2011-05-16

An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload

Bujar Raufi
Technological University Dublin, bujar.raufi@tudublin.ie

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Frontiers in Neuroinformatics

An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload

Bujar Raufi* and Luca Longo

Artificial Intelligence and Cognitive Load Lab, Applied Intelligence Research Centre, School of Computer Science, Technological University Dublin, Dublin, Ireland

Many research works indicate that EEG bands, specifically the alpha and theta bands, have been potentially helpful cognitive load indicators. However, minimal research exists to validate this claim. This study aims to assess and analyze the impact of the alpha-to-theta and the theta-to-alpha band ratios on supporting the creation of models capable of discriminating self-reported perceptions of mental workload. A dataset of raw EEG data was utilized in which 48 subjects performed a resting activity and an induced task demanding exercise in the form of a multitasking SIMKAP test. Band ratios were devised from frontal and parietal electrode clusters. Building and model testing was done with high-level independent features from the frequency and temporal domains extracted from the computed ratios over time. Target features for model training were extracted from the subjective ratings collected after resting and task demand activities. Models were built by employing Logistic Regression, Support Vector Machines and Decision Trees and were evaluated with performance measures including accuracy, recall, precision and f1-score. The results indicate high classification accuracy of those models trained with the high-level features extracted from the alpha-to-theta ratios and theta-to-alpha ratios. Preliminary results also show that models trained with logistic regression and support vector machines can accurately classify self-reported perceptions of mental workload. This research contributes to the body of knowledge by demonstrating the richness of the information in the temporal, spectral and statistical domains extracted from the alpha-to-theta and theta-to-alpha EEG band ratios for the discrimination of self-reported perceptions of mental workload.

Keywords: human mental workload, EEG band ratios, alpha-to-theta ratios, theta-to-alpha ratios, machine learning, classification

1. INTRODUCTION

Human mental workload is a fundamental concept for investigating human performance. It represents an intrinsically complex and multilevel concept, and ambiguities exist in its definition. The most general description of mental workload can be framed as the quantification of a cognitive cost of performing a task in a finite timeframe in order to predict operator, system performance or both (Reid and Nygren, 1988; Rizzo and Longo, 2018; Hancock et al., 2021). Mental workload has

been regarded as an essential factor that substantially influences task performance (Young et al., 2015; Galy, 2018; Longo, 2018a). As a construct, it has been widely applied in the design and evaluation of complex human-machine systems and environments such as in aircraft operation (Hu and Lodewijks, 2020; Yu et al., 2021), train and vehicle operation (Li et al., 2020; Wang et al., 2021), nuclear power plants (Gan et al., 2020; Wu et al., 2020), various human-computer and brain-computer interfaces (Longo, 2012; Asgher et al., 2020; Putze et al., 2020; Bagheri and Power, 2021) and in educational contexts (Moustafa and Longo, 2019; Orru and Longo, 2019; Longo and Orr, 2020; Longo and Rajendran, 2021), to name a few. Mental workload research has accumulated momentum over the last two decades, given the fact that numerous technologies have emerged that engage users in multiple cognitive levels and requirements for different task activities operating in diverse environmental conditions.

Different methods have been proposed to measure human mental workload. These methods can be clustered into three main groups. *Subjective measures* which relies on the analysis of the subjective feedback provided by humans interacting with an underlying task and is usually in the form of a post-task survey. The most well-known subjective measurement techniques are the NASA Task Load Index (NASATLX) (Hart and Staveland, 1988), the Workload profile (WP) (Tsang and Velazquez, 1996), and the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988). *Task performance measures*, often referred to as primary and secondary tasks measures, focus on the objective measurement of a human's performance in an underlying task. Examples of such measures include timely completion of a task, reaction time to secondary tasks, number of errors on the primary task and tapping error. *Physiological measures* are based upon the analysis of physiological responses of the human body. Examples include EEG (electroencephalography), MEG (magnetoencephalography), brain metabolism, endogenous eye blink rates, pupil diameter, heart rate variability (HRV) measures or electrodermal responses such as galvanic skin response (GSR) (Byrne, 2011).

Many research works indicate that EEG data contains information that can help correlate task engagement and mental workload in cognitive processes like vigilance, learning and memory (Berka et al., 2007; Roy et al., 2016), in operating under environmental factors such as temperature (Wang et al., 2019) and in critical systems domains such as transport (Borghini et al., 2014; Diaz-Piedra et al., 2020), nuclear power plants (Choi et al., 2018) and aviation (Wilson et al., 2021). The reason for using EEG is that it offers several benefits compared to imaging techniques or mere behavioral observational approaches. The most important benefit of EEG is its excellent time resolution which offers the possibility to study the precise time-course of cognitive and emotional processing of behavior. Billions of neurons in the human brain are organized in a highly intricate and convoluted fashion exhorting in complex firing patterns. These patterns, accompanied by frequency oscillations, are measurable with EEG reflecting certain cognitive, affective or attentional states. These frequencies, in adults, are usually decomposed in different bands: delta band (1–4 Hz), theta band

(4–8 Hz), the alpha band (8–12 Hz), the beta band (13–25 Hz) and gamma band ($\geq$ 25 Hz) (Mesulam, 1990).

Recent studies seem to indicate changes in frequency band across different brain regions when a subject performs specific tasks (Gevins and Smith, 2003; Schmidt et al., 2013; Borys et al., 2017). The theta band is thought to be linked to mental fatigue and mental workload (Gevins et al., 1995). The increase in theta spectral power is thought to be correlated with the rise in the use of cognitive resources (Tsang and Vidulich, 2006; Xie et al., 2016), task difficulty (Antonenko et al., 2010) and working memory (Borghini et al., 2012). Alpha band tends to show sensitivity in experiments with mental workload (Xie et al., 2016; Puma et al., 2018), cognitive fatigue (Borghini et al., 2012), attention and alertness (Kamzanova et al., 2014).

Even though EEG bands have been proposed as indicators that can discriminate mental workload (Gevins and Smith, 2003; Tsang and Vidulich, 2006; Antonenko et al., 2010; Coelli et al., 2015), it is unclear which of these best contribute to such discrimination. This article aims to identify the impact of the high-level features extracted from alpha and theta band ratios (and their combination) on the discrimination of levels of perception of mental workload self-reported by users. To tackle this aim, an empirical research experiment has been designed to generate time-series of alpha, theta band ratios, and their combinations, and extract high-level features that can be used to build models to classify self-report perceptions of mental workload.

The remainder of this article is organized as follows: Section 2 outlines the related work regarding the specific definition and use of the alpha-to-theta and theta-to-alpha band ratios along with their relationship to mental workload. Section 3 describes the design of an empirical experiment and the methodology employed for answering the above research goal. Section 4 presents the findings followed by a critical discussion while Section 5 concludes this work, proposing future research directions.

2. RELATED WORK

Recent studies analyze EEG bands on various experimental settings designed for specific domains and purposes such as fatigue and drowsiness (Borghini et al., 2014), brain-computer interfaces (Gevins and Smith, 2003; Käthner et al., 2014), learning (Borys et al., 2017; Dan and Reiner, 2017) as well as for specific brain function disorders such as Alzheimer (Schmidt et al., 2013). Most research studies seem to indicate the possibility that EEG signals across various cortical regions can be a helpful tool toward discriminating mental workload while performing experiments with varying degree of task demands (Borghini et al., 2014).

The theta band is thought to be linked to mental fatigue and drowsiness (Gevins et al., 1995; Borghini et al., 2014). Increase of spectral power in the theta band is associated with an increase of demand in cognitive resources (Tsang and Vidulich, 2006; Xie et al., 2016), an increase in task difficulty (Gevins and Smith, 2003; Antonenko et al., 2010; Käthner et al., 2014; Borghini et al., 2015) and an increase in working memory (Borghini et al., 2012,

2014). Particularly, the theta power spectrum seems to increase in cases where a prolonged concentration while executing a task is required (Gevins and Smith, 2003; Borghini et al., 2014; Käthner et al., 2014). Some research even indicates a decrease in vigilance and alertness where a higher power spectrum in theta band is observed (Kamzanova et al., 2014). The brain regions thought to be associated with theta activity are mostly in the frontal cortical area (Gevins and Smith, 2003; Borghini et al., 2014; Dan and Reiner, 2017).

The research on alpha band emerges to indicate sensitivity toward mental workload (Xie et al., 2016; Puma et al., 2018), cognitive fatigue (Borghini et al., 2012, 2014), and an increase in the alpha band activity is associated with a decrease in attention and alertness (Kamzanova et al., 2014). An increase/decrease in the alpha band power spectrum is witnessed during relaxed states with eyes closed and opened, respectively (Antonenko et al., 2010). A continuous suppression in the alpha band seems to be linked with increments of task difficulty (Mazher et al., 2017). The brain regions that are primarily associated with the alpha-band activity are parietal and occipital areas (Borghini et al., 2014; Puma et al., 2018).

The beta band is thought to be linked to visual attention (Wróbel, 2000), short term memory tasks (Palva et al., 2011) and inconclusively it was hypothesized that an increase in the beta band is associated with an increase in working memory (Spitzer and Haegens, 2017). An increase in the beta band spectrum seems to be associated with increased levels of task engagement (Coelli et al., 2015) and concentration (Kakkos et al., 2019). The brain regions that are associated with the beta-band activity are parieto-occipital areas that have been observed during visual working memory task experiments (Mapelli and Özkurt, 2019).

Multiple EEG band combinations and ratios have also been used to improve mental workload assessment. For instance, $beta/(alpha + theta)$ known as engagement index is used to study task human engagement (Mikulka et al., 2002), mental attention (MacLean et al., 2012) and mental effort (Smit et al., 2005). The reduction in the alpha band activity seems to correlate with increased activity in the frontal-parietal areas with an increase in beta power followed by a decrease in theta, which indicates high vigilance states (MacLean et al., 2012). Alpha band activity reduction is also thought to correlate with activities in the parietal brain region where a decrease in beta activity followed by an increase of theta band activity indicate states of drowsiness and low attention (MacLean et al., 2012).

Attempts to assess mental workload and task engagement using the information from the theta and alpha bands in the form of theta-to-alpha band ratios are seen in Gevins and Smith (2003), Käthner et al. (2014), Dan and Reiner (2017), and Xie et al. (2016). This is based on the assumption that an increase in the theta power band in the frontal brain region, and a decrease in the alpha power in parietal region is associated with an increase in mental workload (Käthner et al., 2014). The increase in both alpha and theta power is related to the rise of fatigue (Käthner et al., 2014; Xie et al., 2016). Research seems to indicate that task load manipulations are followed by an increase of theta band activity in frontal brain regions followed by a decrease in alpha power in the parietal areas

(Gevins and Smith, 2003; Käthner et al., 2014; Dan and Reiner, 2017).

