Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review

Yueying Zhou, Shuo Huang, Ziming Xu, Pengpai Wang, Xia Wu, and Daoqiang Zhang

***Abstract*—Machine learning and its subfield deep learning techniques provide opportunities for the development of operator mental state monitoring, especially for cognitive workload recognition using electroencephalogram (EEG) signals. Although a variety of machine learning methods have been proposed for recognizing cognitive workload via EEG recently, there does not yet exist a review that covers in-depth the application of machine learning methods. To alleviate this gap, in this paper, we survey cognitive workload and machine learning literature to identify the approaches and highlight the primary advances. To be specific, we first introduce the concepts of cognitive workload and machine learning. Then, we discuss the steps of classical machine learning for cognitive workload recognition from the following aspects, i.e., EEG data preprocessing, feature extraction and selection, classification method, and evaluation methods. Further, we review the commonly used deep learning methods for this domain. Finally, we expound on the open problem and future outlooks.**

***Index Terms*—Cognitive workload, electroencephalogram (EEG), machine learning, deep learning**

I. Introduction

Currently, with the advanced development of detection devices and machine learning techniques, human mental states monitoring [1,2], such as fatigue [3], emotion [4], and cognitive workload [5], has been widely used in the fields of human-robot interaction and passive brain-computer-interface. As a special case, cognitive workload or mental workload has gained increasing attention. It can be generally described as the ratio of a person's available resource over the task demanded resources [5]. In real environments, overload and underload of work will affect and harm the state of the operator. Therefore, recognizing cognitive workload becomes grossly important.

To date, measurements for cognitive workload can be divided into subjective and objective measures [6]. Subjective measurements are based on the perceived feeling and self-rating of operators, using some questionnaires such as the National Aeronautics and Space Administration-Task Load Index [7] and Subjective Workload Assessment Technique [8] to estimate cognitive workload. Though these methods are easy to implement, they cannot yield real-time, objective, and accurate results. On the other hand, objective measurements are mainly based on the recordings of task performance and physiological signals, which have less interference on the task and can ease the above problems. The commonly used physiological signals can be roughly divided into several categories, i.e., electroencephalogram (EEG), heart rate, eye movement, respiration, electromyogram, and skin [9]. Among them, EEG is one of the most widely used signals, due to its high temporal resolution, convenience, security, and cheapness. Hence, in this paper, we focus on EEG-based cognitive workload recognition.

EEG signals have the characteristics of weak, noisy, and nonstationary among subjects. Considering this, it is still a challenge to find robust features in EEG. The traditional analytical methods are based on the statistical test to validate the difference among features, such as the power change with specific frequency bands [10], which may not provide great modeling power [11], or discover the relationship between cause and effect [12]. In literature, various machine learning methods are proposed to handle those problems [12]. Machine learning can learn discriminative features that well describe the intrinsic rules from data, and build models to predict.

Although several reviews cover cognitive workload assessment using multiple physiological data, to our knowledge, there does not yet exist a review that covers in-depth the application of machine learning methods for recognizing EEG-based cognitive workload. For example, as in [6], twenty-four cognitive workload assessment algorithms using multiple physiological data are reviewed. Several recently published reviews focus on the various physiological measures [9,13] and multi-modal fusion [14] for the cognitive workload. To alleviate the gap, in this paper, we survey cognitive workload and machine learning literature, aiming to highlight the primary advances in the employment of machine learning methods for cognitive workload recognition.

The rest of the review is organized as follows. In Section II, we briefly introduce the concepts of cognitive workload and machine learning. Then, we review the general steps in classical machine learning and the advanced deep learning methods for cognitive workload recognition, in Section III and IV, respectively. In Section V, we discuss the major findings, open problems, and trends that need further investigation. Finally, we conclude the whole paper in Section VI.

This work is supported by the National Key Research and Development Program of China (Nos. 2018YFC2001600, 2018YFC2001602) and the National Natural Science Foundation of China (Nos. 61876082, 61861130366). Corresponding author: D. Zhang (dqzhang@nuaa.edu.cn).

Y. Zhou, S. Huang, Z. Xu, P. Wang and D. Zhang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China (email: zhouyueying@nuaa.edu.cn, huangshuo@nuaa.edu.cn, xuziming@nuaa.edu.cn, pengpaiwang@nuaa.edu.cn, dqzhang@nuaa.edu.cn).

Xia Wu is with School of Artificial Intelligence, Beijing Normal University, Beijing, 100875, China (email: wuxia@bnu.edu.cn).

II. Background

This paper focuses on the recognition of EEG-based cognitive workload using machine learning. In the following, we briefly introduce the definition, applications, and common paradigms of cognitive workload, and machine learning.

A. Cognitive Workload

1) Definition

The cognitive workload is one of the most important mental or functional states of human beings, which can be defined as "the relation between the function relating the mental resources demanded by a task and those resources available to be supplied by the human" [5]. It can be influenced by different factors, such as individual differences, variations of function state, task complexity, social and environmental factors [15].

In real-world environments, the operators in daily work may experience three workload states, i.e., underload, normal, and overload. Keeping the workload in a balanced and fit way can help operators work to safeguard and effectively [12,16].

2) Applications

Cognitive workload recognition is applied to various working environments, such as education [17] (e.g., students learning online [18] and web browsing [19]), public transport (e.g., driving vehicle [20], airplane [21], air traffic management [22]), medical [23], and special working situation (e.g., engineers of nuclear power plants [24]) that needs much attention. In recent years, the cognitive workload is also applied to computer-aided diagnoses, like cancer [25], depression [26], schizophrenia [27], and autism spectrum disorder [28].

3) Common Paradigms

Usually, subjects are required to participate in tasks with various difficulty levels to experience different cognitive workload states. The widely used paradigms to experience cognitive workload are performed either under controlled laboratory conditions, e.g., cognitive task, or operating machine in real or simulated environments. We then categorize the paradigms into cognitive-oriented and operate-oriented.

a) Cognitive-Oriented Task Paradigms. In general, subjects performed the cognitive tasks without too many operations, they are required to keep still. Such paradigms include n-back working memory task [29], Sternberg working memory task [30], mental arithmetic task [31], IQ test [32], silent reading [33], and visual degradation task [34].

Working memory tasks require temporary storage and processing of information items, and it includes the n-back task, Sternberg working memory task, and some variants. In the n-back task, the subject needs to maintain n items and identify whether the current stimulus (like letters, numbers, or graphs) matches a stimulus presented n trials before the current one, here, n is used to set the difficulty level, e.g., 1-back for low level, and 3-back for high level [29]. Two types of n-back tasks, i.e., verbal and spatial n-back are commonly used. Evidence shows that with increasing n, the reaction time is increasing and the accuracy is decreasing. In Sternberg working memory task, the subjects need to remember different numbers of stimulus groups, maintain for seconds, and determine whether the shown stimulus has existed in the memorization groups before. For example, using English letters as the stimulus, the group contains 2,4,6, and 8 letters, corresponding to 4 difficulty levels of the task [30].

In the mental calculations task, the subjects need to maintain the results of the shown formulas (e.g., 5+7) and identify whether the given number (e.g., 11) matches the results they calculated before. This task involves temporal storage of intermediate results and information retrieval held in the mental workspace [31]. Mathematical addition and subtraction are both used in literature. The difficulty levels of addition tasks can be set with multiple digits and carry numbers.

b) Operate-Oriented Task Paradigms. Concerned with the operators involved in various jobs and tasks, operate-oriented task paradigms focus on drivers, air traffic officers, pilots, and surgeons. In such paradigms, participants are required to operate the machine in a simulated or real way. Related paradigms include air traffic management task [35], drive vehicle task [20], the pilot in flight task [21], and complex operation multi-tasks, such as the NASA Multi-Attribute Task Battery (MATB) task [36,37], and automation-enhanced cabin air management system (aCAMS) [38].

The MATB task is a platform designed to evaluate human operator performance and workload in multi-task conditions [36]. It includes four subtasks with light detection or resource management task, monitoring, tracking, and communication task. By varying the demands of each subtask, different difficulty levels of the task can be designed and manipulated.

The air traffic management task is a multi-task activity for air traffic controllers, whereas they are continuously engaged in visual activities of airplane control on the radar and auditory communication with pilots [35]. The workload can be inferred via the changes in traffic manipulation, complexity, or volume.

The aCAMS task provides a micro-world in a spaceflight, where operators perform a safety-critical task with multiple subsystems that control the air quality of a spacious cabin [38]. As long as any automatic control system has a fault, the operator needs to fix the error manually. The difficulty level can be controlled via the number of failed subsystems.

B. Machine Learning

Nowadays, machine learning is one of the most rapidly growing fields, lying at the core of artificial intelligence and data science. It addresses the question of how to build computer systems that learn and improve automatically through experience and data [39]. The general definition of machine learning can be "the problem of improving some measure of performance when executing some task, through some type of training experience" [39].

Generally, based on different tasks, machine learning methods can be categorized into many different types [40], e.g., supervised learning, unsupervised learning, semi-supervised learning, active learning, transfer learning, multi-task learning, and reinforcement learning. Among them, the most common type of machine learning is supervised learning [39].

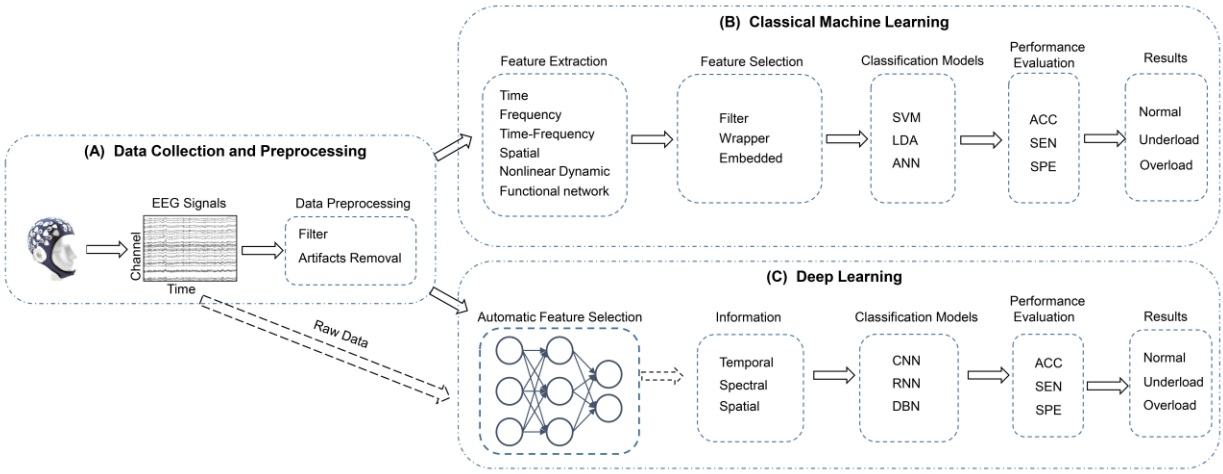

Fig. 1 General steps in machine learning methods with application of EEG-based cognitive workload recognition. (A) is the EEG data collection and preprocessing steps, including filter and artifacts removal steps. (B) is the classical machine learning steps, including feature extraction and selection from various domains, classification methods (e.g., support vector machine (SVM), linear discriminative analysis (LDA), and Artificial Neural Network (ANN)), and performance evaluation metrics (accuracy (ACC), sensitivity (SEN), and specificity (SPE)). (C) is the deep learning methods mainly based on neural networks (e.g., Convolution Neural Network (CNN), Recurrent Neural Network (RNN), and Deep Belief Network (DBN)), where the data processing steps may be unnecessary for them when the model uses the raw data as inputs. In most cases, deep learning methods use computed features to learn temporal, spectral, and spatial information. Taking a three-class classification task as an example, the final result is a prediction of the cognitive workload with normal load, underload, and overload state.

Supervised learning [39] builds a model that can predict outputs from input data, where the labels of all or major data are known. Commonly, the classification and regression methods can be viewed as the supervised learning tasks, e.g., k nearest neighbor, support vector machine, and linear discriminative analysis.

Accumulating studies have provided evidence that machine learning is capable of extracting meaningful information from high-dimensional and noisy EEG data [12,41]. For instance, the Berlin brain-computer-interface group [41] uses a spatial filter and linear discriminative analysis to apply for both brain-computer-interface and arousal monitoring from EEG signals and shows that machine learning can be seen as a key technology to access these fields. In a visual object recognition task [42], using independent component analysis for EEG signal processing and support vector machine for classification, results show that it can correctly distinguish the presence of visual stimuli at around 87% accuracy within single trials.

III. Classical Machine Learning Methods for Cognitive Workload Recognition

Dividing general machine learning into classical one and deep learning firstly, in Fig. 1, we display the major steps of machine learning methods involved in EEG-based cognitive workload recognition.

As shown in Fig. 1, (A) is the data collection and preprocessing step. The combination of (A) and (B) displays the classical machine learning steps including data preprocessing, EEG feature extraction, EEG feature selection methods (i.e., filter, wrapper, and embedded), classification methods (e.g., support vector machine), and performance evaluation (e.g., accuracy). (C) is the deep learning method that can learn features automatically, e.g., Convolution Neural Network, Recurrent Neural Network, and Deep Belief Network. Since deep learning methods can use raw data as inputs, in this way, the data preprocessing steps may be unnecessary for them. In most cases, for this domain, deep learning methods use computed features to learn temporal, spectral, and spatial information.

In the following, we give a brief introduction to the major steps of classical machine learning for workload recognition.

A. Datasets and Data Preprocessing

1) Public Datasets

We summarize some public EEG datasets that can be used for cognitive workload monitoring, as listed in TABLE I.

a) *EEGLearn*[1]. A public dataset was provided by [43,44], where 15 subjects (finally 13, by excluding two subjects due to the noise, aged between 24 and 33 years) performed Sternberg working memory task with English letters, using EEG signals with 64 electrodes and standard 10-10 location. Finally, 2670 trails were acquired and divided into four workload levels of difficulty, corresponding to the set size of memory letters 2, 4, 6, and 8, respectively. The larger the size, the higher the corresponding workload level. Also, the data had been preprocessed well and extracted frequency power as features. This dataset can be used for subject-independent workload classification. Several studies have used this dataset to validate their proposed machine learning models [45,46].

