements were used in [15] to discriminate between four levels of cognitive load using a CNN with four convolutional and two fully connected layers. Apart from the efficiency of transformer network (TN) architectures in natural language processing tasks, it has been applied to discriminate between levels of mental workload [16].

Despite promising results reported in the literature regarding mental workload assessment using EEG, the majority of studies have utilized multichannel EEG data, which increases the wearable complexity. Moreover, this configuration requires coverage over hair-bearing areas of the scalp, making it more susceptible to noise and interference [2]. Consequently, these factors have limited its practical application in real-life scenarios.

Nonetheless, the development of wearable consumer-oriented easy-to-use EEG devices has opened up a new avenue for the detection of mental workload in real-life environments [17]. Utilizing these systems to monitor mental workload can address the limitations associated with using a large number of EEG channels. Yet, despite their potential, these systems have not garnered enough attention. To the best of the authors' knowledge, only a few studies used such systems for mental workload assessment.

Almogbel et al. [18] fed four EEG signals recorded by the Muse headband to a CNN of eight convolutional layers to discriminate between three levels of cognitive load in a driving simulator environment. Based on windows of different lengths, varied accuracies were reported. Arslan et al. [19] utilized the same EEG data recording system, extracting multiple features from the theta band, which were then input into a multilayer perceptron (MLP) classifier. Three different levels of mental workload were classified using leave-one-out cross-validation (CV). Wang et al. [20] utilized the Emotiv EPOC EEG headset, recording data from 14 channels across various brain regions, to classify three distinct levels of mental workload. With a sample size of only nine subjects, the authors detected different levels of mental workload. Liu et al. [21] used the EMOTIV INSIGHT 1.0 headset, which records data from 5 EEG channels, to classify the mental workload of pilots. Their findings indicated that the highest accuracy was attained by employing power spectral density features fed into a k-nearest neighbors (kNNs) classifier. So et al. [22] utilized the Neurosky MindWave EEG headset to record data solely from the Fp1 channel. These data were used to evaluate the mental workload of 20 subjects during four cognitive and motor tasks. A support vector machine (SVM) model was employed for leave-one-subject-out CV (LOSOCV).

While the studies mentioned above have shown promising results, a potential limitation arises when relying solely on one database. This becomes especially critical when employing nonlinear measures, which necessitate parameter tuning before computation. Ensuring the interchangeability of these tuned features for other databases is of paramount importance, which has been overlooked. In this article, we propose a new method for monitoring mental workload, validated on two databases recorded by two different commercial EEG devices. The n-back memory game served as the test setting. Noise and artifacts were removed from the raw EEG data using Wavelet analysis, and features that quantify the entropy and energy of the EEG subbands were extracted. Finally, classification was performed by five classifiers: SVM, kNN, MLP, Adaboost, and TN. The developed methodology relies on two frontal EEG channels, and the feature set is parameter-free, eliminating the need for hyperparameter calibration.

II. Data

In this article, we used two databases with distinctive characteristics to evaluate the performance of the proposed algorithm. We begin by describing the test setting of the n-back memory game. Following that, we provide a detailed description of the EEG data acquisition process.

A. Study Protocol for Inducing Mental Workload

In investigations of mental workload, especially in studies involving working memory performance, the n-back game has gained considerable popularity. The n-back game induces mental workload by requiring participants to constantly monitor and update their working memory [23]. In this task, individuals are presented with a sequence of stimuli, such as letters or numbers, and are required to indicate whether the current stimulus matches the one presented 'n' steps back in the sequence [20]. This real-time information maintenance and updating place a cognitive load on the working memory system, leading to mental workload as individuals engage in continuous attention, memory retrieval, and decision-making processes [24].

In our study, we used n-back tasks which require containing and processing numbers from 0 to 9 temporarily. The numbers needed to be maintained by the subject to recognize whether the current number matches a number presented one (1-back) or two (2-back) steps before. In the 0-back game, the subject had to compare the currently presented number to a certain number given in the task description. The subjects had to respond to the target numbers by clicking the left mouse button. The game levels cause an increasing amount of mental load with 0-back being the easiest (here considered as No Load or NL), 1-back game rated as Mid-Load (or ML), and 2-back the most difficult (High Load, or HL). In our experience, if levels higher than 2 are used, the number of errors increases significantly causing distraction, and were therefore not used in the experiment.

In our trial, each experimental session consisted of nine game rounds, each featuring 50 numbers. A new number was presented every 2 s; thus, each game lasted 100 s. Between the games, there was a relaxation period of 30 s during which the subjects were instructed to relax and keep their eyes closed.

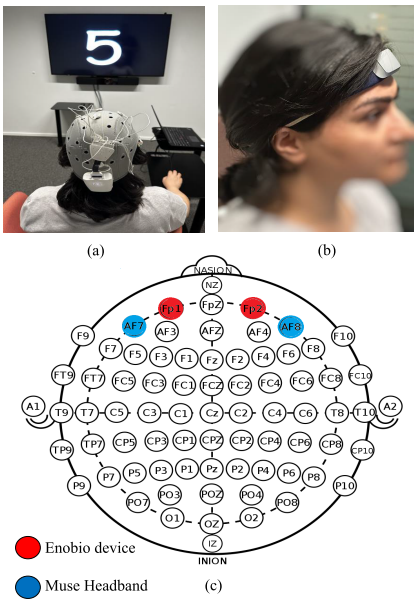

Fig. 1. Illustration of the environment and set up to collect EEG signals while playing an n-back game. (a) Enobio device. (b) Muse headset. (c) Channel configuration.

The recording session included a random sequence of three repetitions of each of the three game levels. Every game event was recorded in a log file containing information on the game level, the displayed number, the subject's response (i.e., mouse click), and the subject's ID.