The motivation for this article arises from the fact that research studies are indicating that band ratios, specifically the theta and alpha bands, are associated with mental workload states (Gevins and Smith, 2003; Borghini et al., 2014) and to some extent, this seems to justify their potential as workload indicators (Fernandez Rojas et al., 2020). While research exists that focuses on the alpha, theta and beta bands as well as their respective ratios such as $beta/(alpha + theta)$ and to some extent, theta-to-alpha, there is an absence of research related to the use of the alpha-to-theta and theta-to-alpha ratios and their role in discriminating self-reported perceptions of mental workload. Therefore, to address the goal as stated in the introductory Section 1 we formulate a research problem focused on the investigation of the importance of high-level features extracted from the alpha-to-theta and the theta-to-alpha EEG band ratios on the discrimination of levels of perception of mental workload. In other words, the research question that can be formulated is: what is the impact of high-level features extracted from alpha and theta band ratios (and their combination) on discriminating of levels of perception of mental workload self-reported by users?
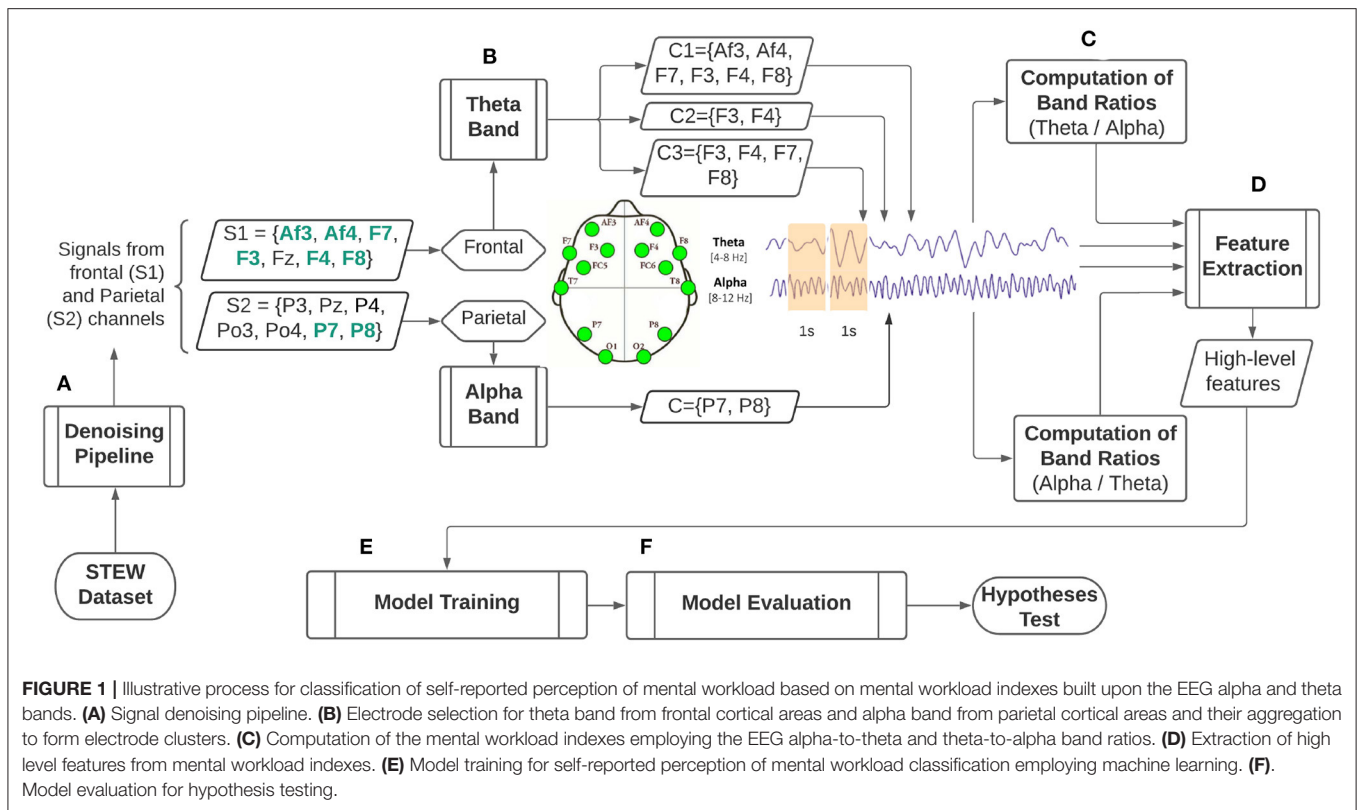
3. DESIGN AND METHODOLOGY

To answer the research problem and research question outlined above, the following research hypotheses are defined:

1. **H1:** If high-level features are extracted from indexes of mental workload built upon alpha-to-theta and theta-to-alpha band ratios, then their discriminatory capacity to self-reported perceptions of mental workload will be higher than those extracted from indexes of mental workload built upon the alpha and theta bands alone.
2. **H2:** If more adjacent EEG electrodes from the respective cortical areas are used to create indexes of mental workload built upon alpha-to-theta and theta-to-alpha band ratios, then they will exhibit higher discriminatory capacity to self-reported perceptions of mental workload than those indexes built with fewer electrodes.

In order to test these research hypotheses, empirical comparative research has been designed based on a process pipeline as illustrated in **Figure 1** with details outlined in the following subsections.

3.1. Experiment Design and Dataset Description

The STEW (Simultaneous Task EEG Workload) (Lim et al., 2018) has been selected for experimental purposes. The dataset consists of raw EEG data collected from 48 subjects across 14 channels in two experimental conditions. In one condition, the EEG data was recorded from subjects in the rest state while not performing any mental activity. In the second condition, a multitasking SIMKAP test was presented to subjects, and EEG data was recorded. In both cases, a sampling frequency of 128 Hz was used with 2.5 min of EEG recordings utilizing the Emotiv EPOC EEG headset.

FIGURE 1 | Illustrative process for classification of self-reported perception of mental workload based on mental workload indexes built upon the EEG alpha and theta bands. **(A)** Signal denoising pipeline. **(B)** Electrode selection for theta band from frontal cortical areas and alpha band from parietal cortical areas and their aggregation to form electrode clusters. **(C)** Computation of the mental workload indexes employing the EEG alpha-to-theta and theta-to-alpha band ratios. **(D)** Extraction of high level features from mental workload indexes. **(E)** Model training for self-reported perception of mental workload classification employing machine learning. **(F)**. Model evaluation for hypothesis testing.

Every recording contains 19,200 data samples (128 samples x 150 s) across the following 14 channels: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4. Additionally, a subjective rating was collected after each task execution whereby users rated their experienced mental workload on the scale 1 to 9. The rating was a likert scale with 1 = "very, very low mental effort"; 2 = "very low mental effort"; 3 = "low mental effort"; 4 = "rather low mental effort"; 5 = "neither low nor high mental effort"; 6 = "rather high mental effort"; 7 = "high mental effort"; 8 = "very high mental effort" and 9 = "very, very high mental effort." The rationale of using a perceived mental workload scores, was a form of subjective validation to verify whether a subject indeed experienced an increase in cognitive load load while performing the SIMKAP condition as compared to the resting condition.

3.2. EEG Denoising Pipeline

Applying a denoising pipeline is an important step to pre-process the raw EEG data and to remove noise from it to facilitate subsequent analysis. In detail, this process follows the Makoto's pre-processing pipeline (Miyakoshi, 2018) including:

- re-referencing channel data to average reference.
- high-pass filtering of each channel at 1hz.
- source separation and artifact removal via Independent Component Analysis (ICA).

The key pre-processing step is the application of ICA which is utilized to separate the 14 EEG signal sources into independent components for each subject. Fourteen components

are generated and used to automatically find and remove artifacts without human intervention using part of the methodology described in Nolan et al. (2010). In detail, the criteria for identifying bad components includes the computation of the z-scores of each component's spectral kurtosis, slope, Hurst exponent and gradient median. Spectral kurtosis is a parameter in the frequency domain indicating the component's impulsiveness variation with frequency. The slope of a component represents its mean slope of the power spectrum over two-time points. The Hurst exponent, also known as a long term memory in time series, tends to measure the tendency of a component to either regress to its mean or to catch up with an upward/downward trend. The gradient median is the median slope of the component's time course. All of the components exhorting values above and below ranges "$z - score \pm 3$" can be considered as artifacts since they are outliers and significantly different from all the others. The value of ± 3 for a $z - score$ was adopted from Nolan et al. (2010) as part of automatic outlier detection taken from the FastR method. Finally, the inverse ICA has been executed to convert the remaining "good" components back in the original neural EEG signal.

3.3. Forming Cluster Combinations

A baseline of initial parietal and frontal electrodes was adopted following the electrode locations from the 10-20 international system to form different alpha and theta clusters for analysis and comparison purposes. These electrode locations were cross-referenced with locations, naming notation and electrode

Cluster notation	Band	Electrodes	Channel aggregation approach
$c1 - \theta$	theta	AF3, AF4, F3, F4, F7, F8	Average
$c2 - \theta$	theta	F3, F4	Average
$c3 - \theta$	theta	F3, F4, F7, F8	Average
$c - \alpha$	alpha	P7, P8	Average

availability from the Emotiv EPOC EEG headset. The initial electrodes that are selected from the frontal and parietal locations are indicated as $S1$ and $S2$ in **Figure 1**. Due to the limited availability of electrodes from the Emotiv EPOC EEG headset (highlighted in green in **Figure 1**), three frontal and one parietal clusters were constructed. In detail, cluster combinations and electrodes, together with the channel aggregation approach used, is shown in **Table 1**.

3.4. Formation of the Mental Workload Indexes From Clusters of EEG Alpha and Theta Bands

Generating band ratios from EEG channels over time follows the methodology used in Borghini et al. (2014). The computation of the alpha-to-theta and theta-to-alpha ratios was done utilizing the average power spectral density (PSD) values from the alpha band from the cluster $c - \alpha$ and the average PSD values from the theta band from clusters $c1 - \theta$, $c2 - \theta$ and $c3 - \theta$ as outlined in **Table 1**. Three alpha-to-theta and theta-to-alpha ratios are set to take different clusters from frontal electrodes and one cluster from parietal electrodes. The computation of the band ratios are given as follows:

$$\frac{\alpha}{\theta} = \frac{avg(\forall e \in c - \alpha)}{avg(\forall e \in c_x - \theta)} \quad (1)$$

$$\frac{\theta}{\alpha} = \frac{avg(\forall e \in c_x - \theta)}{avg(\forall e \in c - \alpha)} \quad (2)$$

where, $c - \alpha$ and $c_x - \theta$ are the respective alpha and theta clusters (from **Table 1**), with e an electrode in them, and x a cluster among those using the theta band ($c1 - \theta, c2 - \theta, c3 - \theta$). The combination of the clusters in **Table 1**, jointly with their individual use, led to the formation of the following possible mental workload indexes (configurations):

$$MWL\ Indexes = \{c1 - \theta, c2 - \theta, c3 - \theta, c - \alpha, at - 1, at - 2,$$
$$at - 3, ta - 1, ta - 2, ta - 3\} \quad (3)$$

where, $at - 1 = \frac{c - \alpha}{c1 - \theta}$, $at - 2 = \frac{c - \alpha}{c2 - \theta}$, $at - 3 = \frac{c - \alpha}{c3 - \theta}$, $ta - 1 = \frac{c1 - \theta}{c - \alpha}$, $ta - 2 = \frac{c2 - \theta}{c - \alpha}$ and $ta - 3 = \frac{c3 - \theta}{c - \alpha}$. In this study, a 1 s non-overlapping sliding widow technique is employed to segment long EEG data, and for each window, an index of mental workload can be calculated.