[1] https://github.com/pbashivan/EEGLearn;

TABLE I
THE OPEN DATASETS RELATED TO COGNITIVE WORKLOAD TASK

Dataset	Ref.	Signals	Subjects Number	Age	Males Number	Elicited Task	Workload Levels	Raw Data
a) *EEGLearn*	[43,44]	EEG	13	24 to 33	--	verbal Sternberg working memory task	Four 2,4,6,8	No
b) *EEGMAT*	[47]	EEG	36	18 to 26	9	Arithmetic task (serial subtractions)	Two	No
c) *Hybrid EEG-NIRS*	[49]	EEG and NIRS	26	17 to 33	9	n-back task, discrimination/selection response task, and word generation tasks	Three	Yes
d) *WM-EEG*	[53]	EEG and intracranial EEG	9	18 to 56	4	verbal Sternberg working memory task	Three 4,6,8	Yes
e) *STEW*	[55]	EEG	48	--	48	SIMKAP task	Three	Yes

b) *EEGMAT*[2]. A public EEG dataset was provided by the National Technical University of Ukraine [47], where 66 (finally 36) college students (aged from 18 to 26 years) performed the mental arithmetic task (serial subtractions), with 23 EEG electrodes placed over the scalp at 10-20 scheme. For task design, participants had 3 min of EEG resting state with eyes closed and then performed a mental arithmetic task for 4 min. The dataset has the EEG recording of the resting state with eyes closed and the workload state of the first minute. Based on the task performance (number of subtractions and accuracy of the result), the subjects were divided into good counters and bad counters. This dataset can be used to analyze neuronal dynamics [48] and cognitive characteristics of the brain during workload tasks.

c) *Hybrid EEG-NIRS* [3]. An open dataset included simultaneous EEG and near-infrared reflectance spectroscopy (NIRS) recordings, where 26 participants (aged from 17 to 33 years, 9 males) performed three cognitive tasks, including n-back (n = 0, 2, 3) task, discrimination/selection response task, and word generation tasks [49]. This dataset can be used for signal processing and feature extraction [50], single EEG or NIRS analysis, signal task or cross task analysis, and hybrid brain-computer-interface research [51,52].

d) *WM-EEG* [4]. An EEG dataset was recorded from 9 epilepsy patients (aged from 28 to 56, 4 males) during a verbal Sternberg working memory task (English letters with 4, 6, and 8), to monitoring intracranial for the localization of epileptic seizures [53]. This dataset enables the investigation of working memory by providing simultaneous scalp EEG and intracranial EEG recordings, which can be used for brain connectivity analysis, along with hard-to-obtain neuronal recordings from humans [54].

e) *STEW*[5]. An open-access EEG dataset was collected for multi-task cognitive workload activity from 48 male college students who participated in a single-session simultaneous capacity (SIMKAP) task [55], with 14 EEG electrodes placed over the scalp at 10-20 scheme. The SIMKAP multitasking test requires subjects to cross out identical items by comparing two separate panes, whilst responding to auditory questions which can be arithmetic, comparison, or data lookup in nature. The dataset had the EEG recording of resting-state (3 min with eyes open) and workload state (2.5 min), combined with a subjective rating (rating scale from 1 to 9). As such, the class label is decided by the range of rating scale, into low, moderate, and high for SIMKAP task, and low and high for resting state. The sizable dataset of 48 subjects is hoped to be facilitated the development of novel EEG data intra-subject and inter-subject classification algorithms [56], which can be used for subjective and objective data analysis.

To summarize, we find that the listed workload datasets are performed on healthy young or college students, except for *WM-EEG*. Also, *EEGLearn* is given well-preprocessed data with extracted features, *EEGMAT* is preprocessed, while the rest three are raw data. We note that *EEGLearn* and *WM-EEG* are using the same task, and then the public dataset combination or cross-dataset analysis should take into consideration in the future. Of note, it is invalid to compare the performances among the studies using different datasets, as they may have different tasks and preprocessing procedures. On the other hand, studies that examine identical datasets using different approaches provide a more meaningful comparison. Thus, for model-performance improvement, data sharing should be encouraged.

2) *Data Preprocessing*

Due to the noisy characteristic of EEG signals, numerous artifacts (some other non-EEG signals or noises) can be induced during the data collection process, making EEG signals not accurately represent signals from the brain. Therefore, it is vital to apply preprocessing and denoising methods to the recorded EEG data, for reducing the influence of the artifacts and getting clean data. With the development of EEG processing methods, some effective EEG signal visualization and processing toolboxes have been proposed, such as EEGLAB [57] and MNE-Python [58].

There are several preprocessing steps commonly used in the EEG recordings, e.g., filtering, re-referencing, segmenting the signals into epochs, removing or interpolating bad channels, and artifact removal [59]. The filtering, epoch extraction, and artifact removal are inevitable for preprocessing.

a) Filtering data. Due to the existence of power line noise (i.e., 50 Hz in Europe and Asia, or 60Hz in the USA), high-frequency noise, and very low-frequency noise, appropriate filtering is first recommended to eliminate these noises [59]. Filtering EEG signals with certain frequencies were popular, including band-pass filter and notch filter. For band-pass filters (e.g., 0.5 Hz to 50Hz are widely used [16]),

[2] https://physionet.org/physiobank/database/eegmat/;
[3] http://doc.ml.tu-erlin.de/simultaneous_EEG_NIRS;
[4] https://doi.org/10.12751/g-node.d76994;
[5] http://dx.doi.org/10.21227/44r8-ya50;

signal frequencies between the range are kept, while frequencies outside the range are attenuated. A 50/60 Hz notch filter is used to reduce the power line noise by attenuating the corresponding frequency.

b) Re-referencing is a linear transformation of the EEG signals, through which noise in the reference electrodes could turn into noise in the scalp [59]. We can re-reference the EEG signals to the common average of the electrodes or a specific channel, such as mastoid [16] or central electrode CZ.

c) Downsampling. The sample rates of different types of equipment are ranged from 1000Hz to 128Hz (i.e., how many data points are recorded in one second). Downsampling aims to reduce the amount of data, e.g., from 1000Hz to 256Hz, thus increasing the calculation speed. This step is not necessary.

d) Epoch extraction. After that, we can extract epochs that are specific to the events of interest, i.e., around the stimulus or correct response [16,59], to facilitate the investigation of task/stimulus-related changes in EEG.

e) Removal or interpolation of bad channels. Bad channels that are not accurately providing the information on brain activities for some reason, can either be directly removed or interpolated with good channels near them. To avoid information loss, interpolation is often used.

f) Artifact removal. Artifact removal methods include artifact rejection and artifact correction [60]. For example, for large and transient artifacts (e.g., blinks), detecting them by visual examination and removing them through discarding the contaminated EEG epochs are artifact rejection (i.e., removal of bad epochs); for small and constant artifacts (e.g., electrocardiogram), artifact correction estimates the influence of these artifacts on the EEG and using correction procedures (e.g., independent component analysis) to remove the artifacts. Further, artifact correction using independent component analysis has several steps, first, decomposing EEG data into independent components, then checkup manually and removing the bad components. In the case of EEGLAB, there are many tools for artifact removal, e.g., ADJUST [61] and artifact subspace reconstruction [62]. Currently, reference [63] compared nine automated EEG movement-related artifact removal algorithms for low and high workload classification in the MATB task under two physical activities. Experiment results state the combinations of artifact subspace reconstruction + ADJUST and artifact subspace reconstruction + wavelet enhanced independent component analysis can be useful concerned with their better performance.

B. Feature Extraction

After the data preprocessing, we now have relatively clean data. In what follows, an important problem is how to extract the salient features of various cognitive workload states from EEG data. Currently, we can categorize the cognitive workload-related features into time domain, frequency domain, time-frequency domain, spatial domain, nonlinear dynamics, and functional connectivity network features. Many general EEG features have been summarized in [64].

1) Time Domain

The time domain analysis is intuitive and easy to obtain. It aims to find the change of signal amplitude or other attributes w.r.t time. Time domain features mainly include event-related potentials (ERP) [21], statistics features (e.g., mean, standard deviation, variance, kurtosis, skewness), higher-order crossing analysis [28], and Hjorth parameter. The time domain analysis was first developed and contained most information on EEG so that many researchers are still using them.

Among them, ERP features are EEG peaks averaged in the time domain and time-locked to discrete stimuli [65]. Several early and late EPR components have been found to respond to task difficulty levels [66]. For instance, as reported, the amplitude of the P300 elicited by targets of concurrent tasks decreased with increasing workload, and it is thought to be a reliable biomarker for workload estimation [65, 66]. Also, the amplitudes of the early N100, N200, and P200 components are reduced as workload increases [30, 66].

However, given the complexity of EEG signals, there are no particularly effective time domain features except ERP.

2) Frequency Domain

The frequency domain analysis is proposed to display the frequency information of EEG, with an assumption that EEG signals are stationary. By converting the time domain signal into the frequency domain firstly, we can decompose the frequency band into several sub-bands that are closely related to human neural activities, such as delta (0.1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz) and gamma band (30-80 Hz). Briefly speaking, delta, theta, alpha, beta, and gamma bands are mainly associated with deep sleep, drowsy, relaxed, engaged, conscious, and active states, respectively [67]. Depending on the researcher's definition, the range of each frequency domain can be adjusted slightly. Such decomposing techniques are Fourier transform [68] based methods.

Several studies revealed associations between varying workload levels and power alterations in EEG frequency bands. Particularly, the power changes of the alpha band (decreased) and theta band (increased) may serve as the discriminant indicators for workload estimation with increasing workload [66, 69]. Besides alpha and theta bands, powers in the delta, beta, and gamma bands had been reported to associate with varying workloads [70,71,72]. In a related study, Zarjam et al. [70] examined the working memory workload discrimination capability of different frontal frequency bands and found reduced delta power was significantly associated with the increasing workload. Moreover, in a driving task, Borghini et al. [73] proposed a workload index of difficulty levels based on theta and alpha power spectra. And the study found the increasing workload was correlated with increased theta band power over prefrontal areas and decreased alpha band power in parietal areas.

Further, frequency domain features can be computed from these frequency sub-bands, such as power spectrum, energy, event-related synchronization/desynchronization [74], and power spectral density (PSD). Among them, PSD represents

the distribution of power as a function of frequency and it is the most commonly used feature [21].

3) Time-Frequency Domain

Time-frequency domain analysis aims to study how spectral change over time since EEG signals are nonstationary, which can both has time and frequency information. Usually, it adopts the sliding windows and assumes the signals within the window are stationary, and then uses frequency domain techniques to compute the features. The related methods are, for instance, short-time Fourier transform [75] with a fixed window function, wavelet transform that changes the window size adaptively with wavelet function (including continuous wavelet transform, and discrete wavelet transform [72]), and empirical mode decomposition (a nonlinear method).

Short-time Fourier transform divides signals into small sequential data frames (by shifting windows) and a fast Fourier transform is then applied to each frame [76]. For example, short-time Fourier transform is performed on 128 Hz downsampled data (5 minutes) with 40 s windows and 35 s overlap, and then computed power features of seven frequency bands, resulting in 133 features (of 19 channels) as classifier inputs [75]. Here, the longtime windows are used for the long data sample. In most cases, the window length of the Fourier transform is varied from 0.5 to 10 s for cognitive tasks, which is specific to the length of data segmented.

Discrete wavelet transform decomposes the signal into a coarse approximation and detailed information. It employs scaling functions and wavelet functions, which are related to low-pass and high-pass filters, respectively [77]. In this way, the approximate and detail coefficients are produced as features. The selection of suitable wavelets and the number of scales is very important. The smoothing feature of the Daubechies wavelet of order 4 made it more appropriate to detect changes in EEG [78]. For example, the Daubechies-4 wavelet and the 3 wavelet scales are used for decomposing data, and wavelet-based features are then computed as inputs for the workload classifiers [31].

Essentially, both the frequency and time-frequency domain are transformed the time domain signals into the frequency domain, and the feature types (such as PSD) calculated by the two are the same. Compared with the frequency domain, the time-frequency domain has some information involved with time and is expected to extract more useful information.

4) Spatial Domain

Spatial domain features are taking into account to offer some spatial information, as the electrodes of scalp EEG are placed with different brain regions. Generally, the distributions of EEG electrodes can be divided into frontal, parietal, occipital, temporal cortex, and central area. Previous studies indicate the frontal and parietal areas are sensitive to workload change [73]. As such, some studies directly use the related channels in these areas for workload analysis [18,79].

To extract the spatial features, we can further use the spatial filter methods, such as principal component analysis, common spatial patterns [80], and canonical correlation analysis [30]. The filters learned from the data are used to project the EEG to a low-dimension informative subspace, and then the final features are constructed from those projections [43]. The spatial filter methods are widely used in EEG-based motor imagery classification problems and now have been applied to enhance mental workload estimation [81].

5) Nonlinear Dynamics

It has long been observed that EEG signals are nonlinear and nonstationary [82]. Here, nonlinearity means the brain oscillation is not a linear combination of frequency band components, and nonstationary means the frequency band components change in amplitude or shape as time evolves [83]. As such, nonlinear dynamic features are applied to characterize the irregular and nonlinear characteristics of EEG. The most commonly used nonlinear features for workload analysis are *complexity features* and *entropy features* [84]. The complexity analysis mainly represents the degree of randomness in time series, such as Lempel-Ziv Complexity [79] and its variants. And entropy features reflect signal fluctuations and unpredictability, e.g., Shannon entropy, approximate entropy, and permutation entropy [31]. For instance, Shannon entropy measures the probability density based on the probability distribution of amplitude values. Approximate entropy measures the predictability of future amplitude values of the signal based on the information of previous amplitude values. For more detailed information, the interested readers are referred to [82,83,84].