B. EEG Data Acquisition

The EEG of the first recording set (Database E) was acquired using the ENOBIO[1] EEG device [25] and the Neuroelectrics[1] Instrument Controller (NIC2) software (see Fig. 1, left panel). The electrode-cap-mounted EEG amplifier was connected wirelessly to the PC running the NIC2 software. Although the EEG cap supported up to 20 channels, for the sake of similarity to consumer-oriented devices, we only utilized the Fp1 and Fp2 channels according to the standard 10–20 system of EEG electrode placement [26], which we refer to as CH1 and CH2. The EEG signal was sampled at 500 Hz. For the other recording set, MUSE[1] EEG-headband by Interaxon Inc. was used [27] (Database M; see Fig. 1, right panel). The headset has textile electrodes at positions AF7, AF8, Tp9, and Tp10 of which the AF7 (ch1) and AF8 (ch2) derivations were analyzed. The reference electrode is at Fpz. Third-party software by Petal Technology LLC, running on a PC, was used to receive the EEG data from Muse over a Bluetooth wireless connection while the sampling frequency of Database M was 256 Hz. Measurements were performed using standard electrodes in contact with the subject's skin, and the data were transferred to the PC via commercial Bluetooth protocol. The comprehensive details regarding the primary settings and specifications of the experimental setup, aimed at comparing two distinct EEG devices, are outlined in Table I for reference and comparison.

[1]Registered trademark.

TABLE I
DETAILS OF MAIN SETTINGS AND SPECIFICATIONS OF THE EXPERIMENTAL SETUP

Experimental Setup Details	Database M	Database E
Number of recording	30	30
Sampling Rate	256 Hz	500 Hz
Electrode Placement	AF7,AF8	Fp1, Fp2
Task Description	N-back game	
Mental Workload Classes	Low, Medium, High	

C. Participants

In this study, data were collected from a total of 60 recordings involving 46 subjects, each wearing one of two different EEG headsets. The participants ranged in age from 18 to 65 years, with a diverse demographic composition. Out of the 46 participants, 20 were male, and 26 were female. The participants had no significant experience with EEG devices prior to the study. They were able to test the n-back game before the recording session. The subjects were drawn from a mix of academic backgrounds and roles. The participant pool was intentionally diverse, encompassing individuals from various backgrounds, ensuring a broad representation of different cultures and perspectives. This research was conducted in strict adherence to the Helsinki Declaration, a set of ethical principles guiding human medical research. All participants provided informed consent before participating in the study, and their confidentiality and privacy were protected throughout the research process. The study was approved by the Human Sciences Ethics Committee of Universities in Satakunta, Finland no. 14.12.2022.

III. METHODS

The block diagram of the proposed method for monitoring mental workload is shown in Fig. 2. The algorithm consists of four stages: acquisition of the EEG data, data preprocessing, feature extraction, and classification. In the following subsections, each stage is described in detail.

A. Preprocessing and Artifact Removal

Initially, the data associated with each game round was divided into three segments, each lasting 30 s. Subsequently, each 30-s window was further divided into three 10-s segments. First, a zero-phase Butterworth bandpass filter with a passband ranging from 0.5 to 40 Hz was applied to remove very low and high-frequency distortions. Following this, the discrete wavelet transform (DWT) method was employed to eliminate eye blink artifacts from the signals. Consistent with our previous studies [28], we utilized db4 as the mother wavelet due to its morphology's similarity to that of eye blinks (refer to the preprocessing part in Fig. 2).

B. EEG Subband Decomposition

By DWT, the EEG signal was first converted into the approximation component $a_1[n]$ and the detail component $d_1[n]$. The $a_1[n]$ component was then decomposed again into second level approximation ($a_2[n]$) and detail ($d_2[n]$) components. The decomposition was continued until the maximum

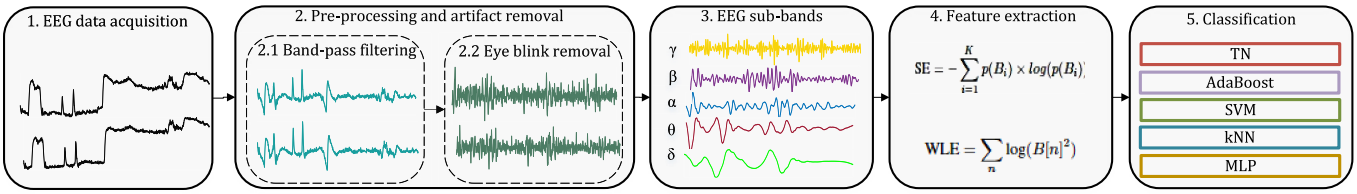

Fig. 2. Proposed framework of mental workload classification in five stages using five different classification methods.

TABLE II
FREQUENCY BANDS USED BY THE DWT COMPONENTS FOR EACH EEG DATABASE

Database M		Database E		Conventional	
DWT level	DWT freq. [Hz]	DWT level	DWT freq. [Hz]	EEG band	EEG freq. [Hz]
a_6	0 . . . 4	a_7	0 . . . 3.9	δ	0 . . . 4
d_6	4 . . . 8	d_7	3.9 . . . 7.8	θ	4 . . . 8
d_5	8 . . . 16	d_6	7.8 . . . 15.6	α	8 . . . 13
d_4	16 . . . 32	d_5	15.6 . . . 31.3	β	13 . . . 30
d_3	32 . . . 64	d_4	31.3 . . . 62.5	γ	> 30

DWT level L was reached. The original signal $x[n]$ can be represented by its components as

$$x[n] = \sum_{l=1}^{L} d_l[n] + a_L[n]. \tag{1}$$

The frequency band of each approximation and detail component can be obtained by

$$a_l = \left[0, \frac{Fs}{2^{l+1}}\right], \quad d_l = \left[\frac{Fs}{2^{l+1}}, \frac{Fs}{2^l}\right] \tag{2}$$

where Fs is the sampling rate. To remove eye blinks, the last approximation components (a_6 and a_7 for Databases M and E, respectively) were denoised using adaptive thresholds, and the corresponding noise component was subtracted from the original signal (see the *Eye blink removal* box in Fig. 2). Because of varying sampling rates, Database E underwent seven levels of decomposition, while Database M underwent six levels. The frequency bands of the decomposed components used in the analysis are outlined in Table II. In the following, we use the notation of standard EEG frequency ranges (i.e., $\delta, \theta, \alpha, \beta$, and γ), even though the cutoff frequencies of the DWT components differ slightly from those of the traditional EEG frequency bands.

