Expert Systems With Applications 249 (2024) 123717

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Application of artificial intelligence in cognitive load analysis using functional near-infrared spectroscopy: A systematic review

Mehshan Ahmed Khan [a,*], Houshyar Asadi [a], Li Zhang [b], Mohammad Reza Chalak Qazani [c], Sam Oladazimi [a], Chu Kiong Loo [d], Chee Peng Lim [a], Saeid Nahavandi [e]

[a] Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong Warun Ponds Campus, Victoria 3216, Australia
[b] Department of Computer Science, Royal Holloway, University of London, Surrey TW20 0EX, United Kingdom
[c] Facuaty of Computing and Information, Sohar University, Albatinah North Governorate 311, Oman,
[d] Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia
[e] Swinburne University of Technonology, Hawthorn, Victoria, 3122, Australia

ARTICLE INFO

Keywords:
Functional Near-InfraRed Spectroscopy (fNIRS)
Deep learning
Machine learning
Cognitive load
Artificial intelligence

ABSTRACT

Cognitive load theory suggests that overloading of working memory may negatively affect the performance of human in cognitively demanding tasks. Evaluation of cognitive load is a difficult task; it is often assessed through feedback and evaluation from experts. Cognitive load classification based on Functional Near-InfraRed Spectroscopy (fNIRS) is now one of the key research areas in recent years, due to its resistance of artefacts, cost-effectiveness, and portability. To make fNIRS more practical in various applications, it is necessary to develop robust algorithms that can automatically classify fNIRS signals and less reliant on trained signals. Many of the analytical tools used in cognitive sciences have used Deep Learning (DL) modalities to uncover relevant information for mental workload classification. This review investigates the research questions on the design and overall effectiveness of DL as well as its key characteristics. We have identified 45 studies published between 2011 and 2023, that specifically proposed Machine Learning (ML) models for classifying cognitive load using data obtained from fNIRS devices. Those studies were analyzed based on type of feature selection methods, input, and DL model architectures. Most of the existing cognitive load studies are based on ML algorithms, which follow signal filtration and hand-crafted features. It is observed that hybrid DL architectures that integrate convolution and LSTM operators performed significantly better in comparison with other models. However, DL models especially hybrid models have not been extensively investigated for the classification of cognitive load captured by fNIRS devices. The current trends and challenges are highlighted to provide directions for the development of DL models pertaining to fNIRS research.

1. Introduction

Cognitive load theory (CLT) has been considered one of the most important learning theories in the field of experimental psychology (Kirschner, Ayres, & Chandler, 2011), educational psychology (Sweller, 2016), developmental psychology (Sepp, Howard, Tindall-Ford, Agostinho, & Paas, 2019), and medical education (Skulmowski & Xu, 2021). CLT signifies that the capacity of the human mind is limited when dealing with novel information (Castro-Alonso, de Koning, Fiorella, & Paas, 2021; Curum & Khedo, 2021). The theory leverages instructional implications and learning procedures of human cognitive structure. Generally, cognitive architecture assumes that all the novel information is initially processed by human's working memory, which has limited capacity and duration. The information is then stored in unlimited long-term memory. However, our working memory is limited when the information is retrieved from the previously organized long-term memory (Buchner, Buntins, & Kerres, 2021). The extent to which mental workload degrades performance depends on the experience of a person

* Corresponding author at: Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong Warun Ponds Campus, Victoria 3216, Australia

E-mail addresses: mehshan.khan@deakin.edu.au (M.A. Khan), houshyar.asadi@deakin.edu.au (H. Asadi), Li.Zhang@rhul.ac.uk (L. Zhang), Mqazani@su.edu.om (M.R.C. Qazani), s.oladazimi@deakin.edu.au (S. Oladazimi), ckloo.um@um.edu.my (C.K. Loo), chee.lim@deakin.edu.au (C.P. Lim), snahavandi@swin.edu.au (S. Nahavandi).

https://doi.org/10.1016/j.eswa.2024.123717
Received 19 July 2023; Received in revised form 9 February 2024; Accepted 17 March 2024
Available online 22 March 2024

working in a particular domain. Increase in cognitive load compromises the performance through decline in motivation and increase in reaction time, fatigue, and error rates. Modern research on behavior sciences emphasizes that the influence of cognitive load must be considered during teaching and learning so that effective knowledge acquisition can take place (Heitmann, Grund, Fries, Berthold, & Roelle, 2022; Tugtekin & Odabasi, 2022).