6) Functional Connectivity Features

Besides the above features, researchers also use functional connectivity networks modeling and graph theory to measure the relationship between different EEG electrodes or brain regions [85]. EEG-related functional connectivity networks can be constructed by some methods [86], such as Pearson correlation coefficient [16], phase-locking value [85] [56], phase-locking index [87], and partial directed coherence. For example, using two mental tasks (N-back and mental arithmetic), Dimitrakopoulos et al. [16] proposed to classify binary workload levels based on multi-band connectivity features computed by Pearson correlation coefficient, and the final discriminative features were mostly located in frontal areas in theta and beta frequency bands.

In brief, there are various features extracted from EEG signals with multiple perspectives, such as time (temporal) features, frequency (spectra) features, and nonlinear features. Among these features, previous studies [21,34,85] indicate that PSD in the frequency domain, ERP in the time domain, entropy-based features, and functional connectivity can be used as effective measures for cognitive workload recognition. Researchers can make their decisions to select appropriate EEG features on hypothesis or data-driven concerns.

C. Feature Selection

After the feature extraction step, a feature selection/ transformation step is generally applied to select or construct a

subset of discriminative features from high-dimensional EEG data, especially when the feature number is larger than the sample number. This step has various benefits, for instance, it can reduce the training time of the learning model, remove the irrelevant and redundant features, thus improving the prediction and population generalization performance (i.e., the ability to identify the classes of unseen data).

Feature selection methods are supervised and generally do not change the feature structure, just reduce the number of irrelevant features, whereas feature transformation methods are unsupervised and may change and reorganize the feature structure to maintain the more important information. Further, we can divide supervised feature selection techniques into the filter, wrapper, and embedded techniques [88,89], whereas feature transformation methods are mostly unsupervised techniques, e.g., principal component analysis, and locality preserving projection. Since feature transformation can be seen as special spatial filter methods, we then focus on feature selection methods related to workload estimation.

1) Filter Techniques

Filter techniques implement feature selections before classification or regression and mostly have two steps [89]. First, using simple statistical measures (e.g. mean, variance, correlation coefficients, or relevance) in detecting group-level differences to rank features. Second, selecting the features with the highest-ranking scores. As such, the selected features are then sent into classifiers. There are many classic filter methods, including t-tests, analysis of variance (ANOVA) [90], Relief [91], minimal-Redundancy-Maximal-Relevance (mRMR) [92], Fisher score, and mutual information.

Statistical hypothesis testing techniques have been used to detect average group-level differences. For example, the t-test can detect the difference between the two groups, whereas ANOVA can select features among multiple groups. These methods are computationally fast and easy to implement. In [90], ANOVA and t-test are both used to assess the efficacy of the ERPs for delineating cognitive workload. However, these techniques consider the information by each feature individually, which may lead to a redundant feature set. The following two methods are more complex and consider multi-feature relationships.

The relief method can determine the importance of features according to their ability to distinguish neighborhood samples, the more important feature should make samples within class closer. In an arithmetical task with three levels [91], using EEG spectral features, the study adopts several ranking methods (eigenvector centrality, ReliefF, mutual information, and Fisher score) as feature selection methods to reduce the input dimension number (95 to 52).

Another popular filter method is mRMR [92], aiming to find the features that satisfy maximum individual relevance to the target class but minimal pairwise redundancy with other features, where the relevance can be measured by mutual information. As displayed in [93], the study extracts various features of time and frequency domains and compares the classification performance with and without channel selection and feature selection steps. Further, with mRMR, the number of features reduced from 54 to 15, and the n-back task with 5 levels confirmed the 4% improvement of accuracy.

2) Wrapper Techniques

However, filter methods may have suboptimal feature subsets due to the features are ranked by their statistical characteristics and not relied on classification or regression models. As such, wrapper techniques are proposed to ease the problem. Wrapper methods first generate different feature subsets and then evaluate them by maximizing or minimizing the objective function from a machine learning model [89]. Related methods include recursive feature elimination (RFE) [94], and forward feature selection [95].

RFE algorithm is worked by eliminating one feature in the feature set and obtaining the performance from the model trained and tested with the remaining features. The eliminated feature corresponding to the best performance is viewed as the smallest contribution one in the set and is removed. An RFE-like feature analysis combined with an artificial neural network has been successfully used for ranking features in a within-task cognitive workload study [96]. Also, in a cross-task cognitive workload study [94], RFE helps to rank and find robust features in the feature set of frequency band features. Comparing with all features and feature selection by RFE, this study validates that feature selection can indeed improve model generalization across tasks.

Forward feature selection is worked by adding features to an empty feature set in a certain order. If the performance is improved, the new features will be retained, and the procedure stopped when no addition of a single feature would yield an improvement in the performance criterion. In a task-independent cognitive workload binary classification task [16], feature scoring stage and sequential forward selection are used to select discriminative functional connectivity features.

Although wrapper methods may have better performance, the training time and computational cost of wrapper methods are relatively high than filter methods, given the evaluation strategy of selection features.

3) Embedded Techniques

Embedded techniques can be viewed as an alternative method for easing filter problems. Embedded techniques select a small subset of relevant features as part of the machine learning process by enforcing certain penalties on the machine learning model [89]. Such methods can reduce the training time and unearth the relationships between the workload labels and features, such as the least absolute shrinkage and selection operator [97] and Elastic net.

The least absolute shrinkage and selection operator adopts an L1 norm penalty to make the feature set sparse. Focusing on the workload of a programmer trying to comprehend code, a study [97] uses the least absolute shrinkage and selection operator to select and decrease the number of possible features of signal power and functional connectivity.

In summary, filter methods are implemented relatively simply and fast using statistical characteristics of features but have no consideration with the learning model. Though wrapper methods rely on the learning model to select features, the training time and computational cost might be higher than the other two types and may cause overfitting problems (i.e., learning properties inherent to the sample rather than relationships in the overall population). Embedded methods combine feature selection and classification, which reduces the training time, but the penalties methods have more parameters than the other two. Moreover, we find the widely used feature selection methods for workload recognition are filter and wrapper methods. Besides the above traditional feature selection methods, new methods such as orthogonal regression (which combined with filter and wrapper) [89] are proposed for EEG-based classification problems. Note that, the specific feature and feature selection methods should be chosen with the concern of the data.

D. Classification Methods

The general steps of constructing the workload recognition model include data collection, preprocessing, feature extraction and selection, classification (or regression) model building, and performance evaluation. After the first three steps, it needs to design an effective workload classification/ regression model. The output of a classifier is a discrete value, whereas the output of a regressor is a continuous variable. Since the majority of research work is handling classification problems, we thus mostly focus on classifiers design.

1) General Classification Methods

Based on the sample set and labels (the corresponding cognitive workload levels), the machine learning classification methods aim to find a rule that can assign samples to one of the levels. Classic classifiers for workload application including support vector machine (SVM), linear discriminative analysis (LDA), linear regression, and artificial neural network (ANN). As follows, we brief introduce SVM, ANN, and LDA.

SVM constructs a hyperplane in a high-dimensional space so that an optimal decision boundary (which maximizes the margin between the boundaries of different classes) can be achieved in this hyperplane [98]. The main novelty of SVM is the use of kernel function for the nonlinear problem, which maps the original nonlinearly-separable feature space into a higher dimensional linearly-separable space. The boundary can be linear or nonlinear, depending on the kernel function. Here, kernel function can be linear, polynomial, or radial basis function\Gaussian. Indeed, the linear and radial basis function kernels are commonly used for workload analysis [16].

ANN is intended to mimic the computation within the human brain and consisted of at least three layers, i.e., input, hidden, and output layers. The input layer contains neurons of extracted features, while the output layer contains several neurons for the corresponding class (e.g., two for low and high levels). The hidden layer contains a nonlinear learning algorithm to tune a set of adaptive weights to determine the contribution of each feature for a class [99].

LDA attempts to project the input features onto a smaller feature space by finding a projection direction to a line [100]. It minimizes the in-class distance and maximizes the between-class distance, thus making all samples from different classes are well separated.

2) Performance Evaluation

To evaluate the performance of recognition models, evaluation types, strategies, and metrics are indispensable.

For cognitive workload recognition, evaluation types can be roughly divided into subject-dependent, subject-independent (or cross-subject), and task-independent (or cross-task). Suppose we have N subjects, in a subject-dependent study, the model is trained and tested on data specific to each subject. Given the large variability between subjects, this is the common and standard way to design the brain-computer interface [81]. The cross-subject study is trained and tested with data from a group of subjects. It may include two cases, case 1) the data is combined from all N subjects; case 2) the model is trained on N-1 subjects and tested on N*th* subject. The difference lies in the training step whether using the data of the target subject or not. In the task-independent study, the model is trained on one task and tested on another similar task. The latter two are more complex and challenging than the first one, but if successful, would enable mental state monitoring in real-world environments [81].

For evaluation strategy, generally, all the data is partitioned into training and testing data via a hold-out or k-fold cross-validation. The hold-out procedure divides the data into the training set (to train the model) and the hold-out set (to test the model), with a certain ratio, such as 80% to 20%. While the k-fold procedure divides the data into k similar-sized and manually-exclusive subsets. And in the *k*th fold, k-1 subsets are used for training the model, while the rest one testing. This step performs k times and the final result is the average of the k fold results. A special case of k-fold cross-validation is leave-one-out cross-validation, in each fold, one sample is left out, and the others are used for training.

Further, metrics for evaluating the performance of the trained classification model, such as accuracy, sensitivity, and specificity, are commonly used. Taking a binary classification for example, we can set the low-level samples as the positive class, and the high-level samples as negative. In this way, accuracy can measure the effectiveness of predicting the true class label, including both positive and negative classes. Sensitivity measures the effectiveness of the positive class label while specificity measures the negative.

Following, we would like to discuss the influence of the number of workload levels, subjects, electrodes used, and data evaluation strategy, on the performance accuracy, and the contribution of an algorithm to the field.

The classification levels of the cognitive workload are ranged from two to seven. In most cases, two to four levels are used. We argue the accuracy of 80% for two workload levels (e.g., easy versus hard) is not as impactful as the same accuracy for three workload levels (e.g., easy, moderate, and hard) since the latter are more complex and hard to classify.

The number of the subject involved in workload tasks is mostly ranged from 6 to 30. With more subjects involved, the number of experimental data getting larger, and the robustness of the proposed models are verified with stronger statistical power. In this case, the accuracy generated from more subjects' data may increase the reliability of the performance.

The number of electrodes in EEG-based workload tasks is ranged from 2 to 128, e.g., the 64 electrodes used in [16]. The high-density EEG equipment has the advantage to capture efficient information over fewer electrodes, but it is time-consuming to build machine learning models, due to the larger number of parameters. The number of electrodes used impacts the sensitivity of a model to overfitting. Achieving a similar accuracy with fewer electrodes is a simpler model, which boosts real-time workload recognition.

For data evaluation strategy, the k-fold cross-validation (k = 5 or 10) is mostly used for subject-dependent studies, and leave-one-subject-out cross-validation is commonly used for subject-independent studies. We argue that the k-fold cross-validation may get higher classification performance than leave-one-subject-out cross-validation since the latter is stricter. However, some papers do not state clearly the data split strategy, thus limiting the power of the studies.

Given different features extracted from EEG signals, an algorithm can help us to find a relatively faster decision boundary to recognize different cognitive workload states. The main difference for varied algorithms is lying in their ability to construct a more accurate and robust boundary.

3) Applications with Cognitive Workload

As displayed in Table II part 1, we review some papers with task parameters, feature extraction and selection, classification models with validation strategy, and evaluation metrics. Here, we first present the cognitive-oriented task. The preprocessing steps, detailed features, and evaluation types are included in Supplementary Table S1. We mainly make this table based on SVM, ANN, LDA, and others.

a) Support Vector Machine (SVM)

We find some papers extract various features and adopt SVM as classifiers, as cases in [12,16,18,20,34,70,74,101,102], indicating the extensiveness and compatibility of SVM methods. For example, Dimitrakopoulos et al. [16] use SVM to classify low and high levels of n-back and arithmetic tasks for 25 subjects, using functional connectivity features and sequential forward selection. The proposed model was able to find the few features of task-independent (17) and task-specific (21 to 23) and provided accuracy of 87% for cross-task, 88% for N-back task, and 86% for arithmetic task. The SVM classifiers with different kernels were employed, and the best performance was obtained with radial basis function kernel. However, the computational time of this model might be high due to the wrapper feature selection. Also, when comparing with other new models, SVM, and some widely used methods (e.g., LDA) can be seen as baseline methods for EEG classification tasks [103].

Besides traditional SVM methods, some variants are also proposed to deal with more complex problems, such as proximal, bound, ensemble, and least square SVMs. To analyze the binary classification of the n-back task (n=0,1,2,3) for nine subjects using 14 channels, a balanced proximal SVM algorithm [29] is employed as a fast and robust alternative to the standard SVM, to alleviate the long time training and data sample unbalanced problem (e.g., 0-back vs 1-,2-,3-back). Combined with mRMR, the accuracies of two-class classification of the lowest level (0-back) and the higher levels (1-, 2-, 3-back) were close to 100%, and the accuracies for 1-back vs 2-back, 1-back vs 3-back were 80%, and 84%, respectively. Some studies of aCAMs task use bound SVM [104], ensemble SVM [105], and least square SVM [106] to classify cognitive workload of multi-levels, with improvement than standard SVM also with the concern of training time. For example, least-square SVM improves the training speed while preserves high generalization capability [106]. Bounded SVM can be trained on a single objective function and has comparable accuracy with a standard SVM. A bounded SVM is used to evaluate the three-class cognitive workload classification, to avoid the complicated one-vs-all SVM classification design [104]. Ensemble SVM classifier is composed of three different multi-kernel SVMs, and this design can effectively mitigate the overfitting problem [105]. Also in [94], considering the workload mismatching problem, the study uses an SVM regression method to construct a cognitive workload estimator, which is trained on the n-back task and can work well under the complex MATB task.