C. Feature Extraction

Features characterizing signal entropy and complexity have been extensively utilized in the assessment of mental and cognitive states of the brain [29], [30]. Mental workload often induces alterations in EEG subband characteristics as it changes the level of focus. Given the inherent complexity and uncertainty of EEG signals during task performance, complexity features were expected to possess discriminative power for distinguishing between various levels of mental workload [31].

In this study, wavelet-log-energy (WLE) and Shannon entropy (SE) [32], [33], calculated from the DWT components indicated in Table II were employed. Prior to implementing the DWT on the cleaned EEG signal, normalization was performed to scale the signal within the range from -1 to 1. SE measures the unpredictability of the signal and is calculated according to

$$\text{SE} = -\sum_{i=1}^{K} p(S_i) \times \log(p(S_i)) \tag{3}$$

where $p(S_i)$ is the probability that the amplitude of signal S (in our case a wavelet component of the EEG) falls within range i. K is the number of bins in the probability histogram obtained from signal S. Therefore, SE estimates the flatness of the probability distribution of the underlying signal segment. The WLE, on the other hand, is calculated from the time domain samples of signal S according to

$$\text{WLE} = \sum_{n} \log(S[n]^2) \tag{4}$$

where n is the sample number of the signal segment. Using two EEG channels and five EEG subbands, a total of 20 features were extracted from each signal. Preprocessing and feature extraction were conducted on segments lasting 10 s. Subsequently, the features from three consecutive segments were averaged to represent 30-s signal segments. This resulted in three feature vectors for each game round (see Section II-A).

D. Classification

In this section, we offer a concise overview of the five classifiers that we offer a concise overview of the five classifiers that are the focus of our study: SVM, kNN, MLP, AdaBoost, and TN. By understanding the fundamental characteristics of these classifiers, we can gain insights into their respective strengths and weaknesses when applied to our research.

1) AdaBoost: AdaBoost is a well-known ensemble learning-based categorization model [34]. To direct subsequent hypotheses on more challenging classification scenarios, a set of weak classifiers or hypotheses is constructed. The outcomes of these calculated hypotheses are then combined through a weighted majority voting scheme. The purpose of training a weak classifier is to create hypotheses that can contribute to the overall classification process. While these individual hypotheses might have limited predictive power, they collectively play a role in the decision-making process of the classifier, and as a result, a portion of the original training data is utilized. The subset utilizes randomly selected samples from the training dataset, and the distribution is updated through iterative processes. The distribution update ensures that examples of extremely challenging training samples are introduced

to the training set. Adaboost's overall output is decided by a weighted majority vote over all the weak classifier outputs obtained by minimizing the error, which can discriminate the intricate patterns of cognitive workload in EEG with greater accuracy than other commonly used classifiers such as SVM or the kNN [35]. However, the efficiency of Adaboost is minimally dependent on the number of estimators and the number of leaves within each decision tree [36].

2) Transformer Network: The TN, at its core, can be a classifier equipped with a self-attention mechanism [37]. Compared to recurrent and convolutional models, it has demonstrated superior performance in natural language processing applications. Recently, TNs have been successfully applied to other domains such as image and time series analysis. The main advantage of the TN lies in its capacity to model interactions among elements within the input sequence, regardless of their spatial or temporal separation. Each input receives an attention score with respect to all other inputs establishing their mutual contributions. The main steps of a transformer-based classifier include embedding, positional encoding, multihead attention, and classification by a fully connected layer. The model was trained using the Adam optimizer with an epsilon of 10^{-8}.

The basic transformer architecture used in this study was presented in Fig. 3. A sequence of eight feature vectors, each representing a 30-s segment of the data, is input to the network. The data were preprocessed, and features were extracted by the methods outlined in Section III-C. TNs were initially designed to operate on tokens, therefore the feature vectors were first converted into strings and tokenized. Subsequently, we implemented the standard embedding and positional encoding operations of the transformer algorithm before inputting the data into the multihead attention block.

3) K-Nearest Neighbors: kNN is a simple yet effective algorithm for classification. It works by finding the k training samples nearest to a new input and classifies the input based on the majority class among its neighbors. The choice of k determines the smoothness of the decision boundary; smaller k values lead to more complex boundaries, potentially capturing noise, while larger k values create smoother, generalized boundaries.

4) Support Vector Machine: SVM is a powerful classification technique that finds an optimal hyperplane in a high-dimensional space to best separate different classes. It aims to maximize the margin between classes, enhancing the model's generalizability. SVMs can handle both linear and nonlinear data by employing kernel functions, transforming the input space into a higher dimensional one, where a linear separation is possible.

5) Multilayer Perceptron: MLP is a type of ANN composed of multiple layers of interconnected nodes, or "neurons." It utilizes an input layer, one or more hidden layers, and an output layer. Each connection between nodes is assigned a weight, which the network learns during training. MLPs use activation functions to introduce nonlinearities, enabling them to learn complex patterns in the data. Through iterative training (backpropagation), MLP adjusts the weights to minimize prediction errors, allowing it to model intricate relationships within the dataset.