3.5. Feature Extraction From Indexes and Selection

Extracting high-level features from MWL indexes is crucial since it allows the finding of distinguishing properties that otherwise would not be possible if a raw index alone is considered. The extraction of such high-level features from the indexes defined in the set Equation 3 is executed using the TSFEL (Time Series feature Extraction Library) (Barandas et al., 2020). The advantage of using TSFEL is that it offers a wide range of statistical properties that can be extracted from multiple domains, including those from frequency and temporal domains. It is useful for identifying peculiar aspects of a signal and its specific properties such as variability, slope or peak to peak, just to name a few. Classes of extracted features span from the most well-known such as statistical/spectral kurtosis, mean and median of a signal, to less frequently employed features such as human range energy ratio, the estimator of the cumulative distribution function (ECDF), variability and peak-to-peak. The idea behind considering a large number of initial features was to assess their individual importance, and subsequently retain only the most informative ones by adopting a systematic feature selection approach rather than selecting them subjectively from intuition. Feature reduction can also facilitate model training in terms of required computational time. The selection criteria were based on the "SelectKBest" feature selection algorithm that ranks the features with the largest ANOVA F-value between a feature vector and a class label. The reason for choosing such an approach is that it offers a better trade-off in terms of accuracy, stability and stopping criteria in comparison to other feature selection algorithms such as SelectPercentile or VarianceThreshold (Powell et al., 2019). Determining the threshold for an optimal number of features is an iterative process of supervised evaluation of model performance with variable numbers of features. Initially, a model with all features was built and its performance metric in terms of accuracy was observed and in subsequent steps, the number of features was reduced by half iteratively as long as the model performance increased. The following iterative step with number of features that would indicate a decrease in model performance served as a stopping criteria. Finally, a Pearson correlation was computed to assess the correlation between selected features in order to reduce multicollinearity among them. Reducing multicollinearity of features is an essential step for retaining the predictive power of each of them. Using highly correlated features very often hamper model training. Experiments conducted by Lieberman and Morris (2014) indicate a correlation threshold of ± 0.5 for optimal model performance.

3.6. Models Training

The modeling and training process aims at learning classification models capable of discriminating self-reported mental workload scores from subjects (target feature), given the features extracted and selected in the previous step D (independent features). The mental workload scores were selected rather than the type of condition ("Simkap" or "Rest") because we wanted a sensible indicator of mental workload, not a task load condition. In other words, a self-reported indicator of mental workload can be considered a more reliable representation of the user experience

than a class representing a certain task load condition. This argument is originated by the fact that, in both task load conditions, users can experience any level of cognitive load. For example, a novice user can experience high mental workload for an easy task load condition when compared to an expert user. Similarly, a skilled user can experience moderate mental workload even while in a resting condition because of significant mind wandering. In research from Charles and Nixon (2019), a distinction between objective elements of the work (taskload) from the subjective perception of mental workload is outlined. Both taskload and subjective perception of mental workload can be mediated by operator experience or time constraint factors. Therefore, it is intuitive that task load conditions are not equivalent to mental workload experiences. In fact, on one hand, the former are strictly defined prior task execution, and are static, meaning they are immutable during task execution. On the other hand, the latter are unknown prior task execution and can change depending on a number of factors, for example including user's prior knowledge, motivation, time of execution, fatigue, stress among the others. To stress further, research has clearly shown that even the same person can execute a task, designed with a specific, static load condition (pre-defined task demands) differently at various times of the day (Hancock et al., 1992).

Additionally, to facilitate subsequent interpretation, we treated model training as a binary classification problem, mainly to use more interpretable evaluation metrics such as precision, recall, accuracy and f1 score. Therefore, the target feature range of 1–9 of the self-reported mental workload scores was mapped into two levels of mental workload, the "suboptimal MWL" and "super optimal MWL." The split was adopted based on the assumption of the parabolic relationship between experienced mental workload and performance as outlined in Longo and Rajendran (2021). This split was done by aggregating the scores from 1 to 4, representing some degree of low mental workload (effort), into the "suboptimal MWL," and all the scores from 6 to 9, for all of those supporting some degree of high mental workload (effort), into the "super optimal MWL." All those scores rated five were discarded because they represent the neutral experience of mental workload.

The learning techniques chosen for achieving this aim are Logistic regression (L-R), Support Vector Machines (SVM) and Decision Trees (DTR). Many research works have considered these three learning techniques for continuous and more prolonged EEG recordings (Berka et al., 2007; Hu and Min, 2018; Doma and Pirouz, 2020). Logistic regression and SVM, as error-based learning techniques, are suitable for binary classification tasks (as in this work). On the other hand, as an information-based technique, decision trees are suitable for distinguishing important features by calculating their information gains during model training.

Due to the fact that a small dataset of 48 subjects was selected, then repeated montecarlo sampling for model training and validation is set in the following order:

1. A randomized 70% of subjects is selected both from the "suboptimal MWL" and the 'super optimal MWL' classes (dependent feature) for model training;
2. The remaining 30% was kept for model testing.

3. The above splits are repeated for 100 iterations to observe random training data, and effectively capture the probability density of the target variable.

A general rule of thumb implies a minimum of $1/5th$ ratio for each feature in the data to increase model accuracy (Friedman, 1997). Given the low number of training instances in each of the target classes ("suboptimal MWL," "super optimal MWL"), the "curse of the dimensionality" problem is anticipated (Verleysen and François, 2005). Therefore, a strategy for generating synthetic data is adopted, which is based on the generation of statistically similar synthetic data that mimics the original data. For this purpose, the Synthetic Data Quality Score based on metrics like Field Correlation Stability, Deep Structure Stability and Field Distribution Stability (Gretel.ai, 2022) is adopted.

The Field Correlation Stability is the correlation between every pair of independent features (fields) in the training data and then in the synthetic data. These values' absolute difference is computed and averaged across all independent features. The lower this average, the higher is the correlation stability of the synthetic data. Deep Structure Stability verifies the statistical integrity of the generated dataset by performing deep, multi-field analysis of distributions and correlations. This is done by executing Principal Component Analysis (PCA) on the original data, and comparing it against that on the synthetic data. A synthetic quality score is created by comparing the distributional distance between the principal components found in the two datasets. The closer the principal components, the higher the quality of the synthetic data. Field distribution stability measures how closely the field distribution in the synthetic data mimic that on the original data. The comparison of two distributions is done using the Jensen-Shannon (JS) distribution distance given as:

$$JSD = H(M) - \frac{1}{2}(H(O) - H(S)) \qquad (4)$$

where $H(O)$ and $H(S)$ are the Shannon entropy values for original (O) and synthetic (S) data respectively and $H(M)$ is the sum of selected weights for probability distributions (π) and dataset probabilities (P) given as $M = \sum_{i=1}^{2} \pi_i P_i$. The lower the distance score on average across all fields, the higher the Field Distribution Stability quality score and consequently the higher the quality of the synthetic data generated.

The Synthetic Data Quality Score represents an arithmetic mean between field correlation stability, deep structure stability and field distribution stability. In this sense, the Synthetic Data Quality Score can be viewed as a confidence score as to whether scientific conclusions drawn from the synthetic dataset would be indistinguishable as if they were to be used in the original data. Synthesizing new data is performed using synthetic generators offered from Gretel.ai[1]. The training process for the combined (original + synthetic) uses the same montecarlo sampling with the same steps as with original data outlined above. Randomized 70% of subjects is selected both from the combined (original + synthetic) for the "suboptimal MWL" and the 'super optimal MWL' classes (dependent feature) for model training, the remaining 30% of combined (original + synthetic) subjects

---

[1]Gretel.ai - Privacy Engineering as a Service for Data Scientists - https://gretel.ai.

was kept for model testing and performing 100 iteration through these randomized splits. During model training, the data was normalized using $z - score$ normalization given as $z = (x - \mu)/\sigma$, where μ is the mean of training samples and σ is the standard deviation. The rationale for using $z - score$ normalization is that it tends to minimize the mean ($\mu = 0$) and maximize the standard deviation ($\sigma = 1$) for the normalized value and makes it suitable reducing extremely peak values in data, by transforming it in such a way that it's no longer a massive outlier.

3.7. Models Evaluation

A set of evaluation metrics were employed to assess the ability of the selected models to generalize on unseen data by learning from the training data. These metrics can be used to measure and summarize the quality of the trained models when tested with previously unseen data. For a binary classification problem, such as in the case, the evaluation of the models is dependent on *True Positives* (*tp*) and *True Negatives* (*tn*) which denote the number of positive and negative instances that are correctly classified. It can be also conducted with the *False Positives* (*fp*) and *False Negatives* (*fn*) that denote the number of miss-classified negative and positive instances, respectively. According to this, several metrics are used to evaluate the performance of the trained models. *The accuracy metric* measures the ratio of correct predictions over the total number of evaluated instances. Accuracy is represented as, $Accuracy = (tp + tn)/(tp + fp + tn + fn)$. *Precision* is used to measure the positive instances that are correctly predicted from the total predicted instances in a positive class, given as $Precision = (tp)/(tp + fp)$. *Recall* measures the fraction of positive instances that are correctly classified, $Recall = (tp)(tp + tn)$. *F-Measure* or *f1-score* is the harmonic mean between recall and precision values represented as, $f1 - score = (2 \cdot Precision \cdot Recall)/(Precision + Recall)$. The proposed evaluation metrics are essential to assess the robustness of the selected models built upon high-level features extracted from the MWL indexes toward the discrimination of self-reported perceptions of mental workload. While precision refers to the percentage of relevant instances, recall refers to the rate of total relevant instances correctly classified by the model. The best model minimizes the value of (*fp*) in precision and (*tn*) in recall, and both come at the cost of each other since we cannot minimize both of them in one metrics. f1-score represent a harmonic mean of precision and recall and takes into account both metrics. Consequently, to bring hypotheses H1 and H2 on provable grounds, the $f1 - score$ metric is adopted too.

4. RESULTS

The results section follows the same order of steps as outlined in the design section.

4.1. EEG Artifact Removal

Artifact removal is performed on each EEG signal for each of the 48 subjects separately for the "Rest" and "Simkap" task load conditions. The average number of ICA components removed from the EEG data associated with each subject is 1.61 for the "Rest" and 1.46 for the "Simkap" condition. The number of

removed artifacts is within limits of the adopted methodology described in Nolan et al. (2010). **Figure 2** depicts the removal occurrence for a total of 14 components across all 48 users for "Rest" condition and "Simkap" condition. As it is possible to see from **Figure 2**, at most, one ICA component per subject that is significantly different from the other components (± 3 standard deviations) exists. These components are removed by zero-ing them, and the EEG multi-channel data is subsequently reconstructed by applying inverse ICA. Since at least one bad component was identified and removed for most subjects, it is possible to reasonably claim that some artifact has been removed from the EEG signal, thus facilitating the subsequent computations of the alpha and theta bands.

4.2. Evaluation of Feature Extraction and Selection

All high-level features are collected from the statistical properties of the mental workload indexes in various domains, including the temporal and frequency domains. The initial number of collected features are 210, and the exhaustive list is provided in the **Supplementary Material** accompanying this article. The ANOVA *F*-Value is computed for each of these features, and those with the highest value are retained for subsequent model training.

Since the SelectKBest algorithm requires an initial number of features, an iterative approach of feature inclusion during model training and the performance of the models with those features is assessed through its accuracy. Iterative optimal feature selection is performed by employing data from the original dataset. **Figure 3** illustrates the convergence on the optimal number of features in relation to model performance grouped by learning techniques (L-R, SVM, DTR). This resulted in a reduced number of features that are kept for model training by employing to the dataset enhanced with synthetic data. **Figure 4** shows the Pearson correlation matrix for the "Rest" and the "Simkap" states for the alpha-to-theta ratios for the case of index *at*-1 (as designed in Section 3.3).