We critique the reviewed paper mainly from the number of subjects, validation strategy, and final performance. For cognitive-oriented tasks, as we can see from Table II part 1, the papers that used SVM methods have similar accuracies but different validation strategies and the number of subjects. Since [18] and [74] have a greater number of subjects and a more robust validation strategy, we argue these approaches are more robust, even though they are elicited by different cognitive task environments. For operate-oriented tasks, the involved subject has a less number than the cognitive-oriented task. Similarly, we argue [105] has a more robust performance.

b) Artificial Neural Network (ANN)

Gevins et al. [107] used frequency band features to train a two-hidden-layered ANN to discriminate three levels of difficulties of two (verbal and spatial) working memory tasks, using 27 channels of EEG from 8 subjects. The average accuracies of four conditions were subject-dependent from 80% to 98%, cross-subject 83%, cross-task 94%, cross-session 95%. Also, ANNs are used to classify workload among highly trained operators on a simulated air traffic management task [108] and a multi-task combination contained in the MATB task [96]. The same authors go on using an ANN-based classification to drive adaptive aiding in an unmanned aerial vehicle task, to distinguish two workload levels [68]. These three studies were focused on subject-dependent analysis and used frequency power as features.

TABLE II
CLASSICAL MACHINE LEARNING MODELS FOR COGNITIVE WORKLOAD RECOGNITION

Part 1

Ref.	Task	#L	#S	Feature domain	FS	Data Spilt	Classifier Model	ACC
SVM								
[70]	Silent Reading	3	5	1,3,5	t-test	Hold-out	SVM	0.83
[101]	Silent Reading	3	5	2,5	KW test	Hold-out	SVM	0.82
[74]	Working Memory and Complex Tasks	3	21	2	--	10-fold	SVM	0.82
[20]	Working Memory, Delayed Match-To-Sample Tasks	3	20	5	Feature fusion	LOSOCV	SVM	--
[34]	Visual Degradation	4	16	2,4	--	4-fold	SVM	0.8
[102]	Cognitive and Motor Tasks	3	20	2	--	10-fold	SVM	0.75
[18]	Multimedia Learning	4	34	5,6	DWT	--	SVM	0.88
[43]	Sternberg Task	4	13	2,5	RF	10-fold	SVM	0.92
[16]	N-Back, and Arithmetic	2	28	6	SFS	LOSOCV	SVM	0.87
[29]	N-Back	2	9	1,2,3	mRMR	10-fold	proximal SVM	0.84
[106]	aCAMS	3	10	2	RFE	Hold-out	least square SVM	0.74
[104]	aCAMS	3	6	2	AES, LPP	10-fold	bounded SVM	0.93
[105]	aCAMS	5	8	2	Laplacian Eigenmaps	2-fold	ensemble SVM	0.77
[94]	N-back and MATB	R	17	2	RFE	3-fold	SVM regression	--
ANN								
[107]	Working Memory	3	8	2	visual inspection	Hold-out	ANN	0.8
[72]	Arithmetic	7	12	3	--	LOSOCV	ANN	0.98
[31]	Arithmetic	7	12	3	KW test	LOSOCV	ANN	0.98
[109]	Working Memory	2	15	2	total power	Hold-out	ANN	0.85
[110]	N-Back	3	12	2	power	--	ANN	0.81
[108]	ATM	4	7	2	SFR	Hold-out	ANN	0.88
[96]	MATB	3	7	2	SFR	Hold-out	ANN	0.86
[68]	Simulated UAV	2	12	2	--	Hold-out	ANN	0.75
LDA								
[112]	N-Back	3	15	2	FFT	4-fold	Regularized LDA	0.63
[30]	Sternberg Task	2	20	1,2	--	10-fold	Fisher LDA	0.91
[21]	Flying Task	2	22	1,3	mRMR	5-fold	Shrinkage LDA	0.7
[87]	Simulated Flight	3	33	6	RFE	10-fold	Ensemble-LDA	0.82
[111]	Real Drive	2	17	2	--	11-fold	LDA	0.92
[22]	ATM	3	12	2	--	Hold-out	SWLDA	--
[113]	ATM	3	37	2	FFT	Hold-out	asSWLDA	--
Others								
[115]	N-Back	2	6	4	MI	10-fold	naive Bayes	0.84
[85]	Arithmetic Task	5	16	6	TSA+LPP	Hold-out	KNN	0.75
[75]	MATB	3	8	3	FT	5-fold	hierarchical Bayes	0.8
[114]	aCAMS	2	8	1,2,5	--	Hold-out	ensemble ELM	0.93

Part 2 Comparison Methods

Ref.	Task	#L	#S	Feature domain	FS	Data Spilt	Classifier Models	ACC
[93]	N-Back	5	10	1,2,3	mRMR	Hold-out	KNN, **RF**, DT, SVM, MLP	0.846
[116]	N-Back	2	22	4	MI	Hold-out	CSP+LDA, FBCSP+LDA, RG, **CNN**	0.727
[32]	Raven Test	R	47	2,5,6	--	Hold-out	LR, RF, **XGB**, ANN	0.713 (PCC)
[91]	Arithmetical	3	11	2	FFT, rank	5-fold	SVM, **KNN**, DT, RF	0.91
[117]	N-Back	2	17	2,3,4	ANOVA	LOSOCV	SVM, KNN, **RF**,GBM	0.778
[12]	Real ATM	3	35	2	FFT	5-fold	LR, LDA, SVM, **KNN**,RF	0.84
[95]	Simulated Flight	3	21	1,2,3	PCA, FFS, mRMR	10-fold	KNN, **LDA**, naïve Bayes, DT, SVM	0.901

Note: #L, the number of workload levels, R means regression; #S, the number of subjects; For tasks, aCAMS, automation-enhanced cabin air management system; MATB, multi-attribute task buttery; ATM, air traffic management; UAV, unmanned aerial vehicle. For feature domain, 1-time, 2-frequency, 3-time-frequency, 4-spatial, 5- nonlinear dynamics, 6- functional connectivity networks; KW test, Kruskal-Wallis test; DFA, discriminant function analysis; CSP, common spatial patterns; FBCSP, filter bank CSP; CCA, canonical correlation analysis; WT, wavelet transform; DWT, discrete wavelet transform; FFT, fast Fourier transform; For feature selection, RF, random forest; PCC, Pearson correlation coefficient; SFS, Sequential Forward Selection; mRMR, minimal-Redundancy-Maximal-Relevance; RFE, recursive feature elimination; AES, adaptive exponential smoothing; LPP, locality preserving projection; MI, mutual information; TSA, tensor subspace analysis; FFS, forward feature selection; SFR, saliency feature reduction; For classification methods, SVM, support vector machine; ANN, Artificial Neural Network; LDA, linear discriminative analysis; KNN, k nearest neighbor; DFA, discriminant function analysis; ELM, extreme learning machine; LR, logistic regression; RG, Riemannian Geometry; XGB, Extreme Gradient Boosting; GBM, gradient boosting machine; CNN, Convolution Neural Network. LOSOCV, leave-one-subject out cross-validation, and n-fold is n-fold cross-validation; In part 2, the best classifier models are bold. The detailed table is in supplementary S1, which includes the preprocessing, the specific features, and the evaluation type.

In [31,72], a two-hidden-layered ANN is used to classify seven workload levels of mental arithmetic addition task for 12 subjects, with a first hidden layer of twenty neurons, a second hidden layer of fourteen neurons, and an output layer of seven neurons. To our best knowledge, seven workload levels are the biggest for the cognitive workload. The two studies were cross-subject works using leave-one-subject out cross-validation, focusing on the frontal areas with single-feature classifications, i.e., the performance of each coefficient, were 97%, 98%, 98%, respectively. As we can see, the two studies have similar performances though with different features as inputs. However, in most cases, we are unavailable to reach such high performance.

For two task-independent studies [109,110], though the ANN classifiers achieved similar over 80% (81% to 85%) accuracies within the task, the cross-task accuracies were around the chance level with 50%. Generally, the performance of cross-task and cross-subject studies are worse than subject-specific, due to the high variability between subjects. In short, among the ANN-based approaches, [31] and [72] are more robust due to the use of leave-one-subject out cross-validation and get higher classification accuracy.

c) Linear Discriminative Analysis (LDA)

In a real-time and real driving task [111], a linear LDA is used to classify low and high workload levels, since nonlinear models have inferior offline performance. This study demonstrates the improved performance of a subject by mitigating the workload in high workload conditions.

Besides this, some other variants [21,22,30,87,112,113], such as fisher LDA, shrinkage LDA, and stepwise LDA, have been used to classify the workload under different conditions.

Roy et al. [30] propose to compare two spatial filter processing chains (ERP with canonical correlation analysis versus PSD with common spatial patterns) combined with a fisher LDA classifier for working memory tasks from 20 participants. The binary classification results show that the ERP feature chain has a better and more stable accuracy performance than the PSD feature, as 91% versus 61%. However, the proposed model does not compare with related methods or use ERP or PSD directly as inputs for the classifiers, thus limiting the impact of the paper. Further, the same team goes on research on a real flight scenario to test the feasibility of the dry electrode system, with ERP and PSD features combined with shrinkage LDA [21]. Here, shrinkage consists of using a regularized estimate of the covariance matrices in LDA, which might be helpful to suppress the calibration time in brain-computer-interfaces. Along with the two flying tasks at low and high levels, 22 pilots are presented with a passive auditory oddball. The experimental result shows that the PSD feature performs better than ERP, with 70% accuracy versus 50%, suggesting ERP is much more sensitive to noise than PSD features. Therefore, different task environments might make a great variety of workload analysis.

In a simulated flight experiment with three mental workload levels [87], the authors used functional connectivity network and compared the reorganization pattern between the computer screen and virtual reality interfaces. The multi-level workload classification accuracies are 82%for both interfaces. In an air traffic management task with 3 workload levels, extracting PSD features, stepwise LDA [22], and automatic stop stepwise LDA [113] are adopted as classifiers since these methods have the advantage of automatic feature selection. To be specific, a three-classes stepwise LDA [22] is used to select the most relevant EEG spectral features, to discriminate the cognitive workload level among the three task conditions (easy, moderate, and hard) of 12 air traffic controllers. And then the linear discriminant function is evaluated to test the reliability of the feature selection criteria. These methods can be seen as the wrapper methods. Further, an EEG-based workload index is computed via the moving average. To summary, [87] proposed a more robust LDA-based approach, due to the subject number, cross-validation, and higher accuracy. Also, comparing different feature selection methods, we argue the wrapper techniques may have a more powerful ability to extract features, such as RFE.

Besides the methods mentioned above, some other classifiers are also used in cognitive workload recognition, such as KNN [85] and the Bayes-based model [75,114,115].

In conclusion, SVM with kernel methods are simple and robust, which can handle nonlinear data, but the modeling power may be reduced when the data scale is big. LDA methods are computed fast with low cost and fewer parameters but have inferior performance when the data is little and nonlinear. Standard SVM and LDA are only supporting binary classification problems, and for multiple classifications, these methods need further setup, e.g., the combination of multiple classifiers. ANN methods are shallow neural networks and can naturally deal with multiple classifications. Recently, as we can see, some variants of these methods are also proposed to improve model performance.

d) Comparative Analysis

Many classifiers have their merits and demerits. It is a great challenge to choose or construct the appropriate models. We list several studies [12,32,91,93,95,116,117] that focus on comparison among feature extraction and classification methods in Table II part 2, to find the best model for specific data and tasks. The table has similar contents in Table II part 1 but includes the best-performance-classifier (which are bold) in the corresponding studies. As we can see, the classifiers are not relatively consistent and stable, due to the data-driven features and methods. For example, in [93], to classify five workload levels of n-back tasks, the study uses channel selection and mRMR to select features from time and frequency domains, combined with several supervised machine learning classification methods, e.g., KNN, random forest, decision trees, SVM, ANN, and LDA. The results confirm the improved accuracy of using channel selection and mRMR, as the final mean accuracies are obtained with random forest of 92.9%, 87%, 87%, and 84.6% for two, three, four, and five classes, respectively, while for the case without these

two steps, the final mean accuracies are 92.3%, 85%, 82%, 80% for two, three, four and five classes, respectively. In [91], to classify three workload levels of an arithmetical task, the study uses a fast Fourier transform to calculate the spectrum power of 5 frequency bands and 4 ranking methods to select features, combined with SVM, KNN, decision trees, and random forest. Results show that KNN and decision trees achieve higher accuracies compared to the other models. Take the number of subjects and validation strategy into consideration, we argue the KNN method in [12] performed better in the comparative analysis.

IV. Deep Learning Methods for Cognitive Workload Recognition

In contrast to the conventional cognitive workload recognition framework, which typically extracts features from the temporal and spectral views separately, deep learning can learn to acquire complex information of multi-domain simultaneously [45,103]. Therefore, researchers have begun using deep learning to learn robust EEG representations.

A. Deep Learning Methods

Deep learning is a particular subcategory of machine learning, which allows computational modules to learn hierarchical representations of input data via successive simple but non-linear transformations [118]. Given bigger training data and greater computer processing power in recent years, deep learning significantly outperforms traditional machine learning algorithms [118], ranging from fields of computer vision, speech recognition, natural language processing, to medical image analysis. Deep learning methods, or deep neural networks, in this review, are defined as neural networks with more than two hidden layers, combined with input and output layers, where each layer contains multiple neurons.

Recently, deep learning methods have become prevalent for the analysis of physiological signals for human functional state monitoring and are widely used in brain-computer-interface systems. For instance, some studies use Convolution Neural Network (CNN) for EEG-based human state assessment, such as emotion recognition [119] and driver fatigue detection [3]. Other deep learning models like, and Recurrent Neural Network (RNN) [120] and Deep Belief Network (DBN) [121] are also applied to these domains. As pointed in [3], the deep learning models should reduce the dimension of EEG signals firstly and then transform them into new representations while not lose any significant information. A study [122] reviews some deep learning models (e.g., RNN, and CNN) and their applications for analyzing EEG data to decode human brain activities and diagnose brain diseases. Also, in [123], the researchers review serval EEG classification task applications using deep learning models, including emotion recognition, motor imagery, mental workload, seizure detection, ERP detection, and sleep scoring.