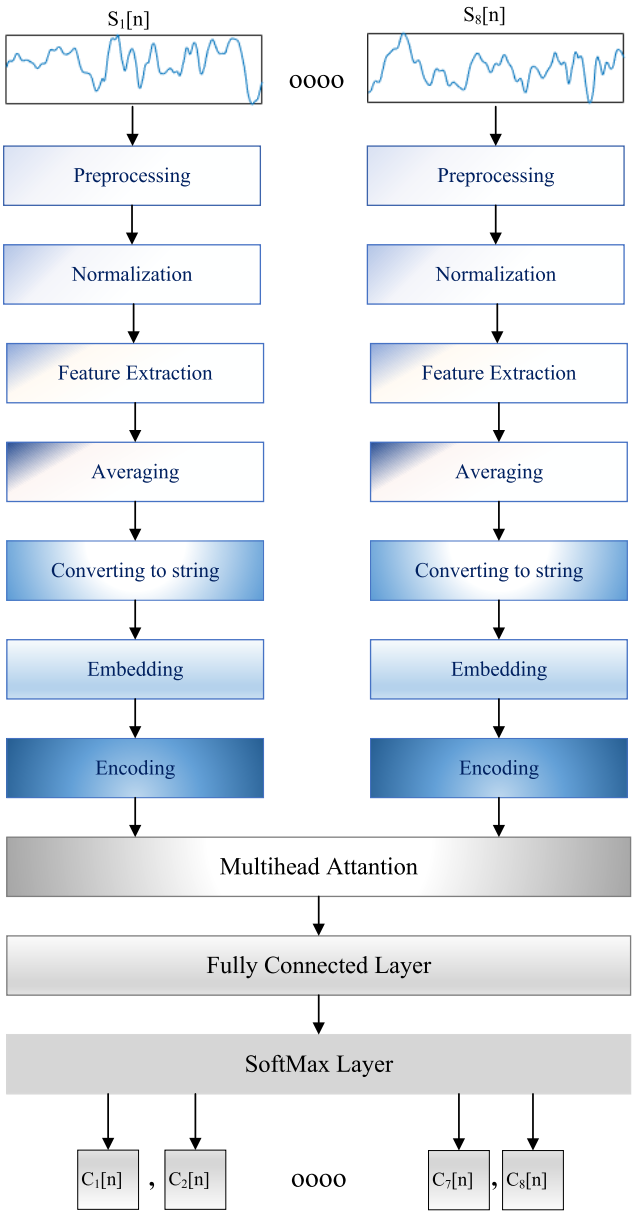

Fig. 3. Structure of the proposed transformer-based algorithm. Here, $Si[n]$ denotes the ith input (i.e., the segment of the raw EEG signal). Preprocessing and feature extraction at each of the eight inputs are performed on three 10-s segments, the feature vectors of which are then averaged.

E. Evaluation

Each feature extracted from the clean data was normalized between 0 and 1. Then, we randomly allocated 80% of the normalized feature vectors for training and validation, while the remaining 20% was reserved for testing. The training-testing procedure was repeated 100 times to ensure the reliability of the classification results.

The effectiveness of each classifier was investigated using the following metrics:

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_N + F_P} \times 100 \tag{5}$$

TABLE III

TESTING RESULTS FOR ALL CLASSIFIERS FOR **DATABASE M**. THE TOTAL ACCURACY IS GIVEN AS THE MEAN ± STANDARD DEVIATION OVER THE 100 RANDOM SAMPLING SETS. IN THE CLASS-WISE RESULTS, ONLY MEAN IS GIVEN

		NL				ML			HL		
	Channel	Accuracy	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AdaBoost	CH 1&2	**76.96±0.03**	80.47	70.83	75.34	78.28	74.10	76.13	73.24	85.94	79.09
	CH 1	72.02±0.03	80.56	63.46	70.98	72.58	74.10	73.32	65.90	78.53	71.66
	CH 2	72.46±0.03	80.57	62.81	70.60	72.50	75.87	74.15	67.04	78.70	72.40
Transformer	CH 1&2	**88.10±0.02**	87.37	81.74	84.46	84.28	86.98	85.61	92.59	95.59	94.06
	CH 1	81.90±0.06	79.72	78.00	78.85	80.07	80.74	80.40	85.81	86.93	86.38
	CH 2	80.88±0.07	81.37	78.18	79.72	78.60	78.77	78.69	82.66	85.66	84.14
SVM	CH 1&2	68.50±0.10	66.17	67.14	64.49	67.98	69.90	66.57	62.29	66.19	68.09
	CH 1	62.82±0.09	59.31	58.61	55.85	61.17	60.39	60.22	58.55	66.25	66.12
	CH 2	60.47±0.07	62.66	56.18	68.72	57.80	58.17	68.69	62.00	55.34	64.74
kNN	CH 1&2	55.71±0.04	55.94	51.83	57.54	50.66	48.31	54.65	56.78	56.68	56.93
	CH 1	49.47±0.05	48.81	47.00	50.13	41.81	42.18	43.09	47.11	46.70	49.56
	CH 2	49.55±0.05	48.81	48.67	50.85	44.21	45.22	45.77	49.77	48.10	50.66
MLP	CH 1&2	66.10±0.02	65.44	59.28	62.55	62.50	64.10	63.85	70.79	71.95	72.46
	CH 1	59.80±0.06	58.32	57.40	58.85	58.70	58.64	58.70	64.51	66.85	66.38
	CH 2	60.18±0.07	61.57	57.28	56.72	57.86	57.87	58.78	62.50	65.14	64.44

TABLE IV

TESTING RESULTS FOR ALL CLASSIFIERS FOR **DATABASE E**. THE TOTAL ACCURACY IS GIVEN AS THE MEAN ± STANDARD DEVIATION OVER THE 100 RANDOM SAMPLING SETS. IN THE CLASS-WISE RESULTS, ONLY MEAN IS GIVEN