Noticeably, most of the features are in the correlation range between -0.5 and $+0.5$, which contributes to reduce multicollinearity and thus potentially being all relevant and with high prediction capability (Lieberman and Morris, 2014). **Figure 4** is an illustration of the results associated to a single mental workload index (at-1). However, results associated to the other indexes are mostly consistent with these, as it is possible to examine in the **Supplementary Figures S1–S9** accompanying this article.

4.3. Evaluation of the Training Set Across Indexes

After the feature selection process, training of the models was conducted with Montecarlo sampling using Logistic Regression (L-R), Support Vector Machines (SVM) and Decision Trees (DTR) as described in design subsection E). Model training suffered from the "curse of dimensionality" issue since it comprised 48 subjects across only the seven selected features. The number of training instances is low compared to the number

FIGURE 2 | The number of components removed across all 48 subjects for "Rest" and "Simkap" task load conditions.
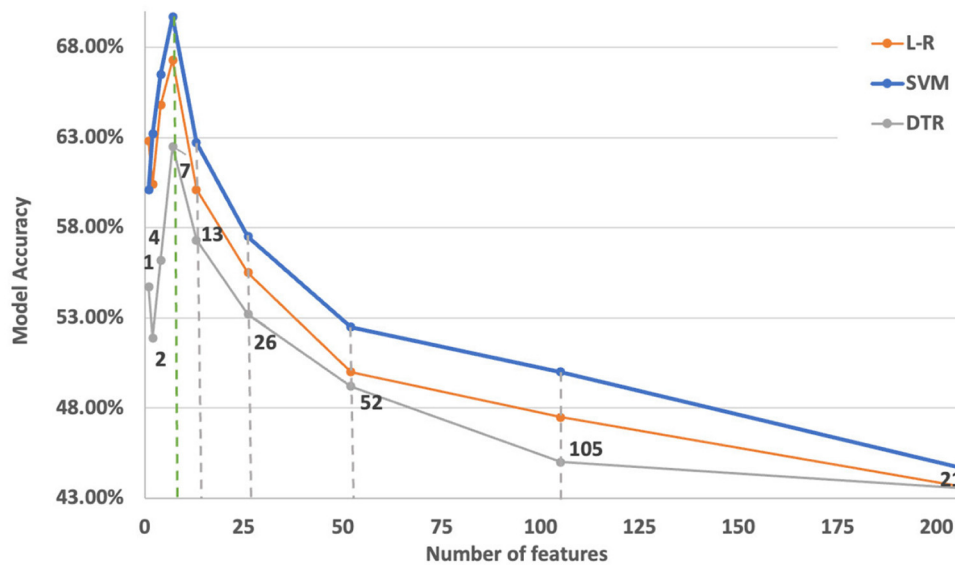

FIGURE 3 | Optimal number of features against model performance with data coming from the Simultaneous Task EEG workload (STEW) dataset. The dashed lines indicate the number of feature considered in the iteration. It can be seen that the optimal number of top features to select is around seven indicated with green dashed line which also acts as a stopping criteria.

of independent features retained for modeling purposes. This is followed by the peak phenomenon of feature inclusion, where the number of features and their cumulative discriminatory effect is essential for the average predictive power of a classifier, which is data-dependent (Zollanvari et al., 2019). The initial model evaluation with test data on the original dataset did not reveal the accuracy of more than 60% for the standalone mental workload indexes built upon the alpha and theta indexes

alone ($c1 - \theta$, $c2 - \theta$, $c3 - \theta$, $c - \alpha$). An accuracy of 70% for the mental workload indexes built upon the alpha-to-theta and theta-to-alpha ratios was observed. An in-depth analysis of the learning curves associated to the classifiers indicated a model underfitting and an inability to generalize from test data. Moreover, analyzing the spectral entropy of the mental workload indexes revealed a small variance variation, as can be seen from the boxplots of **Figure 5**. Small data variance subsequently
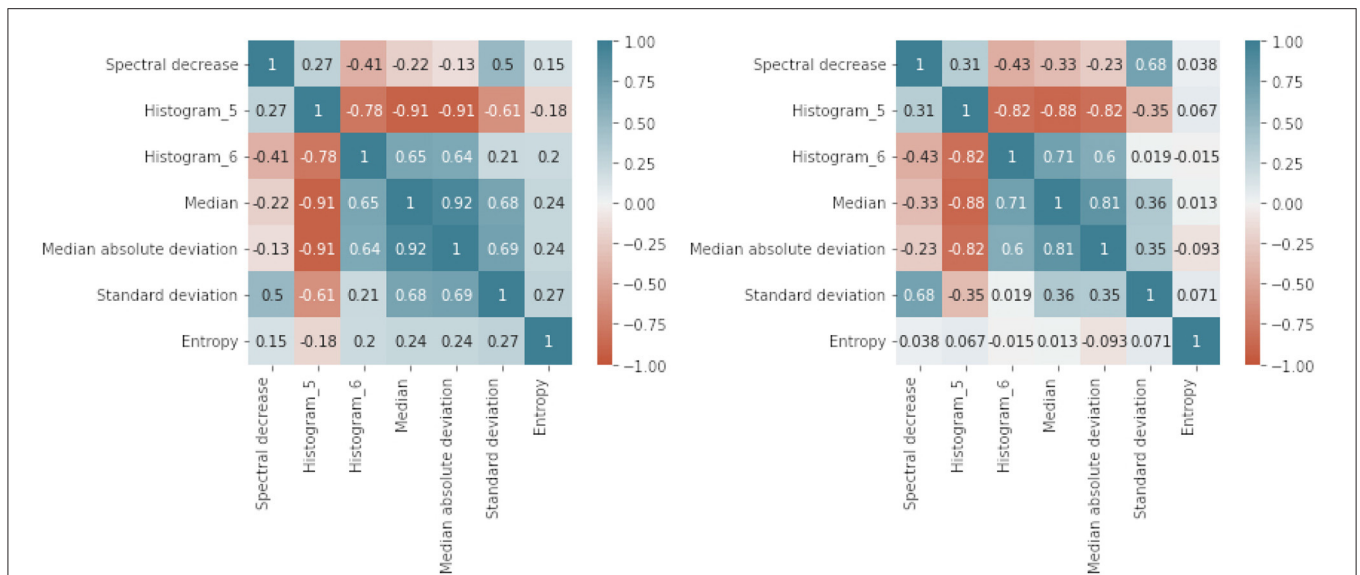
FIGURE 4 | Pearson correlation coefficients matrix for the case of MWL index - at-1: Rest **(Left)** and Simkap **(Right)** task-load conditions. at-1: alpha-to-theta ratios between the indexes $c - \alpha$ and $c1 - \theta$. The scale on the right of the image indicates the Pearson correlation coefficients range.
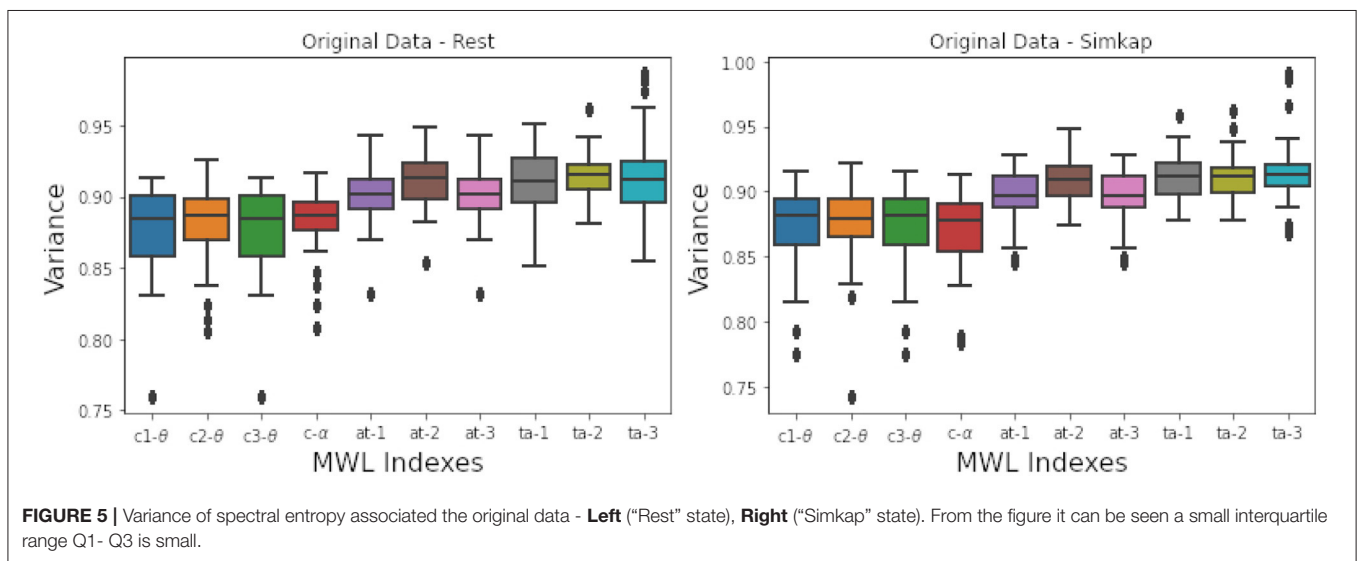

FIGURE 5 | Variance of spectral entropy associated the original data - **Left** ("Rest" state), **Right** ("Simkap" state). From the figure it can be seen a small interquartile range Q1- Q3 is small.

increases the bias influencing the model's ability to generalize. Thus, as expected, synthetic data generation was applied for training robust models.

4.4. Synthetic Data Evaluation

The input for data synthesis was the initial dataset comprised of 48 subjects and 150 data points (2.5 min of EEG data per participant split into segments of 1 s) for each of the indexes designed in Equation 3. Two synthetic datasets are created, one for the "Rest" and one for the "Simkap" task load conditions in order to retain the original dataset's intrinsic properties. **Table 2**

illustrates the overall synthetic quality scores for the mental workload indexes set in Equation 3.

Findings suggest that the overall synthetic data score is always above 87% throughout all the mental workload indexes selected for the comparative analysis. The synthetic quality score was measured in the scale (1–20)%-Very Poor, (20–40)%-Poor, (40–60)%-Moderate, (60–80)%-Good and (80–100)%-Excellent. This suggests that the quality of synthesized data is excellent, and in line with similar studies (Hernandez-Matamoros et al., 2020). Consequently, the synthesized data was of an additional 180 synthesized subjects with 150 data points (2.5 min of EEG activity split into 150 segments of 1 s) for each mental workload index. The final combined dataset with original and synthesized data

TABLE 2 | Synthetic score for different mental workload indexes and two task load conditions ("Rest" and "Simkap").

MWL index:	Field corr. stability		Deep struct. stability		Field dist. stability		Overall score	
	Rest (%)	Simkap (%)	Rest (%)	Simkap (%)	Rest (%)	Simkap (%)	Rest (%)	Simkap (%)
$c1 - \theta$	100	100	95	98	78	70	91	89
$c2 - \theta$	100	96	94	94	78	71	91	87
$c3 - \theta$	94	95	97	100	78	80	89	91
c-α	100	94	97	100	72	81	90	91
$at - 1$	91	94	89	92	73	74	84	87
$at - 2$	100	92	93	94	76	78	90	88
$at - 3$	93	100	95	94	72	74	87	89
$ta - 1$	100	100	93	95	77	75	90	90
$ta - 2$	91	100	96	88	76	73	88	87
$ta - 3$	100	96	85	94	80	82	88	90

is now comprised of 228 subjects with 150 data points for each mental workload index, as defined in set Equation 3.