B. Deep Models for Cognitive Workload Recognition

Different types of hidden layers are used as building blocks in deep neural networks, such as convolutional, or recurrent layers, many of which compute non-linear input-output mappings. As such, the neural networks using convolutional or recurrent layers are referred to as CNN and RNN, respectively [122]. Here, we provide a quick overview of CNN and other deep learning architectures used in workload recognition.

1) Convolution Neural Network (CNN)

CNN is one of the most widely used deep learning methods, even in the cognitive workload recognition domain. Typically, the multiple hidden layers of CNN consist of convolutional, pooling, and fully connected layers. The convolutional layers include several 2D filters to convolve with the input image [124], encouraging the model to learn invariant representations of the data. Downsampling is performed in the pooling layers to merge semantically similar features into one, mainly including max pooling and mean pooling. Neurons in the fully connected layer receive the activations of every single neuron of the preceding layer as input. The network parameters of CNN contain the weights of neurons and convolutional kernels, which are tuned by the backpropagation algorithm. A softmax layer is widely used as a classifier and output layer. All these structures enable CNN to extract multi-level hierarchical features.

In [20], a CNN (8 convolutional and 2 fully connected layers) is used for classifying binary levels of cognitive workload and the context of driving of a vehicle driver performing simulated drive task. The proposed method uses only raw EEG signals as input, which helps reduce the noise effect and the complexity of model training.

We find some papers [45,46,124,125] using EEG data from the dataset *EEGLearn* of working memory task with 4 levels, to validate some deep models and compare with the results in [44]. Specifically, Bashivan et al. [44] used fast Fourier transform to convert EEG data into the frequency domain and map the 3D spatial positions of electrodes to 2D, according to the distribution of the electrodes. The 4–30 Hz frequency domain is divided into theta, alpha, and beta frequency bands, generating 3-channel spectral maps which are sent to deep CNN model (7 convolutional and 1 Long Short-Term Memory (LSTM) layers). This method achieves an accuracy of 91.11%. Jiao et al. [45] proposed deep CNNs with 4 convolutional and 2 fully connected layers on classifying four workload levels. Considering the spectral and temporal information of signals, a fused model of CNNs and a high-order extension of the Restricted Boltzmann Machine are used.

The results showed that the fused CNN model can have competitive performance compared with original deep methods in [44], with 92.37% vs 91.11%. Further, Zhang et al. [46] proposed a 6-channel parallel mechanism of spectral feature-enhanced maps to enhances the expression of structural information, compared with the original 3-channel spectral maps in [44]. The study adopted four CNN structures, AlexNet, VGGNet, ResNet, DenseNet, to measure the effectiveness of the approach. The best accuracy is obtained by ResNet with 93.71%.

TABLE III
DEEP LEARNING MODELS FOR COGNITIVE WORKLOAD RECOGNITION

Ref.	Task	#L	#S	Inputs	Data Spilt	Deep Model	ACC
CNN							
[45] *	Sternberg Task	4	13	power map	LOSOCV	CNN	0.9237
[46] *	Sternberg Task	4	13	spectral maps	LOSOCV	parallel mechanism CNN	0.9371
[125] *	Sternberg Task	4	13	spectral, spatial	LOSOCV	3DCNN	0.939
[125]	Sternberg Task	2	62	spectral, spatial	LOSOCV	3DCNN	0.908
[20]	Simulated Drive	2	1	raw data	Hold-out	CNN	0.9531
RNN							
[126]	Working Memory	4	22	spatial, spectral, temporal	LOSOCV	RNN	0.925
[127]	MATB	3	6	spectral-statistical	5-fold	LSTM	0.93
DBN							
[128]	aCAMS	7	6	temporal	Hold-out	DBN	0.928
[129]	aCAMS	3	8	temporal, spectral	LOSOCV	switching DBN	0.769
DAE							
[133]	aCAMS	2	8	frequency-PSD	Hold-out	SDAE	0.74
[134]	aCAMS	2	7	frequency-PSD	Hold-out	adaptive SDAE	0.8579
[135]	aCAMS	2	8	temporal, spectral	10-fold	ensemble SDAE	0.92
[38]	aCAMS	2	8	temporal, spectral	Hold-out	transfer DAE	0.86
Hybrid							
[44] *	Sternberg Task	4	13	spatial, spectral, temporal	LOSOCV	CNN+RNN	0.911
[124] *	Sternberg Task	4	13	spatial, spectral, temporal	10-fold	CNN+BNTM	0.963
[136]	N-Back Task	3	17	spectral, temporal maps	10-fold	CNN+TCN	0.919
[103]	N-Back, Arithmetic	2	20	spatial, spectral, temporal	Hold-out	RNN+3DCNN	0.889
[56]	SIMKAP	2,3	48	temporal, spectral	--	BLSTM-LSTM	0.86, 0.83
[137]	MATB	2	8	spectral-PSD	LOSOCV	CNN+RNN	0.868

Note: #L, the number of workload levels; #S, the number of subjects; For data split strategy, LOSOCV, leave-one-subject out cross-validation, and n-fold is n-fold cross-validation; DAE, Denoising AutoEncoder; SDAE, Stacked Denoising AutoEncoder; DBN, Deep Belief Networks; CNN, Convolution Neural Network; RNN, Recurrent Neural Network; LSTM, long short-term memory; BNTM, Bidirectional Neural Turing Machine; TCN, Temporal Convolution Network; BLSTM, Bidirectional LSTM; * denotes the used EEG data is from dataset *EEGLearn.* The detailed table is in Supplementary Table S2, which includes the preprocessing, the evaluation type, and the detailed design of deep models.

To learn the spectral, spatial, as well as local and global information, Kwak et al. [125] propose a multi-level feature fusion method based on CNN. The study records the EEG signals from 62 participants during a working memory task with easy and hard levels, also included the *EEGLearn* in [44]. By transforming EEG data into 3D EEG images contained spectral and spatial information firstly, a 3D CNN (4 convolutional) is then constructed based on the feature fusion network. For the private dataset, the accuracy achieves with 90.8%; for the *EEGLearn* dataset, the accuracy is 93.9%. By comparison, we find the proposed 3D CNN model in [125] is the most robust one that using the dataset *EEGLearn.*

There are three input styles for CNN, raw data of EEG signal, or computed features or images. Using raw EEG signals as input may reduce the noise effect and the complexity of training to some extent. However, we find the raw data are rarely used in the cognitive workload analysis. If we use images as the input of CNN, then spectrograms or topographic maps may be the most prevalent choice [44] [123].

2) Other Deep Models

Other deep models for workload recognition include RNN, DBN, denoising autoencoder, and hybrid models.

RNN is an architecture to train sequential processing. The key difference between CNN and RNN is that CNN only considers the current input while RNN considers both current and previous input [95]. Typically, the input of a recurrent layer is from both the current activations of the preceding layer and its activations from a previous time step. Kuanar et al. [126] use RNN to learn robust features and predict four levels of the cognitive workload of working memory tasks. The study first transforms the EEG signals into a sequence of multi-spectral images. Then nine convolutional layers are used to extract spatial and spectral features, then the features are fed into recurrent LSTM layers to extract temporal information. The final accuracy for the cross-subject task is 92.5%. In [127], deep RNN with LSTM cells shows the effectiveness of learning temporal representation from statistical EEG features for the MATB task. The mean, variance, skewness, and kurtosis of frequency power distributions and their combinations are features, and a variety of RNN architectures are used and compared, including LSTM and some variants. The best accuracy is 93% with 2 hidden-layers LSTM.

Besides the above-mentioned supervised methods, DBN and autoencoder are unsupervised methods. DBN is a generative model with connections between the multiple layers of latent variables. DBN consists of one or more

Restricted Boltzmann Machines (RBM) that are trained by contrast divergence algorithms [128]. RBM is to learn a probability distribution over a set of inputs. A typical RBM consists of two layers, i.e., the visible and the hidden layers. A trained DBN can be used to generate training data with maximum probability. In [128], Yin et al. use DBN to assess the operator functional states of cognitive workload with seven classes, and this study mainly focuses on the optimal structure of RBM and channel selection based on the connection weights of the input layer and the first hidden layer. The results of comparison with backpropagation, LDA, and naïve Bayes confirmed the effectiveness of the DBN method, as for seven-class, the mean accuracies are 51%, 76%, 55%, and 93%, respectively. In [129], the same team applied switching DBN with adaptive weights to detect cognitive workload, fatigue, and the coupling effect between them across subjects. An ensemble DBN is also used to classify mental fatigue elicited by similar tasks [130].

Autoencoder is the symmetric single-hidden-layer neural network, where the output and input neurons are the same. It learns a representation by making the outputs approximate to the inputs [131]. Denoising autoencoder (DAE) is a special form of autoencoder, which minimizes the reconstruction loss of corrupted input features and learns a robust representation of the noise inputs [132]. Recently, Yin et al. propose to use the variants of DAE to tackle cognitive workload classification tasks, such as transfer DAE [38], Stacked DAE [133], adaptive Stacked DAE [134], and ensemble Stacked DAE [135]. For instance, each hidden layer of a Stacked DAE is an autoencoder that removed the output layer. Stacked DAE with 5 hidden layers [133] is used on single-channel EEG signals to classify binary workload levels in a subject-dependent way, and the frontal and occipital channels are salient, with the O2 channel getting 74% accuracy. To access cross-sessions workload classification of two days, an adaptive Stacked DAE with 6 hidden layers [134] is used to extract the high-level presentation of PSD features, via the adaptive shallow layer and multiple deep layers. Here, the shallow layers can be seen as filters to capture the optimal combination of PSD features. The workload classification accuracy of adaptive Stacked DAE is 85%, which improved 12% accuracy than Stacked DAE. To improve the performance of subject-dependent workload, an ensemble Stacked DAE [135] is proposed with ensemble learning and adding a locality-preserving-projection hidden layer to preserve local information. The average accuracy of ensemble Stacked DAE is 93%, while 88% for DAE and 85% for Stacked DAE.

Several studies apply hybrid architecture for classifying cognitive workload, which uses a combination of two or more standard deep learning models. The hybrid architectures can help to abstract multi-dimension information from temporal, spectral, and spatial dimensions [44,103], whereas single deep models may have shortcomings to extract these features. On the other hand, hybrid architectures can be used to fuse the salience features in some way. Examples are from the fusion of PSD and ERP features with two-stream Neuro Networks of CNN and temporal convolution networks [136], and spectral feature-enhanced maps combined with various CNNs [46]. Of note, the commonly used hybrid architectures for cognitive workload recognition are CNN combining with RNN layers, as applied in [44,136,103,137,124]. In [124], a Ternary-task Convolutional Bidirectional Neural Turing Machine is designed to classify four workload levels using dataset in [44], which are aimed to increase the inter-class variations and reduce the intra-class variations. The hybrid model includes CNN and Bidirectional Neural Turing Machine structures, where CNN disposes multi-spectral images to preserve the spatial and spectral representation, and the generated features are sent into Bidirectional Neural Turing Machine to extract temporal information. The mean classification accuracy is 96.3% with 10-fold cross-validation. To classify cognitive workload states of a cross-task [103], the study proposes an R3DCNN method combining RNN and 3DCNN. R3DCNN firstly constructs topographic maps of frequency and time information and then uses the proposed method to learn features of temporal, spectral, and spatial dimensions. The final mean cross-task classification accuracy is 88.9%. A recently published paper [56] uses a bidirectional LSTM and LSTM model combined with an evolutionary algorithm for classifying the resting state with two levels and SIMKAP task with three levels, using the STEW dataset in [55]. However, the data split strategy is not mentioned.

In Table III, we summarize the deep learning models for cognitive workload recognition, with the order of the following models, i.e., CNN, RNN, DBN, DAE, and hybrid. The table is similar to Table II, but we remove feature selection and include the network structures, such as the number of hidden layers in deep models (in Supplementary Table S2). We find the most effective deep models are CNN and hybrid models. Specifically, considering the number of subjects and validation strategy, we argue the proposed deep models in [125,126,136] perform more robust in this part. Also, interested readers are referred to a review [123], which gives the most recommended deep models for workload recognition are CNN and DBN, and summarizes characteristics and classifiers recommendations of DBN, CNN, and RNN.

V. Discussion

In the following, we discuss the main findings of the review, the open problems, and future trends of EEG-based cognitive workload recognition that require further investigation.

A. Summary of Major Findings

Here, we give an informed algorithm design aiming to help the readers designing algorithms based on the insights we gained from reviewing all the papers. The data quality, the classification levels, the task objective (subject dependent or not), and the time delay (real-time or not) may impact the design of the algorithm.

To get high-quality EEG signals, data preprocessing steps and feature extraction are necessary. Most papers use filtering, data segmenting, and artificial removal steps [16,18,20,117] in

preprocessing section. For feature extraction, we can use one specific domain features [31,79], e.g., the ERP features in the time domain [21,30], or PSD features in spectral domain [104-107], or extract various features from different domains and combined them as inputs for the classifier [29,32,70], expecting features fusion may provide more information than only one perspective [93,114]. In this way, the feature selection is needed to select the most relevant top N features. We found that feature selection methods have a varied performance and are dependent on the extracted features to some extent. Also, the range and the number of frequency bands are set and changed slightly specific to the studies.

Concerned with classification levels of the cognitive workload, two to seven are widely used. The conventional cognitive workload recognition benefits from hand-crafted feature extraction and supervised learning algorithms [45], and can handle the binary classification task well. In contrast to the classical models, the deep learning models might improve the evaluation performance of multi-class classification through powerful nonlinear feature representation, but require big data and more training time to tune the structure and parameter [103]. Also, the input features for classical models are vectors, whereas the inputs for CNN models are mostly spectral maps or images. Moreover, the deep learning models for EEG workload recognition are commonly shallow networks (e.g., the convolutional layers for CNN are ranged from 4 to 8). And the most effective deep models are CNN and hybrid models.