Measure of cognitive load plays a vital role in enhancing the skill set of a variety of tasks, e.g. in aviation (Wilson, Nair, Scielzo, & Larson, 2021; R. Zhu, Wang, Ma, & You, 2022), semi-autonomous cars (H. Zhang, Zhang, Xiao, & Wu, 2022), defense training (Buckley et al., 2022), aerospace (Magnusdottir, Johannsdottir, Majumdar, & Gudnason, 2022), e-learning (R. Liu, Wang, Koszalka, & Wan, 2022), virtual reality-based trainer (Zhao et al., 2022), and assembly operations (Fournier et al., 2022). In the last few decades, several non-invasive modalities have been exploited to measure cognitive load by acquiring signals from the human body. Changes in cognitive load can be detected via various physiological parameters, e.g. Electroencephalogram (EEG) (Farkish, Bosaghzadeh, Amiri, & Ebrahimpour, 2022), Electrocardiogram (ECG) (Lagomarsino, Lorenzini, De Momi, & Ajoudani, 2022), eye tracking (Yan et al., 2022), Functional Near-InfraRed Spectroscopy (fNIRS) (Agbangla, Audiffren, Pylouster, & Albinet, 2022), skin conductance level (Saha, Jindal, Shakti, Tewary, & Sardana, 2022), and Positron Emission Tomography (PET) (Canário, Jorge, Martins, Santana, & Castelo-Branco, 2022). Each physiological parameter is responsible for observing different biological processes. However, bulkiness, high cost and sensitivity to different disturbances limit the capability of these devices in ubiquitous computing. As an example, while eye tracking is widely used and is unobtrusive, it only provides indirect measure of the brain activity (Anderson et al., 2011). Neuroimaging studies related to fMRI and PET have generated insights into the pathological changes in blood oxygenation and metabolic functions (Catana, Drzezga, Heiss, & Rosen, 2012). Besides being expensive, fMRI and PET require a subject to be immobilized in a tightly restrained environment (Fujikawa et al., 2022; Harauzov, Ivanova, Vasiliev, & Podvigina, 2022). In addition, both modalities expose the subject to hazardous materials and loud noise. Electrodes of EEG are prone to internal and external artifacts, such as heartbeat, movement, and other electromagnetic interference. These disturbances make it challenging to differentiate signals from noise (H. Wang, Guo, Zhang, Gao, & Zheng, 2022). Skin temperature, eye tracking and skin conductance level are also widely used as non-intrusive measures of workload; but the findings suggest insignificant correlation between sensor data and subjective workload measure (Cosme et al., 2022; Žagar et al., 2022).

fNIRS has the potential to overcome the above-mentioned issues, and is useful and usable in a wide range of applications (Klein, Debener, Witt, & Kranczioch, 2022). Being known to be powerful and non-invasive, fNIRS functions as a safe tool to investigate hemodynamic responses in superficial cortical regions. fNIRS uses an optical fiber-based light source to emit infrared between a spectral window of 600 to 1000 nm and detectors to detect optical density changes (Li et al., 2022). Changes in neural activities result in changes in blood oxygenation levels. Based on the principles of the modified Beer-Lambert Law (Baker et al., 2014), fNIRS measures cognitive load by monitoring concentration variation in oxygenated hemoglobin (HbO2) and deoxygenated hemoglobin (dHb) at the cortical microcirculations blood levels, as shown in Fig. 1. The main advantages of fNIRS include high spatial resolution, safety, movement tolerability, portability, and ability for integration with EEG, PET, or ECG (Krampe, 2022; Y. Liu et al., 2022).

Although fMRI provides high-resolution and in-depth information on the blood oxygenation levels, inexpensive fNIRS targets the cortical regions of interest. fNIRS is also tolerant of motion artefacts, which makes it a better candidate for detecting brain activities in cognitive load-related tasks (Zhuang et al., 2022). For these reasons, we focus only on fNIRS-based data collection campaigns that capture the hemodynamics changes in the prefrontal cortex using off-the-shelf equipment in our review. fNIRS signals are naturally complex, non-linear, and have a high dimension. This data format makes it difficult to identify abnormalities with our naked eyes. These properties have made fNIRS data suitable for analysis using Deep Learning (DL) and Machine Learning (ML) models.