4.5. Validation of Models for Discriminating Self-Reported Perceptions of Mental Workload

The training of the models with Logistic Regression and Support Vector Machines learning techniques utilized the linear optimizer since it offers speed and optimum convergence on minimizing a multivariate function by solving univariate optimization problems during repeated training of the model (Fan et al., 2008). In the case of model training with Decision Tree, a Gini index was used to measure the quality of split during the model build.

The classifiers performance is shown in **Figure 6**. The evaluation metrics are shown across all mental workload indexes and are presented in descending order. The best classification accuracy results are observed for those models built with Support Vector Machines (SVM) and Logistic regression (L-R). In order to acknowledge the best learning technique, a two-tailed t-test between the three learning techniques and the employed evaluation metrics was performed. The results indicated no statistically significant difference between Logistic Regression (L-R), Support Vector Machines (SVM) or Decision Trees (DTR). This indicates the validity of the training approach adopted from the design, which means that no matter the learning technique adopted, the results across all applied evaluation metrics are the same. **Table 3** illustrates the p-value significance levels of the t-test between evaluation metrics for each learning technique used in the study. The t-test was conducted with a threshold confidence value of $\alpha = 0.05$.

Further analysis of mental workload indexes of the alpha-to-theta ratios ($at - 1$, $at - 2$) indicates better performance than their respective individual counterparts used for computing those ratios ($c1 - \theta$, $c2 - \theta$ and $c - \alpha$). For example, in the case of all learning techniques (L-R, SVM and DTR), first two alpha-to theta ratio indexes ($at - 1$ and $at - 2$) show better performance than their individual counterparts ($c1 - \theta$, $c2 - \theta$ and $c - \alpha$).

In the case of the theta-to-alpha mental workload indexes, this is also seen in the first two indexes ($ta - 1$ and $ta - 2$). **Figure 7** illustrates the performance of band ratios (alpha-to-theta and theta-to-alpha) across all evaluation metrics, given as density plots for the case of Support Vector Machines (SVM). The density plots for all other learning techniques are available on the **Supplementary Figures S10–S11**. **Table 4** outlines the significance levels of a two-tailed t-test between the alpha-to-theta and theta-to-alpha ratio indexes with indexes used to construct those ratios. A comparison analysis of models average performance between original data and those enhanced with synthetic data is shown in **Table 5**. An analysis of the number of electrodes across alpha and theta bands as given in **Table 1** outlined in the design Section 3 we can see a higher number of electrodes in indexes $c1 - \theta$ and $c3 - \theta$ in comparison to indexes $c2 - \theta$ and $c - \alpha$. To asses the impact of the number of electrodes in overall performance of the models, a cross-plotting between indexes $at - 1$ vs. $at - 2$ and $at - 3$ vs. $at - 2$ as well $ta - 1$ vs. $ta - 2$ and $ta - 3$ vs. $ta - 2$ is analyzed. **Figure 8** illustrates this cross density plot comparison of performance between the alpha-to-theta and theta-to-alpha ratio indexes. Furthermore, a two-tailed significance test between these band ratio indexes ($at - 1$ and $at - 3$ vs. $at - 2$ as well as $ta - 1$ and $ta - 3$ vs. $ta - 2$) reveals a statistically significant difference. **Table 6** presents the p-value significance levels for confidence interval of α=0.05. The p-values are with Bonferroni correction applied, resulting in a significance level set at $\alpha = 0.005$.

5. DISCUSSION

Rapid advancements in various tools and technologies introduced new perspectives in using EEG signals to classify task load conditions using machine learning techniques. The analysis done so far on EEG frequency bands, specifically alpha and theta bands, seems to correlate the changes in these bands to task load (Gevins and Smith, 2003; Borghini et al., 2014).
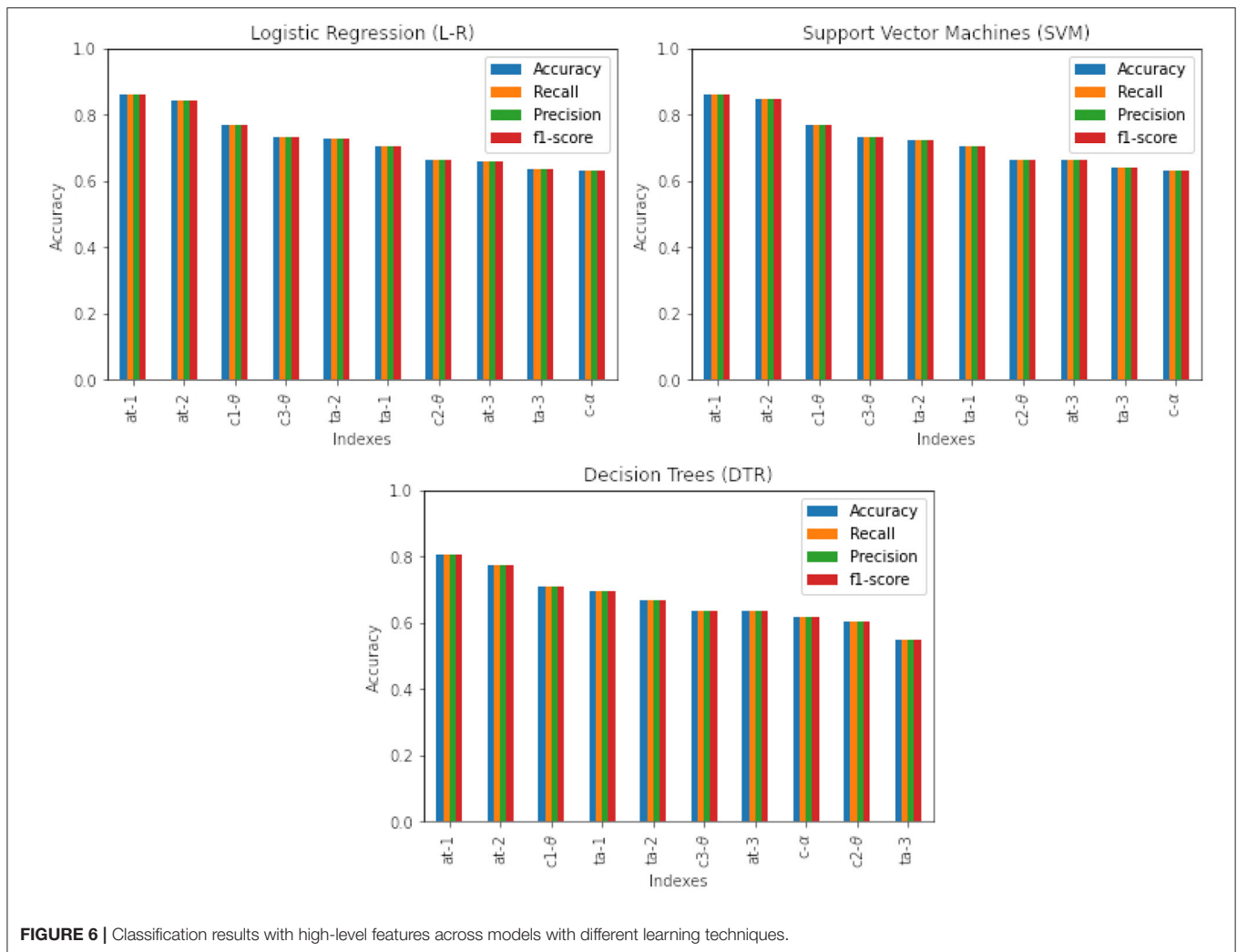
FIGURE 6 | Classification results with high-level features across models with different learning techniques.

TABLE 3 | The two-tailed *t*-test between L-R, SVM and DTR and f1-score, accuracy, recall and precision.

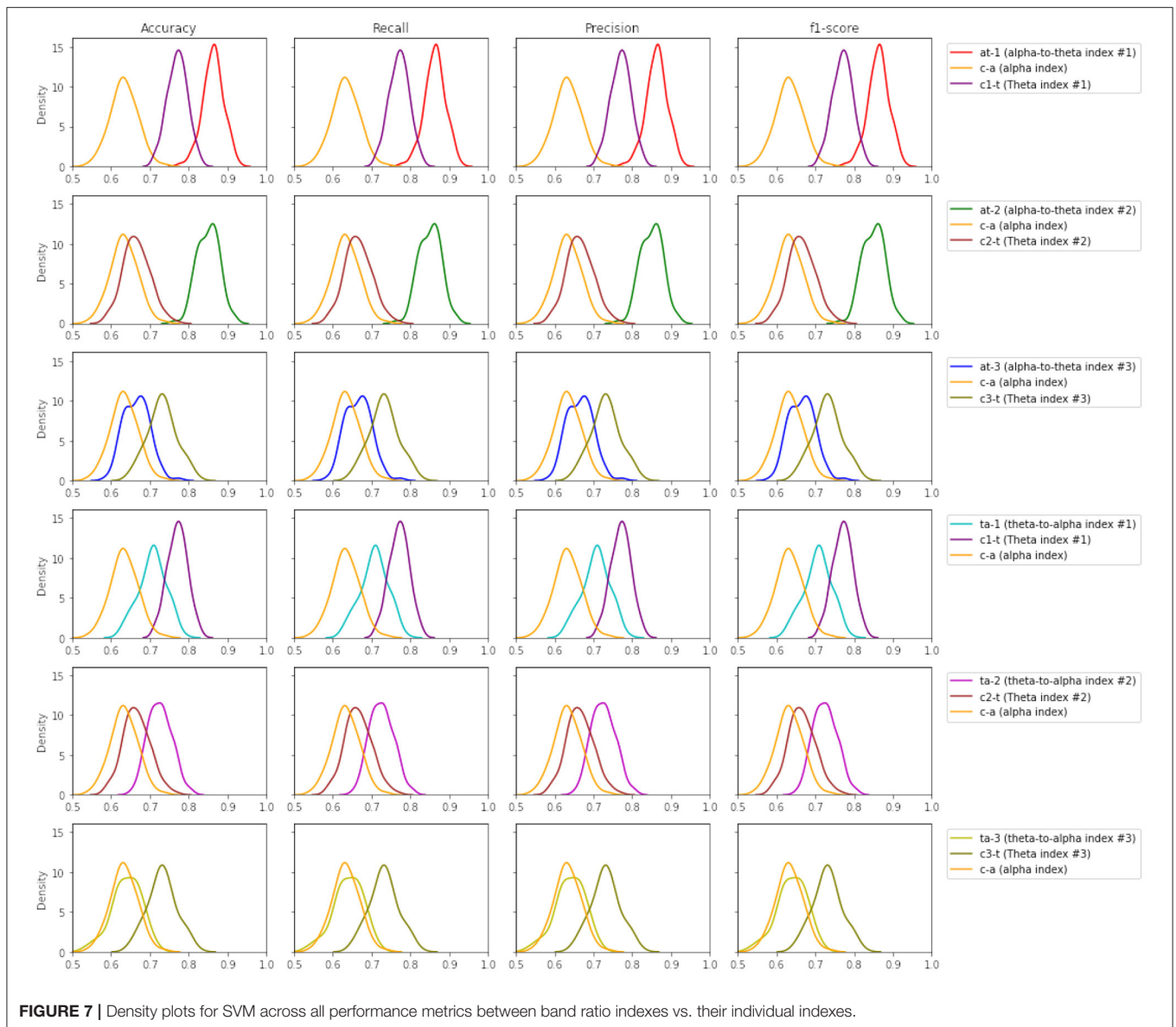|	L-R - SVM	SVM-DTR	L-R - DTR
	t-statistic (*p*-value)		
f1-score	−0.04 (0.96)	1.55 (0.13)	1.50 (0.15)
Accuracy	−0.05 (0.96)	1.55 (0.14)	1.50 (0.15)
Precision	−0.04 (0.96)	1.55 (0.14)	1.50 (0.14)
Recall	−0.04 (0.96)	1.55 (0.14)	1.51 (0.15)

Researchers face many problems in using EEG band ratios for the purpose of mental workload modeling: (i) the limited amount of participants for each conducted empirical experiment (ii) a clear definition of mental workload (iii) a clear EEG measure of mental workload.