		NL				ML			HL		
	Channel	Accuracy	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AdaBoost	CH 1&2	**74.94 ±0.04**	82.39	72.78	77.29	69.65	73.52	71.53	73.81	78.36	76.02
	CH 1	70.17±0.05	77.80	67.55	72.31	67.58	68.26	67.92	66.60	74.70	70.42
	CH 2	71.69±0.03	80.20	62.87	70.50	72.40	73.88	73.13	65.50	78.31	71.33
Transformer	CH 1&2	**85.19±0.04**	81.22	84.33	82.74	81.75	78.35	80.02	92.50	92.88	92.73
	CH 1	79.66±0.05	76.63	78.77	77.69	73.02	70.68	71.83	89.15	89.51	89.33
	CH 2	78.55±0.05	74.08	80.09	76.97	75.40	69.02	72.06	86.23	86.54	86.38
SVM	CH 1&2	67.71±0.10	64.47	65.34	65.55	65.70	67.18	64.57	60.33	65.47	67.80
	CH 1	63.52±0.09	58.91	57.41	56.18	61.17	59.55	60.74	60.48	69.53	63.72
	CH 2	62.71±0.07	63.53	58.27	70.06	59.11	59.717	68.80	63.27	58.25	65.28
kNN	CH 1&2	51.19±0.04	54.48	52.73	59.45	52.12	50.31	58.65	50.28	51.18	50.45
	CH 1	47.47±0.05	47.22	47.56	45.35	45.50	46.14	43.74	49.95	50.70	49.76
	CH 2	50.87±0.05	45.49	46.32	44.00	50.90	51.07	49.57	52.65	50.63	53.74
MLP	CH 1&2	68.10±0.02	67.44	62.73	65.05	65.53	66.12	65.65	72.79	73.95	74.46
	CH 1	63.80±0.06	62.51	61.40	62.85	62.70	64.64	62.70	63.51	69.56	69.21
	CH 2	61.18±0.07	62.57	58.38	57.81	58.28	58.57	59.43	63.56	66.96	65.37

$$\text{Precision} = \frac{T_P}{T_P + F_P} \times 100 \tag{6}$$

$$\text{Recall} = \frac{T_P}{T_P + F_N} \times 100 \tag{7}$$

$$F1\text{-score} = \frac{T_P}{T_P + (F_P + F_N) \times \frac{1}{2}} \times 100 \tag{8}$$

where T_P, T_N, F_P, and F_N denote the number of true positives, true negatives, false positives, and false negatives, respectively.

IV. RESULTS

For Databases M and E, Tables III and IV present the classification results using all classifiers. To provide further clarity, we conducted classification experiments not only with the complete 20-feature vector (comprising two channels, five subbands, and two features) but also with distinct different subsets of features, where each subset included only a single channel. The best classification results were achieved when employing the entire feature set, encompassing all available features and information. The transformer-based classifier exhibited superior performance compared to the other classifiers across both datasets, showing its effectiveness in delivering higher classification accuracy and reliability. In the second place, AdaBoost achieved overall accuracy of 77% and 75% for databases M and E, respectively.

Given the results in Tables III and IV, the confusion matrices (Fig. 4) and the ROC curves (Fig. 5) are shown for the two best-performing classifiers, the transformer and the AdaBoost. The confusion matrices show that the Adaboost classifier tends to make more errors in favor of higher load classes whereas the transformer tends to make more errors in favor of lower load classes. Overall, the HL class has the highest true-positive rates

in all cases. Classified by the transformer, the precision, recall and $F1$-score are all above 90 for the HL class. The AUC values for all three classes and both datasets are significantly higher for the transformer compared to the Adaboost. For the transformer, the AUC values are similar for both datasets whereas for Adaboost better results are obtained for the Muse dataset.

A. Cross Database Performance Analysis

Regarding the cross-database performance, our method demonstrates robust and consistent results. We achieved high accuracies of 88% and 85% for Database M and Database E, respectively, highlighting the effectiveness and generalizability of our approach. Indeed, the conducted t-test revealed no significant differences between these results ($p < 0.05$), affirming the reliability and consistency of our method's performance across different datasets.

B. Comparison With the State-of-the-Art Methods

In the comparative analysis presented in Table V, various studies investigating mental workload assessment are compared based on several key factors, including the number of databases, channels, subjects, devices, tasks, classifiers employed, and the corresponding training-testing strategies, all of which contribute to the reported accuracy. Nonetheless, it should be noted that we have only considered studies that employed consumer-oriented EEG headsets for mental workload assessment. Among the compared studies, the works by So et al. [22] involved 20 subjects using a single channel EEG setup, achieving an accuracy of 65%–75% with a (LOSOCV) strategy. The study by Liu et al. [21], utilizing five channels, achieved an impressive accuracy of 87% with a ten-CV approach in air traffic tasks. The research by Wang et al. [20], incorporating 14 channels, reached an accuracy of 84% in an n-back game using a Proximal SVM classifier with ten-CV. The study by Almoghbel et al. [18], involving four channels, attained an accuracy of 89% in simulated driving tasks using CNN and CV. In comparison, our method was evaluated using two databases (Database M and Database E), each with two channels and 30 subjects. Employing a TN classifier, we achieved an accuracy of 88% for Database M, and 85% for Database E, demonstrating competitive performance in mental workload assessment.

In comparison to existing studies, our research offers several distinctive advantages. Notably, we utilized two diverse databases (Database M and Database E), assuring the robustness and generalizability of our findings. Furthermore, our method consistently demonstrated competitive performance, achieving accuracy rates of 88% and 85% for Database M and Database E, respectively, using random sampling. Remarkably, these high accuracies were achieved with the utilization of only two EEG channels, highlighting the efficiency of our channel allocation strategy. Despite the simplicity in channel usage, our accuracy levels were comparable to, and in some cases, exceeded those obtained by studies employing more channels. This speaks to the efficacy of our chosen features and classification techniques.