DL/ML models have an ability to learn features hierarchically by complex mapping functions directly from data. They are the leading Artificial Intelligence (AI) tools in several domains, such as image processing (Suganyadevi, Seethalakshmi, & Balasamy, 2022), pattern recognition (Bai et al., 2021), image segmentation (Picon et al., 2022), speech analysis (Bhangale & Kothandaraman, 2022) and physiological data processing (Patlar Akbulut, 2022). Signals recorded from fNIRS devices usually contain mixed artifacts and noise. Traditional approaches require the decomposition of fNIRS signals to frequency or wavelet transformation for noise removal. DL models, specifically Artificial Neural Networks (ANNs) or Convolutional Neural Networks (CNNs), sometimes require minimum pre-processing effort by generating machine learned features for classification and pattern recognition (Wani et al., 2022). Success of AI across various engineering fields promises the development of model-free approaches with robust performance. We, therefore, focus on the implementation, validation, and development of wearable fNIRS sensors for logging and tracking of cognitive load during memory demanding tasks in this review.

Although several reviews on the assessment of cognitive load using physiological sensors exist, to the best our knowledge, there is no research paper that cover in-depth applications of DL/ML models for analyzing fNIRS-based cognitive load. Previous survey and review articles within the research domain of cognitive load and physiological signals are thoroughly discussed in Section 2 of this review. These papers have predominantly focused on conventional ML, DL, and statistical techniques, placing particular emphasis on handcrafted feature

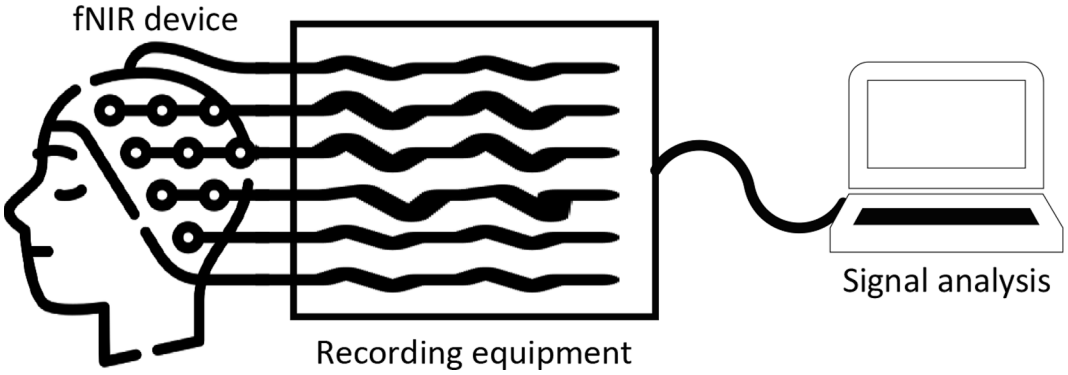

Fig. 1. Multi-channel data acquisition for generating cortical activation maps.

engineering methods for the analysis of fNIRS data. The focus of these articles has been on applications related to neurological disorders, stress, and emotional responses utilizing fNIRS technology. It is noteworthy that different cognitive tasks elicit specific cortical activations in various brain regions, necessitating customized hyperparameters for ML and DL algorithms tailored to each specific task. While existing reviews cover a broad spectrum of AI applications in fNIRS data analysis, there remains a challenge in highlighting and comprehending advancements specifically in ML and DL techniques for analyzing cognitive load data obtained from fNIRS measurements. Recognizing this research gap, we survey cognitive load and fNIRS AI literature with the explicit goal of highlighting progress made in employing ML/DL methods for cognitive load recognition.

The main contributions of this review are as follows:

- A thorough coverage of recent publications of fNIRS from 2011 to 2023 pertaining to classifying cognitive load by using state-of-the art DL and ML methods;
- Fundamental concepts of DL and ML pipelines including design and training of existing models for analysis of fNIRS signals;
- A concise summary of all reviewed publications, along with suggestions for the future development of ML and DL models for inference of cognitive load.

The remaining part of this article is organized as follows. Section 2 provides a review of studies on cognitive load. In Section 3, we outline our literature search strategy, detailing inclusion and exclusion criteria. Methods for inducing cognitive load are summarized in Section 3. Sections 4, 5, and 6 delve into the basics of ML/DL methodologies. Discussion of the reviewed articles is presented in Section 7, while Section 8 discusses future implications and challenges. Finally, Section 9 concludes with closing remarks.