In the case of real-time experimental design, the channel selection to select the most related channels to reduce the cost of time is suggested. We find some papers use all available channels recorded in the EEG signals, in most cases. Some papers [18,31,73,79,101] directly use the channels in frontal or/ and parietal areas, as these areas are validated to be sensitive to cognitive workload. Some papers [93,128] use channel selection to select the optimal related channels, such as the techniques in feature selection. Though the high-density EEG channels are expected to provide more accurate recognition, the adoption of fewer channels or even single-channel will boost real-time workload recognition.

Given various task objectives, such as subject-dependent (subject-specific), subject-independent (cross-subject), and task-independent (cross-task), the studies have different research directions and focus. For example, the subject-dependent studies are focusing on accurate personalized models and ignoring the personal variability, while cross-based studies are getting more attention in the generalized models that can work well under different subjects or tasks. In recent years, regardless of the task environments, the mean accuracies of existing methods to EEG workload recognition are almost over 80%, which seems acceptable for practical applications. However, due to the high variability between subjects, the performance of cross-subject and cross-task are typically inferior to the subject-specific task [74,116]. Thus, for cross-based studies, the reduction of personal variability and then the extraction of shared feature representation from the vast majority are suggested.

B. Open Problems and Future Trends

Though much progress has been made, there are still some major problems and future trends that should be well investigated to further improve the recognition performance.

1) Open Access to Data

Most EEG datasets used in the studies covered by this review are private, and the publically available datasets related to cognitive workload are listed in Table I. Though several studies [45,46,56,124,125] used the public dataset *EEGLearn*, only a handful of studies share their code, as such, many reviewed studies cannot easily be reproduced. We argue the open access to data would ensure the fair comparison of experimental results and enhance the code reproducibility. Thus, one of the core problems is the availability of open database sharing and further the code-sharing using GitHub (https://github.com/) or some open source tools.

To ensure the fair comparison of experimental results across various models, it is thus important to create new benchmark datasets of the workload domain. Inspired by the development of emotion recognition using EEG signals, with widely used public datasets such as SEED [138] and DEAP [139], we thus encourage the researchers in the related domain to make their EEG dataset associated with free licenses or available upon request. Also, for code reproducibility, Roy et al. [140] recommend that the published studies should clearly describe the architecture of models, the data used, or existing datasets used, whenever possible. Besides these, the studies should also include state-of-the-art baselines, and share code.

2) Real-Time and Real-World Designs

In this review, we find many studies are offline [16,29,125], i.e., the training and testing phases of models are not given immediately but cost some time to get the recognition results. Also, wet or gel electrodes need much time for preparation. For real-world applications of operator states monitoring, we expect to recognize the workload levels in an online and real-time way [41,62,79,96,111].

Generally, the experiment of real-time workload detection consists of two consecutive or separated sessions, which are off-line training and online application of the tasks (same tasks [96,111,141] or different tasks [142]), respectively. Some papers adopted channel selection [143] or removal [111] to find the optimal channel sets, then constructed the subject-specific workload detector. A paper [144] constructed a real-time workload assessment system based on the offline classification results, and computed workload index as an indicator based on the frequency band power changes in experimental findings. The above methods are promising for real-time applications. We found deep models were rarely used for this case. Besides, we can also learn from EEG-based real-time emotion recognition [145,146] and fatigue detection used clustering techniques [147], which indicate the real-time models are specific to each subject with stable features. Also, it takes not only the classification accuracy but also the execution time into consideration [146].

On the pathway from laboratory settings to the real-world environments, some studies have adopted the real driving task [111], real air traffic management task [12], outdoor [148], virtual reality [87,112], and the doing-sporting task environment with fewer dry electrodes [63] for cognitive workload elicitation. With the improvement of the sensor of the wireless, wearable, and dry electrode [149], real-time workload detection will become increasingly popular, and beneficial for real scenario applications. In this way, the real-time and real-world mental workload recognition needs carefully hand-crafted design to find the most sensitive biomarkers to construct subject-specific models, that can discriminate and predict the workload change.

3) The Variabilities of Subjects and Tasks

EEG signals have great variabilities among subjects and tasks, including intra-subject, inter-subject, and inter-task variabilities. Here, the intra-subject variability is related to the standard cognitive workload recognition design [81], and many methods have been proposed and achieved acceptable performance [70,101,106,111]. The latter two are more complex and challenging, and typically having inferior performance to the former [16,74,81].

To reduce the intra-subject variabilities, some works try to collect more longitudinal data recordings and consider more accurate personalized models. To reduce inter-subject/task variabilities, several works proposed to first find the common patterns and then construct the subject/task-independent model, assuming a set of invariant features exists across subjects or tasks [16,94,102]. Though many approaches have been proposed, the performance of common features is limited. In this way, we can also learn lessons from other EEG-based classification tasks, such as motor imagery that adopting the data alignment method as a procedure [150], which aligns EEG trials from different subjects in the Euclidean space to make them more similar; and emotion recognition that using transfer learning [151,152], mainly domain adaptation methods [153], that transfer knowledge from a labeled source domain to an unlabeled target domain, aiming to reduce the distribution dependency between the training and testing data, thus making them similar. To our knowledge, domain adaptation has few or not been applied for mental workload systems, thus cross-task variability remaining an open issue, and these methods may be promising.

4) Advanced Machine Learning Techniques

Note that, the EEG signal is essentially high dimensional data with rich temporal, spectral, and spatial information, and has high variability among subjects. In this way, it is difficult to propose a general workload recognition method across different subjects and tasks. Thus, for EEG-based state monitoring, we should pay attention to some advanced machine learning techniques from every step, such as high-density EEG acquisition, new signal decomposing and signal denoising methods in signal preprocessing, new features, more efficient feature selection, and fusion in feature engineering, and classification models. For instance, Sun et al., use a 1D CNN model as a filter for automatic artifact removal, to get clean EEG signals [154]. Concerned with deep models, the limited data and overfitting problem of the model are worth devoting the efforts. More tricks and methods can be found in deep learning theories [118]. Moreover, the ideas in ensemble learning [114], and semi-supervised learning [155] would be helpful for the design of classification methods. For example, ensemble learning algorithms generate many base models, each of which is learned using a traditional algorithm, and combine the predictions of these models. Also, semi-supervised learning can be applied if we aim to use the information contained in the unlabeled testing data.

5) The Generalizability of Models

The generalizability of classical machine learning and deep learning models has gained much attention in many studies. Here, task\population generalizability is defined as the algorithm's ability to generalize across task\subject environment if the algorithm is task\subject-independent and achieves over 80% accuracy. Many works began to pay attention to the generalized model. We found the models in [18,29,31,45,53] are shown good populations generalizability, given nonlinear features and deep CNN models, whereas task generalizability of some models is poor [18,104,109,110]. However, it is difficult to determine the populations\task generalizability of subject- or task-specific algorithms, since the lack of good measure. And the accuracy of subject\task-specific algorithms trained on a population or another task will decrease, due to individual and task differences [103]. More works need to be done to investigate the more complex and cross-task design and develop more robust machine learning models.

6) The Interpretability of Models

Besides the evaluation performance and generalizability of machine learning models, interpretability is also needed to pay attention to. Interpretability is the degree to which a human can understand the cause of a decision and predict the model's result [156]. In contrast to the intrinsic interpretable models, e.g., decision trees, many other existing machine learning approaches are non-interpretable black-box models [156]. Black-box models can only be obtained with the final results, and lack interpretability, especially for deep models.

To increase the interpretability, some studies adopt suitable visualization techniques and statistical tests. For example, to display the feature sensitivity of various workload levels, the visualization of task-related power change in theta and alpha frequency bands is used [10]. One-way ANOVA [128] and Wilcoxon signed-rank test [85] are used to test the statistical significance of different features or classifiers. For deep models, visualizing weights of CNN layers [44] or reconstructing the EEG feature maps via deconvolution [136] is used. These techniques offer good examples to enhance interpretability, also, newer visualization methods and newer interpretable models could be helpful.

VI. Conclusion

In this paper, we comprehensively review the available literature on EEG-based cognitive workload recognition using machine learning. We highlight the general steps of classical machine learning, i.e., data acquisition, preprocessing, feature extraction and selection, classification and evaluation. We also review several widely used deep learning models for workload recognition. Further, we discuss the main findings of this work. Finally, we list the open problems and trends from the aspects of the dataset, experiment design, subject variability, the advancement, generalizability, and interpretability of models, that are required further investigation.

References

[1] R. Mehta, and R. Parasuraman, "Neuroergonomics: a review of applications to physical and cognitive work," *Frontiers in Human Neuroscience*, vol. 7, no. 889, 2013-December-23, 2013.

[2] P. Arico, G. Borghini, G. D. Flumeri, N. Sciaraffa, A. Colosimo, and F. Babiloni, "Passive BCI in Operational Environments: Insights, Recent Advances, and Future Trends," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1431-1436, 2017.

[3] Z. Gao, X. Wang, Y. Yang, C. Mu, Q. Cai, W. Dang, and S. Zuo, "EEG-Based Spatio-Temporal Convolutional Neural Network for Driver Fatigue Evaluation," *IEEE Transactions on Neural Networks & Learning Systems*, pp. 1-9, 2019.

[4] X. Wang, D. Nie, and B. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94-106, 2014.

[5] C. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, pp. 159 - 177, 2002.

[6] J. Heard, C. E. Harriott, and J. A. Adams, "A Survey of Workload Assessment Algorithms," *IEEE Transactions on Human-Machine Systems*, pp. 1-18, 2018.

[7] S. G. Hart, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, no. 6, pp. 139-183, 1988.

[8] G. B. Reid, and T. E. Nygren, "The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload," *Advances in psychology*, vol. 52, pp. 185-218, 1988.

[9] H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang, "A Systematic Review of Physiological Measures of Mental Workload," *International Journal of Environmental Research and Public Health*, vol. 16, pp. 2716, 2019.

[10] M. E. Smith, A. Gevins, H. Brown, A. Karnik, and R. Du, "Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction," *Human Factors*, vol. 43, no. 3, pp. 366-380, 2001.

[11] S. Lemm, B. Blankertz, T. Dickhaus, and K. Muller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387-399, 2011.

[12] N. Sciaraffa, P. Aricò, G. Borghini, G. D. Flumeri, A. D. Florio, and F. Babiloni, "On the Use of Machine Learning for EEG-Based Workload Assessment: Algorithms Comparison in a Realistic Task," *Human Mental Workload: Models and Applications*, pp. 170-185, 2019.

[13] C. R. L., and N. Jim, "Measuring mental workload using physiological measures: A systematic review," *Applied Ergonomics*, vol. 74, pp. 221-232, 2019.

[14] E. Debie, R. F. Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. G. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal Fusion for Objective Assessment of Cognitive Workload: A Review," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 1-14, 2019.

[15] S. Oviatt, "Human-centered design meets cognitive load theory: designing interfaces that help people think." *MM '06: Proceedings of the 14th ACM international conference on Multimedia*, pp. 871-880, 2006.

[16] G. N. Dimitrakopoulos, I. Kakkos, Z. Dai, J. Lim, J. Desouza, A. Bezerianos, and Y. Sun, "Task-Independent Mental Workload Classification Based Upon Common Multiband EEG Cortical Connectivity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1940-1949, 2017.

[17] P. D. Antonenko, F. Paas, R. H. Grabner, and T. Van Gog, "Using Electroencephalography to Measure Cognitive Load," *Educational Psychology Review*, vol. 22, no. 4, pp. 425-438, 2010.

[18] M. Mazher, A. A. Aziz, A. S. Malik, and H. U. Amin, "An EEG-Based Cognitive Load Assessment in Multimedia Learning Using Feature Extraction and Partial Directed Coherence," *IEEE Access*, vol. 5, pp. 14819-14829, 2017.

[19] J. M. Angel, R. Cristian, and L. Hernan, "Using Psychophysiological Sensors to Assess Mental Workload During Web Browsing," *Sensors*, vol. 18, no. 2, pp. 458, 2018.

[20] M. A. Almogbel, A. H. Dang, and W. Kameyama, "Cognitive Workload Detection from Raw EEG-Signals of Vehicle Driver using Deep Learning." *International Conference on Advanced Communication Technology (ICACT)*, pp. 1-6, 2018.

[21] F. Dehais, A. Duprès, S. Blum, N. Drougard, S. Scannella, R. Roy, and F. Lotte, "Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions," *Sensors*, vol. 19, no. 6, 2019.

[22] P. Aricò, G. Borghini, G. D. Flumeri, A. Colosimo, and F. Babiloni, "A passive brain-computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks," *Progress in Brain Research*, vol. 228, pp. 295-328, 2016.

[23] A. Abrantes, E. Comitz, P. Mosaly, and L. Mazur, "Classification of EEG Features for Prediction of Working Memory Load," *Springer International Publishing*, pp. 115-126, 2017.

[24] G. Liang, J. Lin, S. Hwang, F. Huang, T. Yenn, and C. Hsu, "Evaluation and prediction of on-line maintenance workload in nuclear power plants," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 19, no. 1, pp. 64-77, 2009.

[25] S. Mathan, A. Smart, T. Ververs, and M. Feuerstein, "Towards an index of cognitive efficacy EEG-based estimation of cognitive load among individuals experiencing cancer-related cognitive decline," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2010, pp. 6595-6598, 2010.

[26] W. Wu, Y. Zhang, J. Jiang, M. V. Lucas, G. Fonzo, C. E. Rolle, C. Cooper, C. Chinfatt, N. Krepel, and C. Cornelssen, "An electroencephalographic signature predicts antidepressant response in major depression," *Nature Biotechnology*, pp. 1-9, 2020.