In detail, the three aforementioned issues can be overcome and this article is a testament for such a claim. In fact, this research work demonstrates how the first issue can be

tackled by using modern deep-learning methods for synthetic data generation, giving the possibility to expand the often limited cardinality of existing datasets created with EEG data. It also contributes to tackle the second issue by advancing the understanding of mental workload as a construct by means of an empirical experiment with EEG data. In particular, it performs a construction of indexes of mental workload by employing the alpha and theta EEG bands individually and in combination, and the extraction of statistical features from these indexes for the discrimination of self-reported perceptions of mental workload.

Results show that, from an initial highest accuracy of 60% for the individual alpha and theta indexes on the original dataset, we witnessed an increase between 8 and 20% in classifier performance when this data has been augmented with synthetic data.

Regarding mental workload ratio indexes, especially the alpha-to-theta indexes, it was possible to build models with minimum 18.4 − 30.2% higher performance (as measured by f1-score, accuracy, precision, recall) than the other indexes. Furthermore, the results show that mental workload indexes

FIGURE 7 | Density plots for SVM across all performance metrics between band ratio indexes vs. their individual indexes.

$at - 1$, $at - 2$ and $ta - 1$, $ta - 2$ can better discriminate self-reported perceptions of mental workload in comparison to their individual counterparts ($c - \alpha$, $c1 - \theta$ and $c2 - \theta$). This proves our hypothesis **H1** given earlier that alpha-to-theta and theta-to-alpha ratios can significantly discriminate self-reported perceptions of mental workload than the individual use of EEG band power and can be used in designing highly accurate classification models. The accuracy, f1-score, recall and precision evaluation metrics indicate a good classification across almost all alpha-to-theta and theta-to-alpha indexes.

One interesting observation is the impact of the number of electrodes in the selected indexes on the overall accuracy of the classifiers. For example, it can be seen from **Table 1** that $c1 - t$ from theta band has a higher number of electrodes that contribute to the computation of band ratios and indicate the higher accuracy in both alpha-to-theta and theta-to-alpha indexes. Given the results from **Figure 8** and **Table 6**, hypothesis

H2 cannot be conclusively proven that the number of electrodes used for calculating alpha-to-theta and theta-to-alpha better effectuate the predictive power of the classifiers. **Figure 8** indicates better performance of $at - 2$ and $ta - 2$ indexes which are computed from $c2 - \theta$ and $c - \alpha$ individual indexes that, if seen from **Table 1** have lesser electrode numbers. One potential explanation hypothesized by authors lies in the nature of the experiment performed while collecting STEW datasets' EEG recordings, where "Rest" and "Simkap" activities are performed in sequence one after the other. Some research indicates a strong correlation between EEG frequency patterns and the relative levels of distinct neuromodulators (Vakalopoulos, 2014). This sudden change in task load activity may lead to neuromodulation in the parietal region and neuronal suppression on the frontal cortical region, resulting in better performance of band ratio indexes ($at - 2$ and $ta - 2$) with a smaller number of electrodes. Further research is required to validate this claim.

TABLE 4 | The two-tailed t-test between the alpha-to-theta and theta-to-alpha ratio indexes with their individual indexes with Bonferroni(†) correction applied, resulting in a significance level set at $\alpha = 0.005$.

		$c1 - \theta$	$c2 - \theta$	$c3 - \theta$	$c - \alpha$
			t-statistic (*p*-value)		
L-R	$at-1$	$25.1 \ (1.96 \cdot 10^{-63})** \ (2.36 \cdot 10^{-62})^\dagger$	-	-	$51.4 \ (1.32 \cdot 10^{-111})** \ (1.56 \cdot 10^{-115})^\dagger$
	at-2	-	$42.3 \ (3.94 \cdot 10^{-101})** \ (4.73 \cdot 10^{-100})^\dagger$	-	$46.8 \ (4.82 \cdot 10^{-109})** \ (5.78 \cdot 10^{-108})^\dagger$
	at-3	-	-	$15.5 \ (3.93 \cdot 10^{-36})** \ (4.71 \cdot 10^{-35})^\dagger$	$4.90 \ (1.95 \cdot 10^{-6})** \ (2.34 \cdot 10^{-5})^\dagger$
	ta-1	$14.3 \ (1.49 \cdot 10^{-32})** \ (1.79 \cdot 10^{-31})^\dagger$	-	-	$15.1 \ (4.40 \cdot 10^{-35})** \ (5.28 \cdot 10^{-34})^\dagger$
	ta-2	-	$14.9 \ (2.96 \cdot 10^{-34})** \ (3.56 \cdot 10^{-33})^\dagger$	-	$20.7 \ (1.07 \cdot 10^{-51})** \ (1.28 \cdot 10^{-50})^\dagger$
	ta-3	-	-	$19.4 \ (5.81 \cdot 10^{-48}) \ (6.97 \cdot 10^{-47})^\dagger$	$0.49 \ (0.62) \ (1)^\dagger$
SVM	at-1	$25.0 \ (1.99 \cdot 10^{-63})** \ (2.39 \cdot 10^{-62})^\dagger$	-	-	$52.4 \ (4.40 \cdot 10^{-118})** \ (5.28 \cdot 10^{-117})^\dagger$
	at-2	-	$42.2 \ (4.48 \cdot 10^{-99})** \ (5.37 \cdot 10^{-98})^\dagger$	-	$48.1 \ (2.69 \cdot 10^{-111})** \ (3.22 \cdot 10^{-110})^\dagger$
	at-3	-	-	$13.2 \ (4.75 \cdot 10^{-29})** \ (2.79 \cdot 10^{-28})^\dagger$	$6.68 \ (2.39 \cdot 10^{-29})** \ (5.71 \cdot 10^{-28})^\dagger$
	ta-1	$14.8 \ (4.24 \cdot 10^{-34})** \ (5.01 \cdot 10^{-33})^\dagger$	-	-	$14.8 \ (6.07 \cdot 10^{-34})** \ (7.29 \cdot 10^{-33})^\dagger$
	ta-2	-	$13.0 \ (1.98 \cdot 10^{-28})** \ (2.37 \cdot 10^{-27})^\dagger$	-	$19.9 \ (2.74 \cdot 10^{-49})** \ (3.29 \cdot 10^{-48})^\dagger$
	ta-3	-	-	$17.0 \ (1.10 \cdot 10^{-40}) \ (1.32 \cdot 10^{-39})^\dagger$	$1.21 \ (0.22) \ (1)^\dagger$
DTR	at-1	$20.10 \ (1.02 \cdot 10^{-49})** \ (1.22 \cdot 10^{-48})^\dagger$	-	-	$33.44 \ (1.18 \cdot 10^{-83})** \ (2.69 \cdot 10^{-82})^\dagger$
	at-2	-	$33.9 \ (2.06 \cdot 10^{-84})** \ (2.47 \cdot 10^{-83})^\dagger$	-	$27.76 \ (3.43 \cdot 10^{-70})** \ (4.12 \cdot 10^{-69})^\dagger$
	at-3	-	-	$-0.57 \ (0.56) \ (0.15)^\dagger$	$2.51 \ (0.01) \ (1)^\dagger$
	ta-1	$2.49 (0.01) \ (0.15)^\dagger$	-	-	$13.20 \ (5.73 \cdot 10^{-29})** \ (6.88 \cdot 10^{-28})^\dagger$
	ta-2	-	$11.6 \ (2.75 \cdot 10^{-24})** (3.30 \cdot 10^{-23})^\dagger$	-	$8.07 \ (6.16 \cdot 10^{-14})** \ (7.93 \cdot 10^{-13})^\dagger$
	ta-3	-	-	$14.6 \ (2.83 \cdot 10^{-33})** \ (3.40 \cdot 10^{-32})^\dagger$	$11.21 (6.54 \cdot 10^{-23})** \ (7.85 \cdot 10^{-22})^\dagger$

TABLE 5 | Models performance increase across mental workload indexes between original dataset and dataset combined with synthetic data.

	Data	$c1 - \theta(\%)$	$c2 - \theta$ (%)	$c3 - \theta(\%)$	$c - \alpha(\%)$	$at - 1(\%)$	$at - 2(\%)$	$at - 3(\%)$	$ta - 1(\%)$	$ta - 2(\%)$	$ta - 3(\%)$
f1-score	Orig.	57.6	55.9	58.8	38.4	53.8	56.0	44.1	63.1	56.4	71.0
	Orig.+Synth.	74.2	64.6	70.9	62.8	84.5	82.2	65.2	70.3	70.7	60.8
Accuracy	Orig.	57.7	56.6	60.0	50.0	53.3	56.6	47.7	63.3	56.6	71.0
	Orig.+Synth.	74.8	64.7	70.0	62.6%	84.5	82.2	65.2	70.3	70.7	60.8
Precision	Orig.	57.7	56.7	61.3	33.3	65.25	57.2	47.5	63.8	56.6	71.1
	Orig.+Synth.	74.2	64.7	70.6	62.6	84.6	82.2	65.2	70.3	70.7	60.8
Recall	Orig.	57.7	56.6	60.0	50.0	54.4	56.6	47.7	63.3	56.6	71.0
	Orig.+Synth.	74.8	64.5	70.0	62.0	84.5	81.23	65.2	70.3	70.7	60.8
Avg. Increase:		17.2	8.0	10.0	19.9	30.2	25.5	18.4	6.9	14.1	f–10.1

Based on the results above, we can conclude that EEG band ratios, alpha-to-theta and theta-to-alpha ratio mental workload indexes, can significantly discriminate self-reported perceptions of mental workload and be used to design models for detecting such levels of mental workload perception. The observations however cannot conclusively prove that the higher the electrode number, especially in the parietal region, leads to a better discrimination self-reported perceptions of mental workload.

6. CONCLUSION

Various EEG frequency bands indicate a direct correlation to human mental workload. In particular, EEG bands such as alpha and theta bands tend to increase/decrease in the state of mental workload (Borghini et al., 2014). However, no conjoint analysis of both bands in the form of indexes over time has been sufficiently analyzed so far.

This article has empirically demonstrated that EEG band ratios, specifically the alpha-to-theta and theta-to-alpha ratios can be treated as mental workload indexes for the discrimination of self-reported perceptions of mental workload. In detail, a set of higher level features associated to these indexes, have proven useful for the inductive formation of models, employing machine learning, for the discrimination of two levels of mental workload perception ("suboptimal MWL" and "super optimal MWL"). Another important contribution in this research is the analysis of the impact of electrode density in-band ratios on the formation of discriminative models of self-reported perceptions of mental workload.