2. Related work

Over the past few years, numerous researchers have undertaken reviews and surveys on cognitive load, with the aim of understanding current trends in monitoring cognitive load. The findings of these reviews have highlighted the complex nature of cognitive load assessment, revealing that it can be evaluated through various means, including both subjective and physiological measures. While subjective measures, such as questionnaires, have traditionally been a common means for gathering insights into cognitive load, metanalyses conducted by R. A. Block et al. (Block, Hancock, & Zakay, 2010) have indicated certain limitations associated with this approach. Their analysis, encompassing data from 117 experiments, revealed that relying solely on subjective measures can introduce biases and be influenced by individual differences in cognitive ability. Most reviews within this field consistently highlight the significance of employing physiological measures to gain valuable insights into cognitive performance during task execution. These measures include, but are not limited to, ECG, EEG, eye tracking, fNIRS and skin conductance level. These measures provide a direct and objective means of assessing the intricate aspects of cognitive function associated with task performance.

The development of deep learning techniques had a significant impact on the direction of neurology research. The current popularity of deep architectures brings the need to review and analyze existing studies about deep learning in physiological signals domain. Several studies have been conducted to discuss and investigate the role of DL models in analyzing physiological data. For instance, Y. Roy et al. (Roy et al., 2019) emphasizes the role of EEG in clinical applications such as sleep disorder diagnosis, epilepsy monitoring, and brain–computer interfacing. They highlight the increasing adoption of DL to address challenges like automating time-consuming tasks and improving generalization across subjects. The review identifies major trends, including DL's prevalence in EEG classification for various domains. Notably, studies varied widely in data quantity, architecture choices, and the use of raw EEG data. The review suggests a need for targeted investigations into optimal data amounts for DL in EEG processing. Recommendations are provided to enhance result reproducibility, including clear architecture and data descriptions, use of existing datasets, and code sharing. E. Banuelos-Lozoya et al. (Banuelos-Lozoya, Gonzalez-Serna, Gonzalez-Franco, Fragoso-Diaz, & Castro-Sanchez, 2021) highlights the research in the context of Quality of Experience/User Experience (QoE/UX) evaluation, focusing on recognizing cognitive states from various physiological data sources. The study found that while cognitive states such as mental workload, stress, and attention have been analyzed, there is still a need to understand their relationship with specific elements that contribute to the overall user experience. The main findings emphasized the general physiological and behavioral responses to stimuli rather than individual components of interfaces or interactions. Y. Zhou et al. (Y. Zhou et al., 2021) provides a comprehensive review of EEG-based cognitive workload recognition using machine learning. The article covers the steps of classical machine learning, including data acquisition, preprocessing, feature extraction and selection, classification, and evaluation. Additionally, it explores widely used deep learning models for workload recognition. Adil et al. (Saleem et al., 2023) review centers around driver drowsiness detection and emphasizes the complexity of driving, where reduced cognitive performance due to drowsiness can lead to accidents. The study reviews recent techniques and technologies for detecting driver drowsiness, emphasizing the use of physiological signals, particularly EEG and ECG sensors, along with GSR and thermal cameras. This review identifies challenges such as the lack of customized deep learning architectures, limited multimodal approaches due to complexity and real-time constraints, and difficulties in comparing performance across heterogeneous hardware sensors. The authors suggest the need for novel solutions, including IoT and mobile devices, non-invasive sensors, transfer learning, and customized deep learning architectures to enhance the robustness, reliability, resilience, and real-time capabilities of driver drowsiness detection systems.

Similarly, numerous other reviews and surveys on DL/ML, which focus on specific fields or applications. These encompass in-depth explorations of deep learning methodologies applied to various domains, such as eye-tracking, ECG, EEG and fNIRS and specific tasks like stress, emotion recognition, sleep disorders, cognitive load, anemia and multimedia learning. These comprehensive review papers have primarily focused on the diverse applications of ML/DL in analyzing various physiological signals. Despite the wealth of literature exploring the application of ML/DL techniques for cognitive load analysis using physiological measures, a notable gap exists in the systematic examination of the use of these techniques specifically for fNIRS signals. To the best of our knowledge, we did not find any in-depth literature review comprehensively covering the application of ML/DL techniques in the context of cognitive load analysis using fNIRS signals. While existing reviews delve into the applications of ML/DL for cognitive load assessment using EEG and other physiological signals, there is a lack of literature addressing the unique characteristics and challenges posed by fNIRS signals in this domain. It is worth mentioning that a review conducted by C. Eastmond et al. (Eastmond, Subedi, De, & Intes, 2022) has provided a broader analysis of the progress made in the application of DL techniques for analyzing fNIRS signals. However, this study did not explore the specific intricacies related to cognitive load assessment using fNIRS. Secondly, it is noteworthy that these reviews have examined studies that analyze physiological signals by either utilizing publicly available datasets or repurposing data from prior studies. However, these reviews do not bring attention to the possible challenges and issues linked to the initial data collection processes utilized for subsequent ML and DL analyses. Therefore, to address the existing gap in the literature, our review aims to highlight the significant advancements made in the application of ML and DL methodologies for the recognition of cognitive workload using fNIRS signals. This involves an examination of all studies

published within this specific domain, by providing information on the development of techniques, methodologies, and findings.