[27] X. Du, J. Li, D. Xiong, Z. Pan, F. Wu, Y. Ning, J. Chen, and K. Wu, "Research on electroencephalogram specifics in patients with schizophrenia under cognitive load," *Journal of Biomedical Engineering*, vol. 37, no. 1, pp. 45, 2020.

[28] L. Zhang, J. Wade, D. Bian, J. Fan, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Cognitive Load Measurement in a Virtual Reality-Based Driving System for Autism Intervention," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 176-189, 2017.

[29] S. Wang, J. Gwizdka, and W. A. Chaovalitwongse, "Using Wireless EEG Signals to Assess Memory Workload in the -Back Task," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 424-435, 2016.

[30] R. N. Roy, S. Charbonnier, A. Campagne, and S. Bonnet, "Efficient mental workload estimation using task-independent EEG features," *Journal of Neural Engineering*, vol. 13, no. 2, pp. 026019, 2016.

[31] P. Zarjam, J. Epps, and N. H. Lovell, "Beyond Subjective Self-Rating: EEG Signal Classification of Cognitive Workload," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 4, pp. 301-310, 2015.

[32] N. Friedman, T. Fekete, K. Gal, and O. Shriki, "EEG-Based Prediction of Cognitive Load in Intelligence Tests," *Frontiers in Human Neuroscience*, vol. 13, no. 191, 2019.

[33] A. Knoll, Y. Wang, F. Chen, J. Xu, N. Ruiz, J. Epps, and P. Zarjam, "Measuring cognitive workload with low-cost electroencephalograph." *International Conference on Human Computer Interaction*, pp. 568-571, 2011.

[34] K. Yu, I. Prasad, H. Mir, N. Thakor, and H. Al-Nashash, "Cognitive workload modulation through degraded visual stimuli: a single-trial EEG study," *Journal of Neural Engineering*, vol. 12, no. 4, pp. 046020, 2015.

[35] P. Arico, G. Borghini, G. D. Flumeri, A. Colosimo, and F. Babiloni, "Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks." *2015 37th Annual*

International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 7242-7245, 2015.

[36] L. R. Fournier, G. F. Wilson, and C. R. Swain, "Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training," *International Journal of Psychophysiology*, vol. 31, no. 2, pp. 129-145, 1999.

[37] L. J. Prinzel, F. G. Freeman, M. W. Scerbo, P. J. Mikulka, and A. T. Pope, "Effects of a Psychophysiological System for Adaptive Automation on Performance, Workload, and the Event-Related Potential P300 Component," *Human Factors*, vol. 45, no. 4, pp. 601-614, 2003.

[38] Z. Yin, M. Zhao, W. Zhang, Y. Wang, Y. Wang, and J. Zhang, "Physiological-signal-based mental workload estimation via transfer dynamical autoencoders in a deep learning framework," *Neurocomputing*, vol. 347, pp. 212-229, 2019.

[39] M. I. Jordan, and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255-260, 2015.

[40] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery," *Chemical Reviews*, vol. 119, no. 18, pp. 10520-10594, 2019.

[41] K. R. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, "Machine learning for real-time single-trial EEG-analysis: From brain–computer interfacing to mental state monitoring," *J Neurosci Methods*, vol. 167, no. 1, pp. 82-90, 2008.

[42] A. X. Stewart, A. Nuthmann, and G. Sanguinetti, "Single-trial classification of EEG in a visual object task using ICA and machine learning," *Journal of Neuroscience Methods*, vol. 228, pp. 1-14, 2014.

[43] P. Bashivan, M. Yeasin, and G. M. Bidelman, "Single trial prediction of normal and excessive cognitive load through EEG feature fusion," *IEEE Signal Processing in Medicine and Biology Symposium*, pp. 1-5, 2015.

[44] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks," arXiv preprint arXiv: 1511.06448, 2015.

[45] Jiao, Zhicheng, Gao, Xinbo, Wang, Ying, Li, Jie, Xu, and Haojun, "Deep Convolutional Neural Networks for mental load classification based on EEG data," *Pattern Recognition*, vol. 76, pp. 582-595, 2018.

[46] Y. Zhang, and Y. Shen, "Parallel Mechanism of Spectral Feature-Enhanced Maps in EEG-Based Cognitive Workload Classification," *Sensors*, vol. 19, no. 4, pp. 808, 2019.

[47] I. Zyma, S. Tukaev, I. Seleznov, K. Kiyono, A. Popov, M. Chernykh, and O. Shpenkov, "Electroencephalograms during Mental Arithmetic Task Performance," *Data*, vol. 4, no. 1, 2019.

[48] I. Seleznov, I. Zyma, K. Kiyono, S. Tukaiev, and O. Shpenkov, "Detrended Fluctuation, Coherence, and Spectral Power Analysis of Activation Rearrangement in EEG Dynamics During Cognitive Workload," *Frontiers in Human Neuroscience*, vol. 13, 2019.

[49] J. Shin, A. Von Lühmann, D. W. Kim, J. Mehnert, H. J. Hwang, and K. R. Müller, "Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open-access dataset," *Sci Data*, vol. 5, pp. 180003, 2018.

[50] P. K. Liu, W. K. Beh, C. Y. Shih, Y. T. Chen, and A. Y. A. Wu, "Entropy and Complexity Assisted EEG-based Mental Workload Assessment System," *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE*, 2019.

[51] M. Saadati, J. Nelson, and H. Ayaz, "Multimodal fNIRS-EEG Classification Using Deep Learning Algorithms for Brain-Computer Interfaces Purposes," *Advances in Neuroergonomics and Cognitive Engineering*, pp. 209-220, 2019.

[52] M. Saadati, J. Nelson, and H. Ayaz, "Convolutional Neural Network for Hybrid fNIRS-EEG Mental Workload Classification," *Advances in Neuroergonomics and Cognitive Engineering*, pp. 221-232, 2019.

[53] E. Boran, T. Fedele, A. Steiner, P. Hilfiker, and J. Sarnthein, "Dataset of human medial temporal lobe neurons, scalp and intracranial EEG during a verbal working memory task," *Scientific Data*, vol. 7, no. 1, pp. 30, 2020.

[54] Ece, Boran, Tommaso, Fedele, Peter, Klaver, Hilfiker, Lennart, Stieglitz, and Thomas, "Persistent hippocampal neural firing and hippocampal-cortical coupling predict verbal working memory load," *Science Advances*, 2019.

[55] W. L. Lim, O. Sourina, and L. P. Wang, "STEW: Simultaneous Task EEG Workload Data Set," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2106-2114, 2018.

[56] D. D. Chakladar, S. Dey, P. P. Roy, and D. P. Dogra, "EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm," *Biomedical Signal Processing and Control*, vol. 60, pp. 101989, 2020.

[57] A. Delorme, and S. Makeig, "EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9-21, 2004.

[58] G. Alexandre, "MEG and EEG data analysis with MNE-Python," *Front Neuro*, vol. 7, no. 7, pp. 267, 2013.

[59] W. Peng, "EEG Preprocessing and Denoising," *EEG Signal Processing and Feature Extraction*, L. Hu and Z. Zhang, eds., pp. 71-87, Singapore: Springer Singapore, 2019.

[60] Luck SJ, "An introduction to the event-related potential technique," Cambridge, MA: MIT Press, 2014.

[61] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229-240, 2011.

[62] T. Mullen, C. Kothe, Y. M. Chi, A. Ojeda, T. Kerth, S. Makeig, G. Cauwenberghs, and T.-P. Jung, "Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG," Annual International Conference of the IEEE Engineering in Medicine and Biology Society. *IEEE Engineering in Medicine and Biology Society*, vol. 2013, pp. 2184-2187, 2013

[63] O. Rosanne, I. Albuquerque, J. F. Gagnon, S. Tremblay, and T. H. Falk, "Performance Comparison of Automated EEG Enhancement Algorithms for Mental Workload Assessment of Ambulant Users," *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019.

[64] R. Jenke, A. Peer, and M. Buss, "Feature Extraction and Selection for Emotion Recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327-339, 2014.

[65] A. Kok, "On the utility of P3 amplitude as a measure of processing capacity," *Psychophysiology*, vol. 38, no. 3, pp. 557-577, 2010.

[66] A. M. Brouwer, M. A. Hogervorst, J. B. F. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using EEG spectral power and ERPs in the n-back task," *Journal of Neural Engineering*, vol. 9, no. 4, pp. 045008, 2012.

[67] A. Kabbara, "Brain network estimation from dense EEG signals: application to neurological disorders," *Neurons and Cognition,* 2018.

[68] G. F. Wilson, and C. A. Russell, "Performance Enhancement in an Uninhabited Air Vehicle Task Using Psychophysiologically Determined Adaptive Aiding," *Human Factors*, vol. 49, no. 6, pp. 1005-1018, 2007.

[69] C. Mühl, C. Jeunet, and F. Lotte, "EEG-based workload estimation across affective contexts," *Frontiers Neurosci*., vol. 8, p. 114, Jun. 2014.

[70] P. Zarjam, J. Epps, and F. Chen, "Characterizing working memory load using EEG delta activity," *in Proc. 19th Eur. Signal Process*. Conf., Aug. pp. 1554–1558, 2011.

[71] C. Dijksterhuis, D. de Waard, K. Brookhuis, B. Mulder, and R. de Jong, "Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns," *Frontiers Neurosci*., vol. 7, p. 149, Aug. 2013.

[72] P. Zarjam, J. Epps, F. Chen, and N. H. Lovell, "Estimating cognitive workload using wavelet entropy-based features during an arithmetic task," *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2186-2195, 2013.

[73] G. Borghini et al., "Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices," *in Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2012, pp. 6442–6445.

[74] C. Walter, S. Schmidt, W. Rosenstiel, P. Gerjets, and M. Bogdan, "Using Cross-Task Classification for Classifying Workload Levels in Complex Learning Tasks," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 876-881, 2013.

[75] Z. Wang, R. M. Hope, Z. Wang, Q. Ji, and W. D. Gray, "Cross-subject workload classification with a hierarchical Bayes model," *Neuroimage*, vol. 59, no. 1, pp. 64-69, 2012.

[76] M. K. Kıymık, İ. Güler, A. Dizibüyük, and M. Akın, "Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application," *Computers in Biology and Medicine*, vol. 35, no. 7, pp. 603-616, 2005.

[77] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084-1093, 2007.

[78] A. Medl, "Time Frequency and Wavelets in Biomedical Signal Processing," *IEEE Engineering in Medicine & Biology Magazine*, vol. 17, no. 6, pp. 15-97, 1998.

[79] Y. Tian, H. Zhang, Y. Jiang, P. Li, and Y. Li, "A Fusion Feature for Enhancing the Performance of Classification in Working Memory Load With Single-Trial Detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 10, pp. 1985-1993, 2019.

[80] R. N. Roy, S. Bonnet, S. Charbonnier, and A. Campagne, "Mental fatigue and working memory load estimation: Interaction and implications for EEG based passive BCI," *IEEE Engineering in Medicine and Biology Society*, vol. 2013, pp. 6607-6610, 2013.

[81] A. Appriou, A. Cichocki and F. Lotte, "Modern Machine-Learning Algorithms: For Classifying Cognitive and Affective States From Electroencephalography Signals," *in IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 3, pp. 29-38, July 2020.

[82] D. P. Subha, P. K. Joseph, A. U. Rajendra, and C. M. Lim, "EEG signal analysis: a survey," *Journal of Medical Systems*, vol. 34, no. 2, pp. 195-212, 2010.

[83] Y. Ma, W. Shi, C. K. Peng, and A. C. Yang, "Nonlinear dynamical analysis of sleep electroencephalography using fractal and entropy approaches," *Sleep Medicine Reviews*, pp. S1087079217300187, 2017.

[84] Y. Bai, X. Li, and Z. Liang, "Nonlinear Neural Dynamics," *EEG Signal Processing and Feature Extraction*, L. Hu and Z. Zhang, eds., pp. 215-240, Singapore: Springer Singapore, 2019.

[85] S. I. Dimitriadis, Y. Sun, K. Kwok, N. A. Laskaris, N. Thakor, and A. Bezerianos, "102," *Annals of Biomedical Engineering*, vol. 43, no. 4, pp. 977-989, 2015.

[86] V. Sakkalis, "Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG," *Computers in Biology and Medicine*, vol. 41, no. 12, pp. 1110-1117, 2011.

[87] I. Kakkos, G. N. Dimitrakopoulos, L. Gao, Y. Zhang, P. Qi, G. K. Matsopoulos, N. Thakor, A. Bezerianos, and Y. Sun, "Mental Workload Drives Different Reorganizations of Functional Cortical Connectivity Between 2D and 3D Simulated Flight Experiments," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 9, pp. 1704-1713, 2019.

[88] B. Mwangi, T. S. Tian, and J. C. Soares, "A Review of Feature Reduction Techniques in Neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229-244, 2014.

[89] X. Wu, X. Xu, J. Liu, H. Wang, B. Hu, and F. Nie, "Supervised feature selection with orthogonal regression and feature weighting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1-8.

[90] D. Sean, C. Caitlin, H. Levi, P. Todd, J. Arun, and J. Blake, "A Simple ERP Method for Quantitative Analysis of Cognitive Workload in Myoelectric Prosthesis Control and Human-Machine Interaction," *Plos One*, vol. 9, no. 11, pp. e112091-, 2014.

[91] M. Plechawska-Wójcik, M. Tokovarov, M. Kaczorowska, and D. Zapaa, "A Three-Class Classification of Cognitive Workload Based on EEG Spectral Data," *Applied sciences*, vol. 9, no. 24, pp. 5340, 2019.

[92] H. Peng, F. Long, and C. Ding, "Feature Selection Based On Mutual Information: Criteria of Max-Dependency Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[93] B. S. Cheema, S. Samima, M. Sarma, and D. Samanta, "Mental Workload Estimation from EEG Signals Using Machine Learning Algorithms," *Engineering Psychology and Cognitive Ergonomics*. pp. 265-284, 2018.