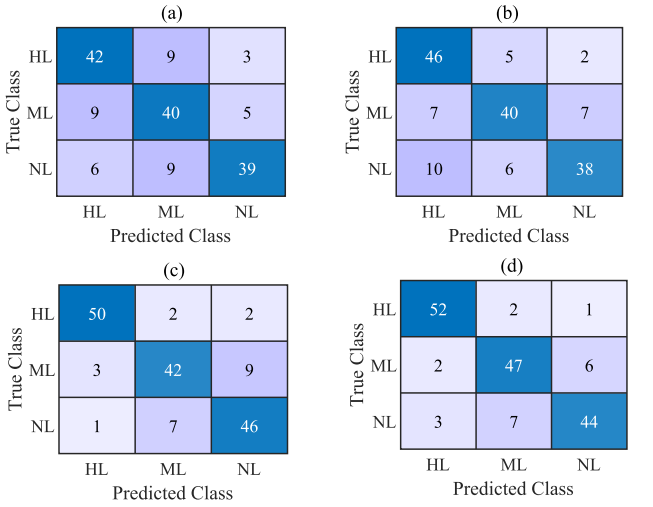

Fig. 4. Confusion matrices for classification by AdaBoost for (a) Database E and (b) Database M. Confusion matrices for classification by the transformer for (c) Database E and (d) Database M. The values represent mean results over 100 validation sets, rounded to integer values.

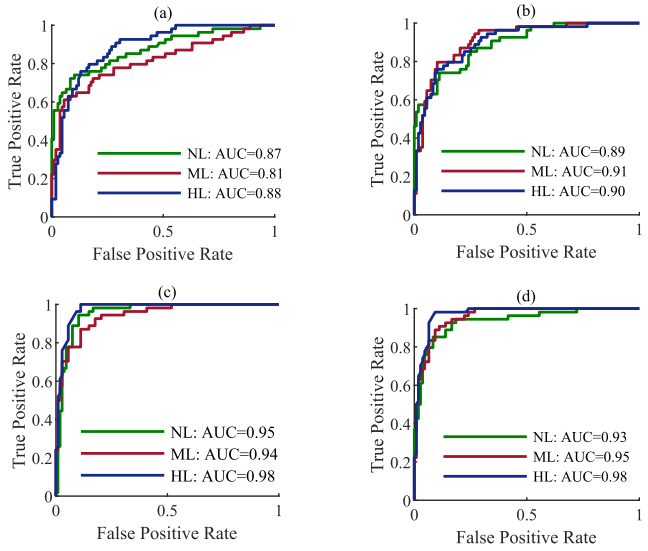

Fig. 5. ROC curves and AUC values for classification by AdaBoost for (a) Database E and (b) Database M. ROC curves and AUC values for classification by the transformer for (c) Database E and (d) Database M. The AUC values and ROC curves represent mean results over 100 validation sets.

V. Discussion

This study introduces an innovative method designed for consumer-oriented equipment, enabling automated classification of discrete levels of mental workload. We employed low-channel EEG signals specifically sourced from the prefrontal cortex, an area linked to cognitive functions and mental workload. In addition, this region is hairless, which can potentially lead to better signal quality. Furthermore, we utilized two parameter-free features: SE and WLE, with the latter being notably absent in prior studies within this domain. At last, we employed a deep learning model based on transformer architecture, which has not been investigated in mental workload assessment. The mentioned innovations open up new avenues for research and practical applications in mental workload assessment.

TABLE V
COMPARISON BETWEEN THE PROPOSED AND STATE-OF-THE-ART METHODS

Study	Task	Classifier	Training-testing strategy	No. databases	No. Channels	No. subjects	Accuracy
So et al. [22]	Four cognitive and motor tasks	SVM	LOSOCV	1	1	20	65-75%
Liu et al. [21]	Airfield Traffic Pattern tasks	kNN	10-CV	1	5	21	87%
Wang et al. [20]	n-back game task	Proximal SVM	10-CV	1	14	9	84%
Arsalan et al. [19]	Public speaking task	MLP	LOOCV	1	4	28	64%
Almoghbel et al. [18]	Simulated driving task	CNN	CV	1	4	1	89%
Proposed	n-back game task	TN	Random sampling	Database M	2	30	88%
				Database E	2	30	85%

Several studies have tackled cognitive load classification using various biomedical signals, including EEG, ECG, heart rate, galvanic skin response, and respiratory rate [7], [38]. The brain is undoubtedly the primary source of mental states, whereas the other biomedical signals signify the autonomous nervous system's response to cognitive load. Consequently, detecting mental states directly from the primary source is paramount for real-life applications. A variety of algorithms based on EEG data has been proposed for the detection and monitoring of mental workload such as [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [15], [16], and [38]. However, these algorithms are hardly practical in real-life scenarios due to their high complexity. On one hand, utilizing multichannel EEG signals presents challenges due to the need for individuals to wear EEG caps with numerous electrodes and attachments. Additionally, multichannel EEG setups often include channels from areas with hair, making recordings uncomfortable for users [2], [13], [15], [30]. On the other hand, studies that relied on low-channel EEG data typically examine a single EEG dataset, resulting in a limited focus on specific devices [7], [16], [17], [18], [19], [20], [21], [22], [23].

In our study, we implemented a configuration utilizing just two prefrontal EEG channels to address these issues. The prefrontal region of the brain is involved in crucial cognitive processes. Its role in executive functions, attention regulation, emotional control, and information integration makes it a reliable source of activity for the assessment of mental workload variations [39]. Furthermore, we employed two distinct datasets: one acquired using a user-friendly mobile device that is clinically validated, enabling complete montage recording, and another dataset collected using a basic consumer-oriented EEG headband. The classification accuracy for both datasets was comparable, demonstrating the applicability of the proposed algorithm across various devices. Surprisingly, slightly better results were obtained for the consumer-oriented Muse dataset. The explanation could lie in that the reference electrode in the Muse headband is positioned at Fpz, which effectively cancels out artifacts originating from the frontal part of the head, such as eye blinks, forehead muscle activity, and eye movements. An alternative reason could be that we likely exercised greater caution in visually assessing signal quality with the Muse device before commencing recordings, as the electrodes took longer to stabilize.