3. Materials and methods

This review covers studies on cerebral activities during cognitive demanding tasks according to guidelines provided by the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) (Page et al., 2021) protocol. We formulate a comprehensible search strategy with the aim to answer specific research questions. To maintain the focus on neuroergonomics studies related to ML and DL, we first identify the keywords for our preliminary search. Therefore, we included the common terms of cognitive load together with fNIRS in the final search string presented in Table 1.

We particularly limit the publications to those in 2011 onward in well-established sources, namely the ACM digital library, Web of Science, PubMed, IEEE Explore, Scopus, PubMed, Google Scholar, and EuropePMC. We use the search keywords in these electronic databases, and then initially titles and abstracts have been screened based on the following inclusion and exclusion criteria.

3.1. Inclusion criteria

The objective of this review is to explore ML/DL-based techniques to decode brain activities from fNIRS signals. The studies included in this article should meet the following criteria:

- Cognitive demanding tasks;
- AI-based techniques used for analysis.
- AI models trained with fNIRS signals;
- Data sets consider healthy subjects so that the true potential of fNIRS signals in developing applications related to human performance, load management and training purposes can be explored.
- Articles published between 1st January 2011 and 31st December 2023 in peer-reviewed articles and highly-cited conference proceedings;

3.2. Exclusion criteria

The following criteria have been considered to determine whether an article needs to be excluded:

- Articles that do not include sufficient details to gauge the research quality or that appear only in an abstract form.
- Dissertations, case studies, thesis, pre-prints, overviews, and book chapters;
- Studies on patients;
- Systematic studies on publicly available data sets.
- Research based on statistical analysis of data.
- Research in languages other than English.

Table 1
Search strings used for each topic.

Topic	Search terms
Cognitive load	"cognitive load" OR "dual task" OR "cogniti*" OR "working memory" OR "attention" OR "load" OR "mental load" OR "overload" OR "mental effort" OR "germane load" OR "germane" OR "intrinsic load" OR "intrinsic cognitive load" OR "extraneous cognitive load"
Artificial intelligence	"deep learning" OR "machine learning" OR "artificial intelligence"
Functional Near Infrared Spectroscopy	"fNIRS" OR "functional near infrared spectroscopy"

3.3. Search results

A selection process has been conducted in two main steps. The first step involves the removal of all duplicates; while the second applies the inclusion and exclusion criteria specified earlier. Articles that have no information on feature analysis, comparisons, study designs and outcomes have also been excluded. Fig. 2 summarizes the precise steps involved in the identification, screening, and eligibility processes.

A total of 1428 studies have been retrieved according to the keyword search, and almost 280 duplicates studies have been removed. Then, a total of 410 studies that meet the exclusion criteria have been deleted, while studies that meet the full inclusion criteria, information regarding cognitive tasks, model designs and outcomes have been extracted. Over 50 % of the articles included in this review have been published in the last three years. In addition, the major results of all 45 articles on cognitive load with fNIRS and ML/DL are summarized in subsequent sections.

4. Cognitive activity capture with fNIRS

A cognitive activity indicates an evaluation of a task based on the performance outcome. Although the main aim of our research is to investigate the physiological measures of cognitive load, researchers have used subjective measure for its analysis. Subjective measure requires the participants to rate different aspects of the learning process using a multi-item scale. Particularly, NASA's Task Load Index (TLX) (Hart & Staveland, 1988) is considered as a gold standard to measure workload in human-system evaluation. The NASA-TLX measure calculates a global index score based on mental demand, physical demand, temporal demand, performance, frustration, and effort. These scores are converted in the range of 0–100 (Nasirizad Moghadam et al., 2021) for task evaluation purposes. However, when many cognitive processes interact one another, learners may not be able to identify different forms of cognitive load. The usefulness of subjective measures has been questioned due to a lack of correspondence and assessment of events in the external world in correlation with the simulated cognitive environment. Therefore, it is important to improve the credibility of subjective measures so that the external world as well as internal sensation and feeling can be correlated in cognitive load measurement.