Future research work will outline the usage of the alpha-to-theta and theta-to-alpha ratio indexes related to the following issues:
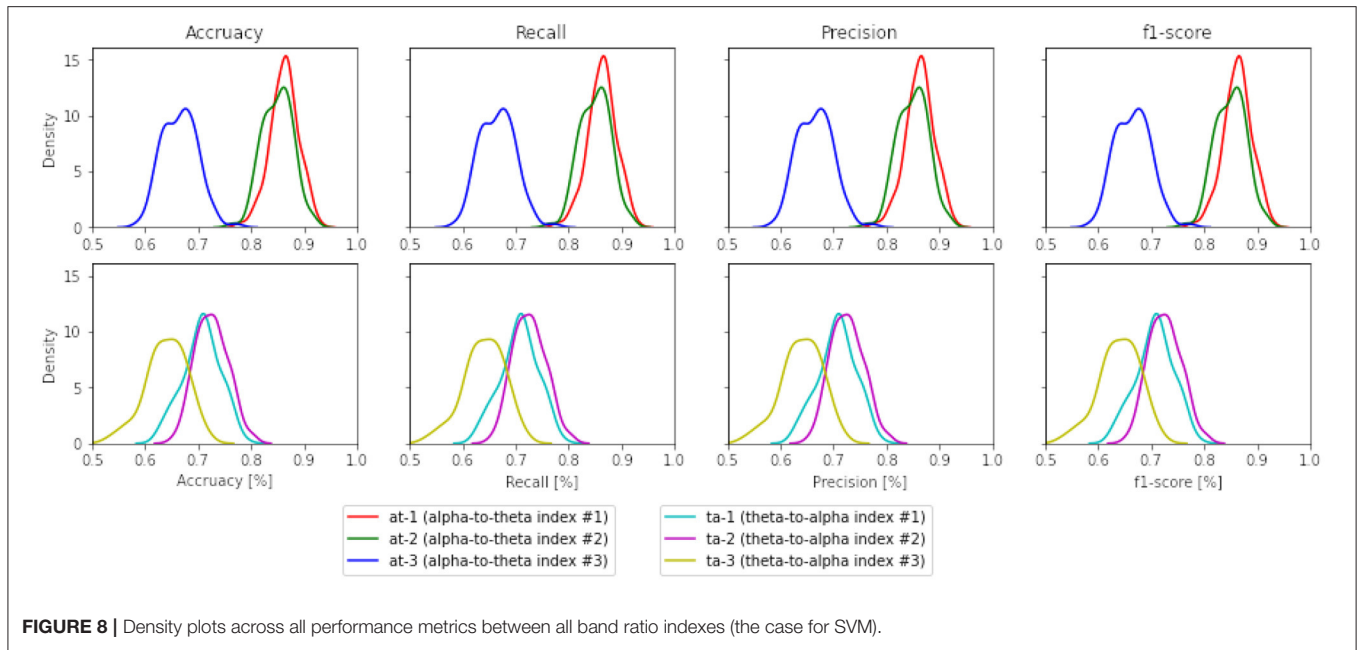
FIGURE 8 | Density plots across all performance metrics between all band ratio indexes (the case for SVM).

TABLE 6 | The two-tailed *t*-test between alpha and theta band ratios: $at-1$ vs. $at-2$: 2-tail test value between indexes $at-1$ and $at-2$.

	Learning technique	$at-1$ vs. $at-2$	$at-3$ vs. $at-2$	$ta-1$ vs. $ta-2$	$ta-3$ vs. $ta-2$
t-statistic (*p*-value)	L-R	$4.82 \, (2.79 \cdot 10^{-6})**$ $(1.11 \cdot 10^{-5})^{\dagger}$	$-42.29 \, (4.27 \cdot 10^{-101})**$ $(1.71 \cdot 10^{-100})^{\dagger}$	$-4.98 \, (1.32 \cdot 10^{-6})**$ $(5.29 \cdot 10^{-6})^{\dagger}$	$-20.09 \, (1.07 \cdot 10^{-49})**$ $(4.31 \cdot 10^{-49})^{\dagger}$
	SVM	$3.31 \, (1 \cdot 10^{-2})**$ $(4.34 \cdot 10^{-3})^{\dagger}$	$-42.31 \, (3.91 \cdot 10^{-101})**$ $(1.56 \cdot 10^{-100})^{\dagger}$	$-4.10 \, (6.00 \cdot 10^{-5})**$ $(2.40 \cdot 10^{-4})^{\dagger}$	$-17.44 \, (6.75 \cdot 10^{-42})**$ $(2.70 \cdot 10^{-41})^{\dagger}$
	DTR	$6.96 \, (4.76 \cdot 10^{-11})**$ $(1.90 \cdot 10^{-10})^{\dagger}$	$-29.21 \, (9.99 \cdot 10^{-74})**$ $(3.93 \cdot 10^{-73})^{\dagger}$	$5.73 \, (3.64 \cdot 10^{-8})**$ $(1.45 \cdot 10^{-7})^{\dagger}$	$-20.99 \, (2.93 \cdot 10^{-52})**$ $(1.17 \cdot 10^{-51})^{\dagger}$

$at-2$ vs. $at-3$: 2-tail test value between between indexes $at-2$ and $at-3$. $ta-1$ vs. $ta-2$: 2-tail test value between indexes ta-1 and ta-2. $ta-2$ vs. $ta-3$: 2-tail test value between indexes $ta-2$ and $ta-3$. The (†) sign indicate the p-value with Bonferroni correction applied, resulting in a significance level set at $\alpha = 0.005$.

- replication of the experiment conducted in this research with additional public available datasets, to further validate the contribution to knowledge.
- evaluation of human tasks different than those employed in this research, as for instance those conducted in the automobile industry (Di Flumeri et al., 2018), in the context of Human-Computer Interaction (HCI) (Longo, 2012) and in education (Longo, 2018b; Longo and Orru, 2018).
- use of multi-channel EEG data collected from a larger pool of electrodes, and thus formation and evaluation of additional mental workload indexes built with different clusters of electrodes for the alpha and theta bands.
- the design of a novel experiment with additional task load conditions of incremental complexity, for example by employing the multiple resource theory of Wickens (Wickens, 2008) and the definition of objective task performance measures that can be used as dependent features from indexes of mental workload.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

LL and BR designed the study. BR conducted the experiment. All authors reviewed and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2022.861967/full#supplementary-material

REFERENCES

Antonenko, P., Paas, F., Grabner, R., and Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22, 425–438. doi: 10.1007/s10648-010-9130-y

Asgher, U., Khalil, K., Khan, M. J., Ahmad, R., Butt, S. I., Ayaz, Y., et al. (2020). Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain-computer interface. *Front. Neurosci.* 14, 584. doi: 10.3389/fnins.2020.00584

Bagheri, M., and Power, S. D. (2021). Investigating hierarchical and ensemble classification approaches to mitigate the negative effect of varying stress state on eeg-based detection of mental workload level-and vice versa. *Brain Comput. Interfaces* 8, 26–37. doi: 10.1080/2326263X.2021.1948756

Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., et al. (2020). Tsfel: Time series feature extraction library. *SoftwareX* 11, 100456. doi: 10.1016/j.softx.2020.100456

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat Space Environ. Med.* 78, B231–B244.

Borghini, G., Aricò, P., Di Flumeri, G., Salinari, S., Colosimo, A., Bonelli, S., et al. (2015). "Avionic technology testing by using a cognitive neurometric index: a study with professional helicopter pilots," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan: IEEE), 6182–6185.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Borghini, G., Vecchiato, G., Toppi, J., Astolfi, L., Maglione, A., Isabella, R., et al. (2012). Assessment of mental fatigue during car driving by using high resolution eeg activity and neurophysiologic indices. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2012, 6442–6445. doi: 10.1109/EMBC.2012.6347469

Borys, M., Tokovarov, M., Wawrzyk, M., Wesoowska, K., Plechawska-Wjcik, M., Dmytruk, R., et al. (2017). "An analysis of eye-tracking and electroencephalography data for cognitive load measurement during arithmetic tasks," in *2017 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE)* (Bucharest: IEEE), 287–292.

Byrne, A. (2011). Measurement of mental workload in clinical medicine: a review study. *Anesthesiol. Pain Med.* 1, 90. doi: 10.5812/aapm.2045

Charles, R. L., and Nixon, J. (2019). Measuring mental workload using physiological measures: a systematic review. *Appl. Ergon.* 74, 221–232. doi: 10.1016/j.apergo.2018.08.028

Choi, M. K., Lee, S. M., Ha, J. S., and Seong, P. H. (2018). Development of an eeg-based workload measurement method in nuclear power plants. *Ann. Nuclear Energy* 111, 595–607. doi: 10.1016/j.anucene.2017.08.032

Coelli, S., Sclocco, R., Barbieri, R., Reni, G., Zucca, C., and Bianchi, A. M. (2015). "EEG-based index for engagement level monitoring during sustained attention," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan: IEEE), 1512–1515. doi: 10.1109/EMBC.2015.7318658

Dan, A., and Reiner, M. (2017). Real time eeg based measurements of cognitive load indicates mental states during learning. *J. Educ. Data Min.* 9, 31–44. doi: 10.5281/zenodo.3554719

Di Flumeri, G., Borghini, G., Aricò, P., Sciaraffa, N., Lanzi, P., Pozzi, S., et al. (2018). Eeg-based mental workload neurometric to evaluate the impact of different traffic and road conditions in real driving settings. *Front. Hum. Neurosci.* 12, 509. doi: 10.3389/fnhum.2018.00509

Diaz-Piedra, C., Sebastián, M. V., and Di Stasi, L. L. (2020). Eeg theta power activity reflects workload among army combat drivers: an experimental study. *Brain Sci.* 10, 199. doi: 10.3390/brainsci10040199

Doma, V., and Pirouz, M. (2020). A comparative analysis of machine learning methods for emotion recognition using eeg and peripheral physiological signals. *J. Big Data* 7, 1–21. doi: 10.1186/s40537-020-00289-7

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874.

Fernandez Rojas, R., Debie, E., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., et al. (2020). Electroencephalographic workload indicators during teleoperation of an unmanned aerial vehicle shepherding a swarm of unmanned ground vehicles in contested environments. *Front. Neurosci.* 14, 40. doi: 10.3389/fnins.2020.00040

Friedman, J. H. (1997). On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.* 1, 55–77. doi: 10.1023/A:1009778005914

Galy, E. (2018). Consideration of several mental workload categories: perspectives for elaboration of new ergonomic recommendations concerning shiftwork. *Theor. Issues Ergonom. Sci.* 19, 483–497. doi: 10.1080/1463922X.2017.1381777

Gan, Y., Dong, X., Zhang, Y., Zhang, X., Jia, M., Liu, Z., et al. (2020). Workload measurement using physiological and activity measures for validation test: a case study for the main control room of a nuclear power plant. *Int. J. Ind. Ergon.* 78, 102974. doi: 10.1016/j.ergon.2020.102974

Gevins, A., Leong, H., Du, R., Smith, M. E., Le, J., DuRousseau, D., et al. (1995). Towards measurement of brain function in operational environments. *Biol. Psychol.* 40, 169–186. doi: 10.1016/0301-0511(95)05105-8

Gevins, A., and Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theor. Issues Ergonom. Sci.* 4, 113–131. doi: 10.1080/14639220210159717

Gretel.ai (2022). *Gretel.ai – The Developer Stack for Synthetic Data [Online]*. Available online at: https://gretel.ai/ (accessed May 3, 2022).

Hancock, G., Longo, L., Young, M., and Hancock, P. (2021). "Mental workload," in *Handbook of Human Factors and Ergonomics* (New York, NY), 203–226.