When we assess the performance of five classifiers, it becomes evident that the transformer consistently outperformed others across a wide range of classification metrics. In a prior investigation that employed a transformer-based algorithm [16] with a dataset comprising 14 EEG channels, there was no significant improvement in accuracy observed compared to other classification methods when tasked with multistage mental workloads. The comparison between our obtained results and those reported in [16] suggests that the benefits conferred by the multihead attention mechanism might be more pronounced when dealing with lower level devices that have limited data available for classification. Nonetheless, this observation is based on the comparison of these two studies and leaves room for further investigation. In the context of algorithms designed for consumer-oriented EEG headsets, low complexity is of paramount importance. Although we conducted research involving a network based on transformer architecture, it is worth noting that our proposed method demonstrates reduced complexity when compared to the deep learning methodologies outlined in [9], [16], and [18] primarily due to the utilization of a smaller number of EEG channels. This inherent efficiency renders the proposed algorithm more streamlined, resulting in both increased speed and efficiency.

A. Direction for Future Work

Although the reported results are promising, this research does possess certain limitations that require attention in future investigations. First, employing a more extensive database with a greater cohort diversity could potentially further improve the reliability of the results. Second, the study did not explore the method's effectiveness on an individual subject basis. Indeed, individual differences in working memory capacity can impact n-back task performance, potentially leading to variations in workload assessments. Third, it would be worthwhile to explore alternative deep learning methods like the combination of TN and CNN for potential applicability. At last, the algorithm's performance might diverge when applying techniques like LOSOCV during the training-testing process. Nonetheless, it should be noted that such a strategy requires a larger number of labels for all three classes from each individual subject in order to obtain fair results.

VI. CONCLUSION

In this article, we propose an algorithm for assessing the mental workload by utilizing only prefrontal EEG data recorded from commercially available EEG headsets, addressing a prevailing constraint in the state-of-the-art approaches that involve an extensive array of EEG channels spanning various brain regions. The main advantage of the proposed method is its proven reliability, demonstrated by achieving comparable results in terms of classification metrics across

two databases with distinct characteristics. Indeed, the TN achieved an accuracy of 88.1 and 85.2 for Databases M and E, respectively.

REFERENCES

[1] L. Longo, C. D. Wickens, G. Hancock, and P. A. Hancock, "Human mental workload: A survey and a novel inclusive definition," *Frontiers Psychol.*, vol. 13, pp. 1–13, Jun. 2022.

[2] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84045–84066, 2021.

[3] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52. Amsterdam, The Netherlands: Elsevier, 1988, pp. 139–183.

[4] G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: A scaling procedure for measuring mental workload," in *Advances in Psychology*. Amsterdam, The Netherlands: Elsevier, 1988, pp. 185–218.

[5] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 5, pp. 434–451, Oct. 2018.

[6] A. Giorgi et al., "Wearable technologies for mental workload, stress, and emotional state assessment during working-like tasks: A comparison with laboratory technologies," *Sensors*, vol. 21, no. 7, p. 2332, Mar. 2021.

[7] S. Betti et al., "Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 8, pp. 1748–1758, Aug. 2018.

[8] R. Katmah, F. Al-Shargie, U. Tariq, F. Babiloni, F. Al-Mughairbi, and H. Al-Nashash, "A review on mental stress assessment methods using EEG signals," *Sensors*, vol. 21, no. 15, p. 5043, Jul. 2021, doi: 10.3390/s21155043.

[9] Y. Zhou et al., "Cross-subject cognitive workload recognition based on EEG and deep domain adaptation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[10] K. Guan, Z. Zhang, T. Liu, and H. Niu, "Cross-task mental workload recognition based on EEG tensor representation and transfer learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2632–2639, 2023.

[11] I. Kakkos et al., "EEG fingerprints of task-independent mental workload discrimination," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3824–3833, Oct. 2021.

[12] P. Zarjam, J. Epps, F. Chen, and N. H. Lovell, "Estimating cognitive workload using wavelet entropy-based features during an arithmetic task," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2186–2195, Dec. 2013.

[13] P. Zarjam, J. Epps, and N. H. Lovell, "Beyond subjective self-rating: EEG signal classification of cognitive workload," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 4, pp. 301–310, Dec. 2015.

[14] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–31, 2023.

[15] Z. Jiao, X. Gao, Y. Wang, J. Li, and H. Xu, "Deep convolutional neural networks for mental load classification based on EEG data," *Pattern Recognit.*, vol. 76, pp. 582–595, Apr. 2018.

[16] G. Siddhad, A. Gupta, D. P. Dogra, and P. P. Roy, "Efficacy of transformer networks for classification of EEG data," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105488.

[17] B. Hu et al., "Signal quality assessment model for wearable EEG sensor on prediction of mental stress," *IEEE Trans. Nanobiosci.*, vol. 14, no. 5, pp. 553–561, Jul. 2015.

[18] M. A. Almogbel, A. H. Dang, and W. Kameyama, "Cognitive workload detection from raw EEG-signals of vehicle driver using deep learning," in *Proc. 21st Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2019, pp. 1–6.

[19] A. Arsalan, M. Majid, A. R. Butt, and S. M. Anwar, "Classification of perceived mental stress using a commercially available EEG headband," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2257–2264, Nov. 2019.

[20] S. Wang, J. Gwizdka, and W. A. Chaovalitwongse, "Using wireless EEG signals to assess memory workload in the n-back task," *IEEE Trans. Hum.-Mach. Syst.*, vol. 46, no. 3, pp. 424–435, Jun. 2016.