In contrast, physiological measures especially fNIRS provide uninterrupted evaluation, offering a more objective workload assessment. fNIRS-based systems have been widely used to study neural changes in simulated cognitive environments. Concentration changes in Hbo2 and dHb are proportional to the change in the cerebral blood volume, providing a useful measure of neural activities. Some studies have implemented subjective surveys and categorised fNIRS signals using DL and ML classification techniques (Asgher et al., 2020; Keles, Cengiz, Demiral, Ozmen, & Omurtag, 2021). The main reason to rely only on physiological signals is that surveys interrupt the underlying operation flow, lengthen the time of operations, and are only available post-task (T. Zhou et al., 2020), leading to intra and inter-subject variability, inconsistency, disruption and inadequacy pertaining to measurements for the scenarios discussed in this article.

The foremost step in developing an fNIRS-based system is the selection of brain regions from where the brain signals are generated. The signals are generally acquired from the pre-frontal cortex or motor cortices. Motor cortices mostly respond to the movement of body parts, e.g., legs, arms, fingers, hands, etc., In comparison, most of the included studies in this survey indicate that signals from the pre-frontal cortex are highly correlated with cognitive tasks. In addition, the signals acquired from the pre-frontal cortex are less sensitive to motion artifacts and high-frequency influence (Gemignani & Gervain, 2021). Fig. 3 depicts the distribution of studies in this review based on cognitive tasks. Cognitively demanding activities contribute to changes in Hbo2 and dHB over the pre-frontal region of the brain can be categorized into four groups: mental arithmetic (16 %), n-back task (24 %), Stroop task (5 %)