[94] Y. Ke, H. Qi, H. Feng, L. Shuang, Z. Xin, Z. Peng, L. Zhang, and M. Dong, "An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task," *Frontiers in Human Neuroscience*, vol. 8, pp. 703, 2014.

[95] J. A. Blanco, M. K. Johnson, K. J. Jaquess, H. Oh, L. Lo, R. J. Gentili, and B. D. Hatfield, "Quantifying Cognitive Workload in Simulated Flight Using Passive, Dry EEG Measurements," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 373-383, 2018.

[96] G. F. Wilson, and C. A. Russell, "Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks," *Human Factors*, vol. 45, no. 4, pp. 635-644, 2003.

[97] M. V. Kosti, K. Georgiadis, D. A. Adamos, N. Laskaris, D. Spinellis, and L. Angelis, "Towards an affordable brain computer interface for the assessment of programmers' mental workload," *International Journal of Human-computer Studies*, vol. 115, pp. 52-66, 2018.

[98] A. Subasi, and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8659-8666, 2010.

[99] C. M. Bishop, Pattern Recognition, M. Jordan and J. Kleinberg, Eds. New York, NY, USA: Springer-Verlog, vol. 128, 2006.

[100] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," *Proc. Adv. Neural Inf. Process. Syst*., 2004, pp. 1569–1576.

[101] P. Zarjam, J. Epps, and F. Chen, "Spectral EEG features for evaluating cognitive load," *International Conference of the IEEE IEEE Engineering in Medicine and Biology Society (EMBS)*, Boston, Massachusetts USA, August 30 - September 3, pp. 3841-3844, 2011.

[102] W. K. Y. So, S. W. H. Wong, J. N. Mak, R. H. M. Chan, and M. Emmanuel, "An evaluation of mental workload with frontal EEG," *Plos One*, vol. 12, no. 4, pp. e0174949, 2017.

[103] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 1, pp. 31-42, 2019.

[104] J. Zhang, Z. Yin, and R. Wang, "Recognition of Mental Workload Levels Under Complex Human–Machine Collaboration by Using Physiological Features and Adaptive Support Vector Machines," *IEEE Transactions on Human Machine Systems*, vol. 45, no. 2, pp. 200-214, 2015.

[105] J. Zhang, Z. Yin, and R. Wang, "Pattern Classification of Instantaneous Cognitive Task-load Through GMM Clustering, Laplacian Eigenmap, and Ensemble SVMs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, pp. 947-965, 2017.

[106] Z. Yin, and J. Zhang, "Operator functional state classification using least-square support vector machine based recursive feature elimination technique," *Computer Methods & Programs in Biomedicine*, vol. 113, no. 1, pp. 101-115, 2014.

[107] A. Gevins, M. E. Smith, H. Leong, L. K. Mcevoy, S. Whitfield, R. Du, and G. Rush, "Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods," *Human Factors*, vol. 40, no. 1, pp. 79-91, 1998.

[108] G. F. Wilson, and C. A. Russell, "Operator Functional State Classification Using Multiple Psychophysiological Features in an Air Traffic Control Task," *Human Factors*, vol. 45, no. 3, pp. p.381-389, 2003.

[109] C. L. Baldwin, and B. N. Penaranda, "Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification," *Neuroimage*, vol. 59, no. 1, pp. 48-56, 2012.

[110] B. N. Penaranda, and C. L. Baldwin, "Temporal Factors of EEG and Artificial Neural Network Classifiers of Mental Workload," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1, pp. 188-192, 2012.

[111] D. Guido, R. M. José del, H. Thilo, J. M. Dennis, and M. Klaus-Robert, "Improving Human Performance in a Real Operating Environment through Real-Time Mental Workload Detection," *Toward Brain-Computer Interfacing*, pp. 409-422: MITP, 2007.

[112] C. Tremmel, C. Herff, T. Sato, K. Rechowicz, Y. Yamani, and D. J. Krusienski, "Estimating Cognitive Workload in an Interactive Virtual Reality Environment Using EEG," *Frontiers in Human Neuroscience*, vol. 13, no. 401, 2019.

[113] G. Borghini, P. Aricò, G. D. Flumeri, G. Cartocci, A. Colosimo, S. Bonelli, A. Golfetti, J. P. Imbert, G. Granger, and R. Benhacene, "EEG-Based Cognitive Control Behavior Assessment: an Ecological study with Professional Air Traffic Controllers," *Sci Rep*, vol. 7, no. 1, pp. 547, 2017.

[114] J. Tao, Z. Yin, L. Liu, Y. Tian, Z. Sun, and J. Zhang, "Individual-Specific Classification of Mental Workload Levels Via an

Ensemble Heterogeneous Extreme Learning Machine for EEG Modeling," *Symmetry*, vol. 11, no. 7, pp. 944, 2019.

[115] M. Arvaneh, A. Umilta, and I. H. Robertson, "Filter bank common spatial patterns in mental workload estimation," *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4749-4752, 2015.

[116] A. Appriou, A. Cichocki, and F. Lotte, "Towards Robust Neuroadaptive HCI: Exploring Modern Machine Learning Methods to Estimate Mental Workload from EEG Signals," *Human Factors in Computing Systems*, 2018.

[117] Z. Dai, B. Anastasios, S. H. Chen, and Y. Sun, "Mental workload classification in n-back tasks based on single-trial EEG," *Chinese Journal of Scientific Instrument*, vol. 38, pp. 1335-1344, 2017.

[118] Y. Lecun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[119] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. J. Junior, X. Baro, and H. Demirel, "Dominant and Complementary Emotion Recognition from Still Images of Faces," *IEEE Access*, vol. 6, pp. 26391-26403, 2018.

[120] X. Ma, S. Qiu, C. Du, J. Xing, and H. He, "Improving EEG-Based Motor Imagery Classification via Spatial and Temporal Recurrent Neural Networks," *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1903-1906, 2018.

[121] W. L. Zheng, J. Y. Zhu, Y. Peng, and B. L. Lu, "EEG-Based Emotion Classification Using Deep Belief Networks," *International Conference on Multimedia and Expo*, pp. 1-6, 2014.

[122] G. Li, C. H. Lee, J. J. Jung, Y. C. Youn, and D. Camacho, "Deep learning for EEG data analytics: A survey," *Concurrency and Computation: Practice and Experience*, e519, 2019.

[123] A. Craik, Y. He, and J. L. Contreras-vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, pp. 031001, 2019.

[124] W. Qiao, and X. Bi, "Ternary-task convolutional bidirectional neural turing machine for assessment of EEG-based cognitive workload," *Biomedical Signal Processing and Control*, vol. 57, pp. 101745, 2020.

[125] Y. Kwak, K. Kong, W.-J. Song, B.-K. Min, and S.-E. Kim, "Multilevel Feature Fusion with 3D Convolutional Neural Network for EEG Based Workload Estimation," *IEEE Access*, no. 99, pp. 16009-16021, 2020.

[126] S. Kuanar, V. Athitsos, N. Pradhan, A. Mishra, and K. R. Rao, "Cognitive Analysis of Working Memory Load from EEG by a Deep Recurrent Neural Network," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 2576-2580, 2018.

[127] R. G. Hefron, B. J. Borghetti, J. C. Christensen, and C. M. S. Kabban, "Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation," *Pattern Recognition Letters*, vol. 94, pp. 96-104, 2017.

[128] J. Zhang, and S. Li, "A deep learning scheme for mental workload classification based on restricted Boltzmann machines," *Cognition, Technology & Work*, vol. 19, no. 4, pp. 607-631, 2017.

[129] Z. Yin, and J. Zhang, "Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights," *Neurocomputing*, vol. 260, no. 18, pp. 349-366, 2017.

[130] Z. Yin, and J. Zhang, "Cross-subject classification of mental fatigue by neurophysiological signals and ensemble deep belief networks," *Chinese Control Conference*, pp. 10966-10971, 2017.

[131] Y. J. Fan, "Autoencoder node saliency: Selecting relevant latent representations," Pattern Recognition, vol. 88, pp. 643-653, 2019.

[132] J. Li, Z. Struzik, L. Zhang, and A. Cichocki, "Feature Learning from Incomplete EEG with Denoising Autoencoder," *Neurocomputing*, vol. 165, no. OCT.1, pp. 23-31, 2015.

[133] Y. Zhong, and J. Zhang, "Recognition of Cognitive Task Load Levels Using Single Channel EEG and Stacked Denoising Autoencoder," *Proceedings of the 35th Chinese Control Conference IEEE*, pp. 3907-3912, 2016.

[134] Z. Yin, and J. Zhang, "Cross-session classification of mental workload levels using EEG and an adaptive deep learning model," *Biomedical Signal Processing and Control*, vol. 33, pp. 30-47, 2017.

[135] S. Yang, Z. Yin, Y. Wang, W. Zhang, Y. Wang, and J. Zhang, "Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders," *Computers in Biology and Medicine*, pp. 159-170, 2019.

[136] P. Zhang, X. Wang, J. Chen, W. You, and W. Zhang, "Spectral and Temporal Feature Learning with Two-Stream Neural Networks for Mental Workload Assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1149-1159, 2019.

[137] R. G. Hefron, B. J. Borghetti, C. M. S. Kabban, J. C. Christensen, and J. R. Estepp, "Cross-Participant EEG-Based Assessment of Cognitive Workload Using Multi-Path Convolutional Recurrent Neural Networks," *Sensors*, vol. 18, no. 5, pp. 1339, 2018.

[138] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18-31, 2012.

[139] W. Zheng, and B. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162-175, 2015.

[140] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, pp. 051001, 2019.

[141] R. Matthews, P. J. Turner, N. J. McDonald, K. Ermolaev, T. M. Manus, R. A. Shelby, and M. Steindorf, "Real time workload classification from an ambulatory wireless EEG system using hybrid EEG electrodes." *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vancouver, BC, pp. 5871-5875, 2008.

[142] G. Zhao, Y. Liu, and Y. Shi, "Real-Time Assessment of the Cross-Task Mental Workload Using Physiological Measures During Anomaly Detection," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 149-160, 2018.

[143] T. Hwang, M. Kim, M. Hwangbo, and E. Oh, "Optimal set of EEG electrodes for real-time cognitive workload monitoring." *The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014)*, JeJu Island, pp. 1-2, 2014.

[144] Shen, R., et al. Real-Time Workload Assessment Using EEG Signals in Virtual Reality Environment. *Springer International Publishing*, 2016.

[145] Y. Liu, O. Sourina and M. K. Nguyen, "Real-Time EEG-Based Human Emotion Recognition and Visualization," *2010 International Conference on Cyberworlds*, Singapore, pp. 262-269, 2010.

[146] Y. Liu, M. Yu, G. Zhao, J. Song, Y. Ge and Y. Shi, "Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 550-562, 2018.

[147] E. Q. Wu, L. M. Zhu, W. M. Zhang, P. Y. Deng, and G. R. Zhou, "Novel Nonlinear Approach for Real-Time Fatigue EEG Data: An Infinitely Warped Model of Weighted Permutation Entropy," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1-12, 2019.

[148] J. E. Reiser, E. Wascher, and S. Arnau, "Recording mobile EEG in an outdoor environment reveals cognitive-motor interference dependent on movement complexity," *Scientific Reports*, vol. 9, no. 1, pp. 1-14, 2019.

[149] R. J. Gentili, K. J. Jaquess, I. M. Shuggi, E. P. Shaw, H. Oh, L. Lo, Y. Y. Tan, C. A. Domingues, J. A. Blanco, and J. C. Rietschel, "Combined assessment of attentional reserve and cognitive-motor effort under various levels of challenge with a dry EEG system," *Psychophysiology*, vol. 55, no. 6, 2018.

[150] H. He, and D. Wu, "Transfer Learning for Brain-Computer Interfaces: A Euclidean Space Data Alignment Approach," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 399-410, 2020.

[151] W. Zheng, and B. Lu, "Personalizing EEG-based affective models with transfer learning," *International Joint Conference on Artificial Intelligence*, pp. 2732-2738, 2016.

[152] Wu D, Xu Y, Lu B L. Transfer Learning for EEG-Based Brain-Computer Interfaces: A Review of Progress Made Since 2016[J]. *IEEE Transactions on Cognitive and Developmental Systems*, PP(99):1-1, 2020.

[153] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain Adaptation Techniques for EEG-Based Emotion Recognition: A Comparative Study on Two Public Datasets," *IEEE*

Transactions on Cognitive and Developmental Systems, vol. 11, no. 1, pp. 85-94, 2019.

[154] W. Sun, Y. Su, X. Wu, and X. Wu, "A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals," *Neurocomputing*, vol. 404, 2020.

[155] J. Zhang, J. Li, and R. Wang, "Instantaneous mental workload assessment using time-frequency analysis and semi-supervised learning," *Cognitive Neurodynamics*, 2020.

[156] Christoph Molnar, "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable", 2019.

Yueying Zhou is currently pursuing a Ph.D. degree in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China.
Her research interests include brain-computer interface, EEG analysis, and machine learning.

Shuo Huang is currently pursuing a Ph.D. degree in software engineering with NUAA.
His current research interests include machine learning and human brain decoding.

Ziming Xu is currently working toward a master's degree in the College of Computer Science and Technology, NUAA.
His research interests include EEG analysis and machine learning.

Pengpai Wang is currently working toward a Ph.D. degree in the College of Computer Science and Technology, NUAA.
His research interests include EEG analysis and deep learning.

Xia Wu received the B.S., M.S., and Ph.D. degrees from Beijing Normal University, China, in 2001, 2004, and 2008 respectively. She is currently a Professor at the School of Artificial Intelligence, Beijing Normal University, China.
Her research interests include intelligent signal processing and EEG analysis.

Daoqiang Zhang received the B.S. and Ph.D. degrees in computer science from NUAA, China, in 1999, and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a lecturer in 2004 and is a professor at present.
His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published more than 200 scientific articles with more than 12,000 citations by Google Scholar. He is a fellow of the International Association for Pattern Recognition (IAPR).