[21] C. Liu et al., "Detection of Pilot's mental workload using a wireless EEG headset in airfield traffic pattern tasks," *Entropy*, vol. 25, no. 7, p. 1035, Jul. 2023.

[22] W. K. Y. So, S. W. H. Wong, J. N. Mak, and R. H. M. Chan, "An evaluation of mental workload with frontal EEG," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0174949.

[23] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning Spatial–Spectral–Temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 1, pp. 31–42, Jan. 2019.

[24] C. Scharinger, L. Prislan, K. Bernecker, and M. Ninaus, "Gamification of an n-back working memory task – is it worth the effort? An EEG and eye-tracking study," *Biol. Psychol.*, vol. 179, Apr. 2023, Art. no. 108545.

[25] (2024). *Enobio*. [Online]. Available: https://www.neuroelectrics.com/solutions/enobio

[26] J. Hh, "Report of the committee on methods of clinical examination in electroencephalography: 1957," *Electroencephalogr. Clin. Neurophysiology*, vol. 10, no. 2, pp. 370–375, 1957.

[27] *Muse*. [Online]. Available: https://choosemuse.com/

[28] M. Shahbakhti et al., "Fusion of EEG and eye blink analysis for detection of driver fatigue," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2037–2046, 2023.

[29] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang, "Cognitive workload recognition using EEG signals and machine learning: A review," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 799–818, Sep. 2022.

[30] R. Ferenets, T. Lipping, A. Anier, V. Jantti, S. Melto, and S. Hovilehto, "Comparison of entropy and complexity measures for the assessment of depth of sedation," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1067–1077, Jun. 2006.

[31] P. Liu, W. Beh, C. Shih, Y. Chen, and A. A. Wu, "Entropy and complexity assisted EEG-based mental workload assessment system," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2019, pp. 1–4.

[32] J. Min, C. Xiong, Y. Zhang, and M. Cai, "Driver fatigue detection based on prefrontal EEG using multi-entropy measures and hybrid model," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102857.

[33] A. Nalwaya, K. Das, and R. B. Pachori, "Automated emotion identification using Fourier–Bessel domain-based entropies," *Entropy*, vol. 24, no. 10, p. 1322, Sep. 2022.

[34] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[35] M. Beiramvand, T. Lipping, N. Karttunen, and R. Koivula, "Mental workload assessment using low-channel prefrontal EEG signals," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2023, pp. 1–5.

[36] N. Pusarla, A. Singh, and S. Tripathi, "Normal inverse Gaussian features for EEG-based automatic emotion recognition," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[37] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[38] J. Minguillon, E. Perez, M. Lopez-Gordo, F. Pelayo, and M. Sanchez-Carrion, "Portable system for real-time detection of stress level," *Sensors*, vol. 18, no. 8, p. 2504, Aug. 2018.

[39] Y. Zhu, Q. Wang, and L. Zhang, "Study of EEG characteristics while solving scientific problems with different mental effort," *Sci. Rep.*, vol. 11, no. 1, p. 23783, Dec. 2021.

Matin Beiramvand (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in biomedical engineering from Islamic Azad University, Tehran, Iran, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Faculty of Information and Communication Technology, Tampere University, Pori, Finland.

Since 2023, she has been a Researcher with the Faculty of Information and Communication Technology, Tampere University, where she conducts research on mental workload detection using consumer-oriented devices. Her research interests include applying biomedical signal processing and machine learning methods in medical wearable equipment for real-life applications.

Mohammad Shahbakhti (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the Biomedical Engineering Institute, Kaunas University of Technology, Kaunas, Lithuania, where he conducts research on monitoring consciousness levels using wearable EEG systems.

Since 2022, he has been a full-time Researcher with Artinis Medical Systems B.V., Elst, The Netherlands, focusing on developing algorithms for estimating complementary physiological parameters from fNIRS data.

Nina Karttunen received the Registered Nurse degree from Diaconia University of Applied Sciences, Helsinki, Finland, in 2014, and the master's degree in health promotion from Satakunta Hospital District, Pori, Satakunta, Finland, in 2022.

From 2014 to 2020, she was with Satakunta Hospital District, as a Registered Nurse. Since 2022, she has been a Researcher and the Project Manager with Satakunta University of Applied Sciences, Pori, Finland, on various health-related projects, utilizing her nursing background.

Reijo Koivula has over ten years of experience in business development with various companies and has been an entrepreneur for more than 25 years. His research interests lie in education, the creation of learning methods, the development and implementation of learning platform exports and training exports, the manufacturing industry, and various areas of information technology in Europe and the Middle East.

Jari Turunen received the Ph.D. degree in signal processing from Tampere University of Technology, Tampere, Finland, in 2003.

He is currently a University Lecturer with Tampere University, Pori, Finland. His main research interests include time series, machine learning, and statistical data analysis.

Tarmo Lipping (Senior Member, IEEE) received the Dr.Tech. degree in signal processing and the M.BA. degree from Tampere University of Technology, Tampere, Finland, in 2001 and 2013, respectively.

From 2001 to 2002, he was a Post-Doctoral Research Associate with Dartmouth College, Hanover, NH, USA. From 2002 to 2003, he was the Director of the Biomedical Engineering Center and Held Professorship in Biomedical Engineering, Tallinn University of Technology, Tallinn, Estonia. Since 2004, he is a Professor of Signal Processing, Tampere University, Pori, Finland. From 2019 to 2023, he was the Director of Pori University Consortium, Pori. During his career, he has been Principal Investigator of numerous academic and industry-related research projects; he is the author or coauthor of over 100 research publications and has supervised over 60 Master's and seven Doctoral theses. His research interests include applying machine learning and artificial intelligence tools to the monitoring of cognitive and mental states in real-life situations.