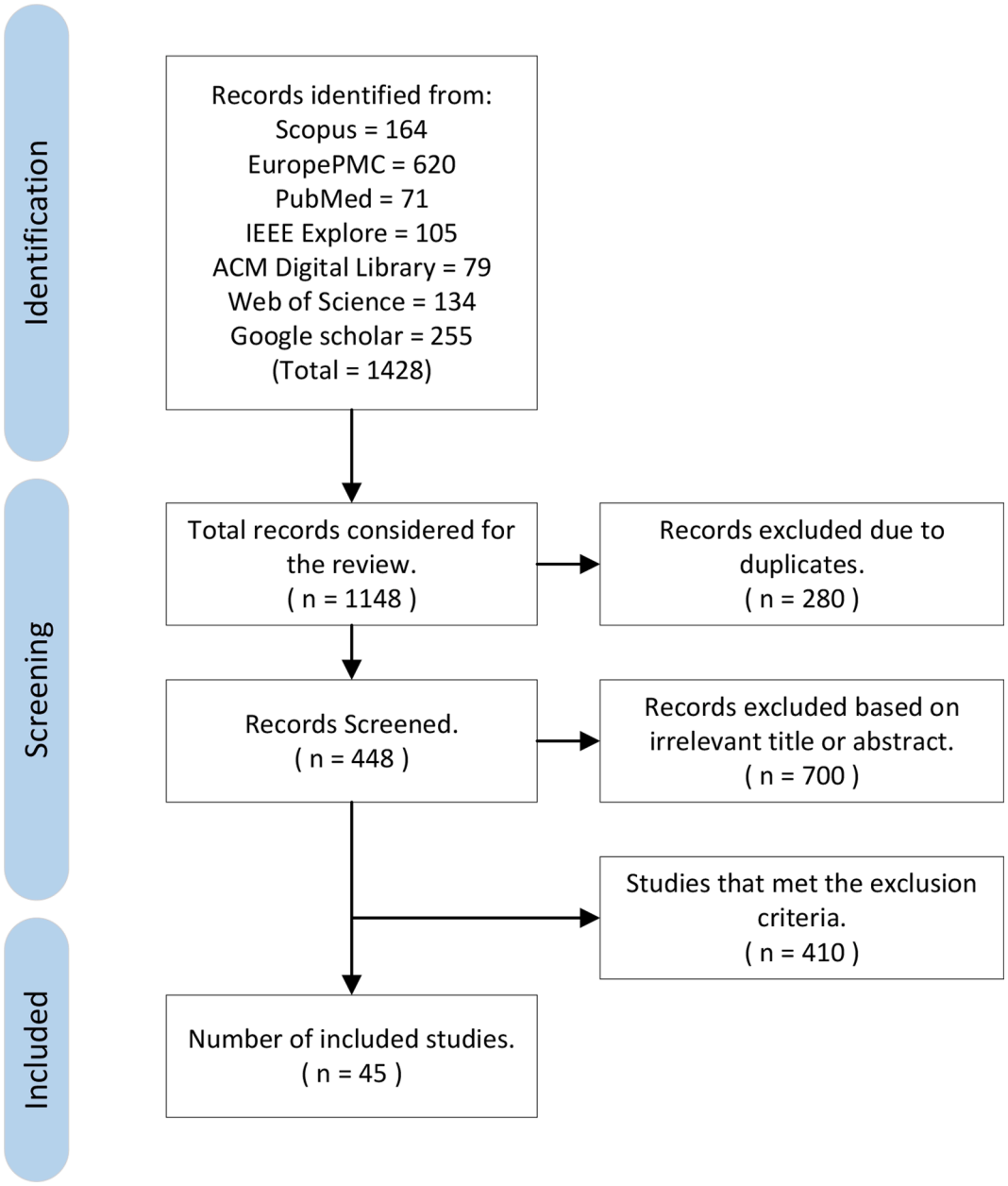

Fig. 2. A flow diagram of the literature search according to the PRISMA (Page et al., 2021) guidelines.

and simulation-based tasks (55 %). A description of the general protocols for these tasks is as follows:

4.1. Mental arithmetic

Arithmetic tasks involve performing mathematical calculations without the help of using a paper, calculator, or computer. Arithmetic tasks usually consist of presenting a sequence of numbers to participants for performing addition, subtraction, multiplication, or division in a predefined duration. Mathematical equations of different complexity levels require simultaneous mental processing and information storage, which induces both low and high levels of mental workload in addressing complex experimental scenarios.

4.2. N-back Tasks

Introduced by Kirchner (Kirchner, 1958) in 1958, n-back tasks have been most extensively used in neuroscience to understand the neural basis of working memory. As visual-spatial tasks, researchers in neuroimaging have leveraged n-back tasks to induce different levels of memory load. It serves as a visual or auditory stimulus to participants with a series of several random numbers, pictures, or digits. Participants need to remember them and then, when enquired, need to determine the matches with stimuli of N items seen before. Cognitive load can be modified by varying the value of N ($N = 0, 1, 2, \cdots, n$). In the 0-back task participants are required to identify single pre-specified digit, letter, or image. In the 1-back task, each new item is identical to the one preceding it. Similarly, for a 2-back, 3-back, …, or n-back task, each new item is identical to item presented 2, 3, …, or n trails back. Fig. 4 shows a schematic of 1-, 2-, and 3-back tasks. Varying the value of N systematically increases the processing load, which results in changes in reaction time and accuracy (Lamichhane, Westbrook, Cole, & Braver, 2020).