Hancock, P., Vercruyssen, M., and Rodenburg, G. (1992). The effect of gender and time-of-day on time perception and mental workload. *Curr. Psychol.* 11, 203–225. doi: 10.1007/BF02686841

Hart, S. G., and Staveland, L. E. (1988). Development of nasa-tlx (task load index): results of empirical and theoretical research. *Adv. Psychol.* 52, 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hernandez-Matamoros, A., Fujita, H., and Perez-Meana, H. (2020). A novel approach to create synthetic biomedical signals using birnn. *Inf. Sci.* 541, 218–241. doi: 10.1016/j.ins.2020.06.019

Hu, J., and Min, J. (2018). Automated detection of driver fatigue based on eeg signals using gradient boosting decision tree model. *Cogn. Neurodyn.* 12, 431–440. doi: 10.1007/s11571-018-9485-1

Hu, X., and Lodewijks, G. (2020). Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures: the value of differentiation of sleepiness and mental fatigue. *J. Safety Res.* 72, 173–187. doi: 10.1016/j.jsr.2019.12.015

Kakkos, I., Dimitrakopoulos, G. N., Gao, L., Zhang, Y., Qi, P., Matsopoulos, G. K., et al. (2019). Mental workload drives different reorganizations of functional cortical connectivity between 2d and 3d simulated flight experiments. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 1704–1713. doi: 10.1109/TNSRE.2019.2930082

Kamzanova, A. T., Kustubayeva, A. M., and Matthews, G. (2014). Use of eeg workload indices for diagnostic monitoring of vigilance decrement. *Hum. Factors* 56, 1136–1149. doi: 10.1177/0018720814526617

Käthner, I., Wriessnegger, S. C., Müller-Putz, G. R., Kübler, A., and Halder, S. (2014). Effects of mental workload and fatigue on the p300, alpha and theta band power during operation of an erp (p300) brain-computer interface. *Biol. Psychol.* 102, 118–129. doi: 10.1016/j.biopsycho.2014.07.014

Li, X., Vaezipour, A., Rakotonirainy, A., Demmel, S., and Oviedo-Trespalacios, O. (2020). Exploring drivers' mental workload and visual demand while using an in-vehicle hmi for eco-safe driving. *Accident Anal. Prevent.* 146, 105756. doi: 10.1016/j.aap.2020.105756

Lieberman, M. G., and Morris, J. D. (2014). The precise effect of multicollinearity on classification prediction. *Multiple Linear Regression Viewpoints* 40, 5–10.

Lim, W., Sourina, O., and Wang, L. (2018). Stew: simultaneous task eeg workload data set. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 2106–2114. doi: 10.1109/TNSRE.2018.2872924

Longo, L. (2012). "Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design," in *International Conference on User Modeling, Adaptation, and Personalization* (Berlin; Heidelberg: Springer), 369–373.

Longo, L. (2018a). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLoS ONE* 13, e0199661. doi: 10.1371/journal.pone.0199661

Longo, L. (2018b). "On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: a case study in a third-level course," in *CSEDU (2)* (Funchal), 166–178.

Longo, L., and Orr, G. (2020). Evaluating instructional designs with mental workload assessments in university classrooms. *Behav. Inf. Technol.* 1–31. doi: 10.1080/0144929X.2020.1864019

Longo, L., and Orru, G. (2018). "An evaluation of the reliability, validity and sensitivity of three human mental workload measures under different instructional conditions in third-level education," in *International Conference on Computer Supported Education* (Funchal: Springer), 384–413.

Longo, L., and Rajendran, M. (2021). "A novel parabolic model of instructional efficiency grounded on ideal mental workload and performance," in *Human Mental Workload: Models and Applications*, eds L. Longo and M. C. Leva (Cham: Springer International Publishing), 11–36.

MacLean, M. H., Arnell, K. M., and Cote, K. A. (2012). Resting eeg in alpha and beta bands predicts individual differences in attentional blink magnitude. *Brain Cogn.* 78, 218–229. doi: 10.1016/j.bandc.2011.12.010

Mapelli, I., and Özkurt, T. E. (2019). Brain oscillatory correlates of visual short-term memory errors. *Front. Hum. Neurosci.* 13, 33. doi: 10.3389/fnhum.2019.00033

Mazher, M., Abd Aziz, A., Malik, A. S., and Amin, H. U. (2017). An eeg-based cognitive load assessment in multimedia learning using feature extraction and partial directed coherence. *IEEE Access.* 5, 14819–14829. doi: 10.1109/ACCESS.2017.2731784

Mesulam, M. (1990). Report of ifcn committee on basic mechanisms of cerebral rhythmic activities. *Electroencephalogr. Clin. Neurophysiol* 76, 481508. doi: 10.1016/0013-4694(90)90001-Z

Mikulka, P. J., Scerbo, M. W., and Freeman, F. G. (2002). Effects of a biocybernetic system on vigilance performance. *Hum. Factors* 44, 654–664. doi: 10.1518/0018720024496944

Miyakoshi, M. (2018). *Makoto's Preprocessing Pipeline*. Available online at: https://sccn. ucsd. edu/wiki/Makotos_preprocessing_pipeline (accessed on February 1, 2019).

Moustafa, K., and Longo, L. (2019). "Analysing the impact of machine learning to model subjective mental workload: a case study in third-level education," in *Human Mental Workload: Models and Applications*, eds L. Longo and M. C. Leva (Cham: Springer International Publishing), 92–111.

Nolan, H., Whelan, R., and Reilly, R. (2010). Faster: Fully automated statistical thresholding for eeg artifact rejection. *J. Neurosci. Methods* 192, 152–162. doi: 10.1016/j.jneumeth.2010.07.015

Orru, G., and Longo, L. (2019). "Direct instruction and its extension with a community of inquiry: a comparison of mental workload, performance and efficiency," in *Proceedings of the 11th International Conference on Computer Supported Education, CSEDU 2019, Heraklion, Crete, Greece, May 2-4, 2019, Volume 1*, 436–444.

Palva, S., Kulashekhar, S., Hämäläinen, M., and Palva, J. M. (2011). Localization of cortical phase and amplitude dynamics during visual working memory encoding and retention. *J. Neurosci.* 31, 5013–5025. doi: 10.1523/JNEUROSCI.5592-10.2011

Powell, A., Bates, D., Van Wyk, C., and de Abreu, D. (2019). "A cross-comparison of feature selection algorithms on multiple cyber security data-sets," in *FAIR* (Cape Town), 196–207.

Puma, S., Matton, N., Paubel, P.-V., Raufaste, É., and El-Yagoubi, R. (2018). Using theta and alpha band power to assess cognitive workload in multitasking environments. *Int. J. Psychophysiol.* 123, 111–120. doi: 10.1016/j.ijpsycho.2017.10.004

Putze, F., Vourvopoulos, A., Lécuyer, A., Krusienski, D., Bermúdez i Badia, S., Mullen, T., et al. (2020). Brain-computer interfaces and augmented/virtual reality. *Front. Hum. Neurosci.* 14, 144. doi: 10.3389/fnhum.2020.00144

Reid, G. B., and Nygren, T. E. (1988). "The subjective workload assessment technique: a scaling procedure for measuring mental workload," in *Advances in Psychology, volume 52 of Human Mental Workload*, eds P.A. Hancock and N. Meshkati (North-Holland), 185–218.

Rizzo, L., and Longo, L. (2018). "Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study," in *Proceedings of the 2nd Workshop on Advances In Argumentation In Artificial Intelligence, co-located with XVII International Conference of the Italian Association for Artificial Intelligence, AI³@AI*IA 2018, 20-23 November 2018, Trento, Italy*, 11–26.

Roy, R. N., Charbonnier, S., Campagne, A., and Bonnet, S. (2016). Efficient mental workload estimation using task-independent EEG features. *J. Neural Eng.* 13, 026019. doi: 10.1088/1741-2560/13/2/026019

Schmidt, M., Kanda, P., Basile, L., da Silva Lopes, H., Baratho, R., Demario, J., et al. (2013). Index of alpha/theta ratio of the electroencephalogram: a new marker for alzheimer's disease. *Front. Aging Neurosci.* 5, 60. doi: 10.3389/fnagi.2013.00060

Smit, A. S., Eling, P. A., Hopman, M. T., and Coenen, A. M. (2005). Mental and physical effort affect vigilance differently. *Int. J. Psychophysiol.* 57, 211–217. doi: 10.1016/j.ijpsycho.2005.02.001

Spitzer, B., and Haegens, S. (2017). Beyond the status quo: a role for beta oscillations in endogenous content (re) activation. *eNeuro* 4, ENEURO.0170-17.2017. doi: 10.1523/ENEURO.0170-17.2017

Tsang, P. S., and Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 39, 358–381. doi: 10.1080/00140139608964470

Tsang, P. S., and Vidulich, M. A. (2006). Mental workload and situation awareness. *Proc. Hum. Factors Ergonom. Soc. Ann. Meeting* 44, 3–463. doi: 10.1002/0470048204.ch9

Vakalopoulos, C. (2014). The eeg as an index of neuromodulator balance in memory and mental illness. *Front. Neurosci.* 8, 63. doi: 10.3389/fnins.2014.00063

Verleysen, M., and François, D. (2005). "The curse of dimensionality in data mining and time series prediction," in *International Work-Conference on Artificial Neural Networks* (Springer), 758–770.

Wang, P., Fang, W., and Guo, B. (2021). Mental workload evaluation and its application in train driving multitasking scheduling: a timed petri net-based model. *Cogn. Technol. Work* 23, 299–313. doi: 10.1007/s10111-019-00608-w

Wang, X., Li, D., Menassa, C. C., and Kamat, V. R. (2019). Investigating the effect of indoor thermal environment on occupants' mental workload and task performance using electroencephalogram. *Build. Environ.* 158, 120–132. doi: 10.1016/j.buildenv.2019.05.012

Wickens, C. D. (2008). Multiple resources and mental workload. *Hum. Factors* 50, 449–455. doi: 10.1518/001872008X288394

Wilson, N., Gorji, H. T., VanBree, J., Hoffman, B., Tavakolian, K., and Petros, T. (2021). "Identifying opportunities for augmented cognition during live flight scenario: an analysis of pilot mental workload using EEG," in *94th International Symposium on Aviation Psychology* (Oregon), 444.

Wróbel, A. (2000). Beta activity: a carrier for visual attention. *Acta Neurobiol. Exp.* 60, 247–260.

Wu, Y., Liu, Z., Jia, M., Tran, C. C., and Yan, S. (2020). Using artificial neural networks for predicting mental workload in nuclear power plants based on eye tracking. *Nucl. Technol.* 206, 94–106. doi: 10.1080/00295450.2019.1620055

Xie, J., Xu, G., Wang, J., Li, M., Han, C., and Jia, Y. (2016). Effects of mental load and fatigue on steady-state evoked potential based brain computer interface tasks: a comparison of periodic flickering and motion-reversal based visual attention. *PLoS ONE* 11, e0163426. doi: 10.1371/journal.pone.0163426

Young, M. S., Brookhuis, K. A., Wickens, C. D., and Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics* 58, 1–17. doi: 10.1080/00140139.2014.956151

Yu, D., Antonik, C. W., Webber, F., Watz, E., and Bennett, W. (2021). Correction to: multi-modal physiological sensing approach for distinguishing high workload events in remotely piloted aircraft simulation. *Hum. Intell. Syst. Integr.* 3, 201–211. doi: 10.1007/s42454-021-00033-3

Zollanvari, A., James, A. P., and Sameni, R. (2019). A theoretical analysis of the peaking phenomenon in classification. *J. Classif.* 37, 421–434. doi: 10.1007/s00357-019-09327-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.