4.3. Stroop task

A Stroop task (Stroop, 1935) was developed in 1935 to study the

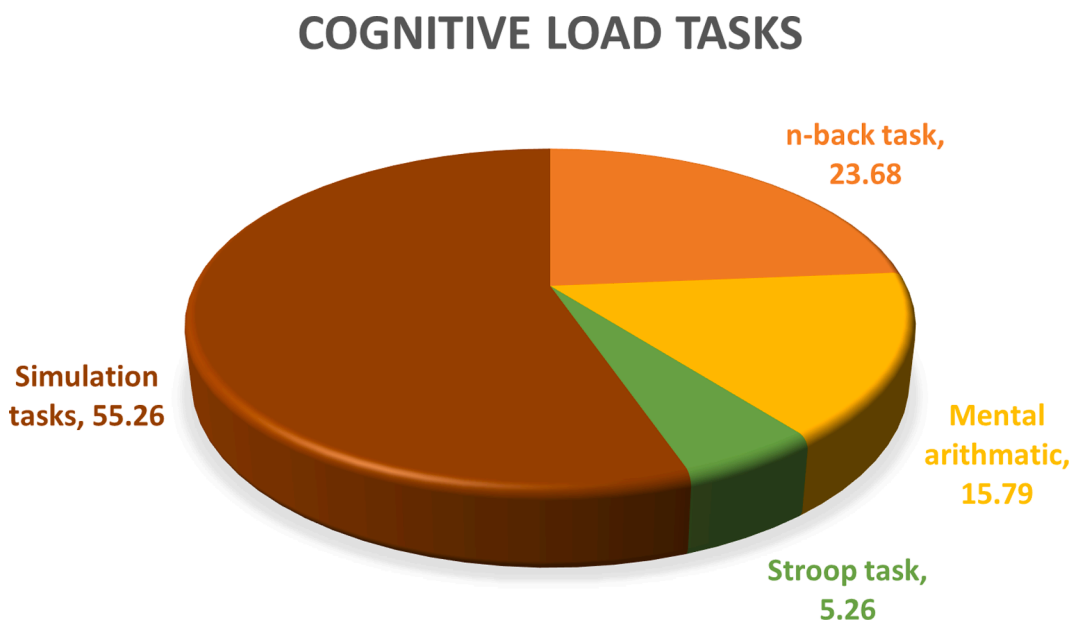

Fig. 3. Task based distribution of studies.

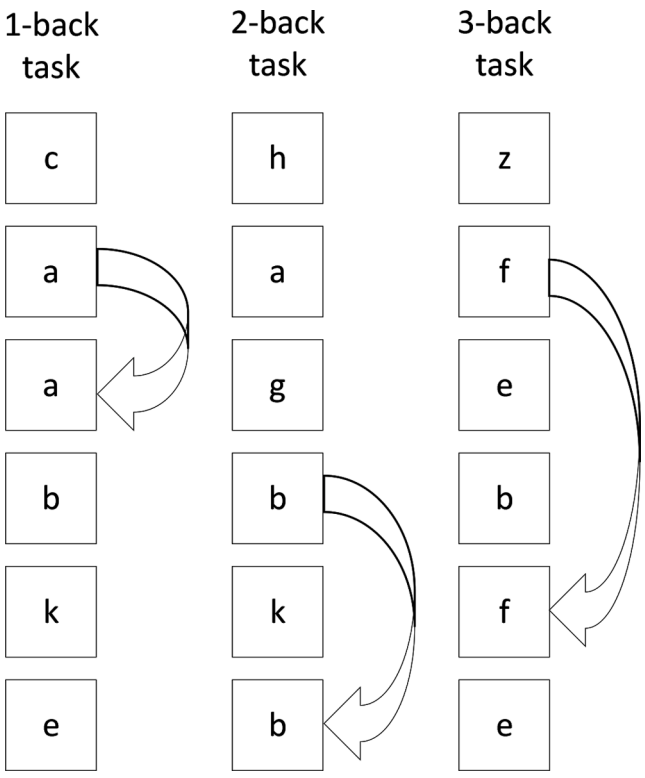

Fig. 4. Schematic of 1, 2 and 3 back tasks.

effect of cognitive inhibition. Since then, many variants of Stroop task have been proposed. Some of them have been used in clinical neuropsychology to study neurological disorders of patients (Fischer-Jbali, Montoro, Montoya, Halder, & Duschek, 2022; Lewis, Garcia, Price, Schweizer, & Nixon, 2022). A traditional Stroop task, as shown in Fig. 5, entails the presentation of four-colour words displayed in red, green, blue, and yellow. As an example, the word green could be displayed in green, yellow, red, or blue. The Stroop effect has been extensively used in neurological studies with an opportunity to earn reward points for accurate and fast responses. In the Stroop test, participants are instructed to identify the font color while ignoring the word. This results in a delayed identification of colors, a slower response time and an increased cognitive workload.

4.4. Cognitive load simulator-based studies

Studies of human brain in a simulator-based environment offer the safest way to expose participants to simulated dangers without risking life or losing property (Frederiksen et al., 2020). Technologies such as driving/flying simulators (Asadi et al., 2023; Asadi et al., 2019), virtual reality (Kooijman, Asadi, Mohamed, & Nahavandi, 2022, 2023), and cognitively demanding games can be used to create simulation where the surroundings from a real environment are integrated into a virtual system. These simulations, as shown in Fig. 6, have a high level of connectivity with different types of commercial joystick or customized controllers. Furthermore, distractions during simulation such as visibility, turbulence, mental state, or pre-programmable handling qualities add cognitive load to participants. In simulator-based studies, a flying/driving task constitutes the majority of neuro-ergonomics application (e.g., aircraft control systems, driving a car, or flying a plane in complex simulated scenarios) (Mejia-Puig & Chandrasekera, 2022; Reddy et al., 2022). Human attention is then monitored and assessed pertaining to complex cognitive tasks (e.g., surgery simulation, video lectures, identification of hazards in lab environment). Nonetheless, unrealistic scenarios that cannot be easily replicated present a detrimental impact on the cognitive and performance outcomes.

Fig. 5. Classical Stroop test.

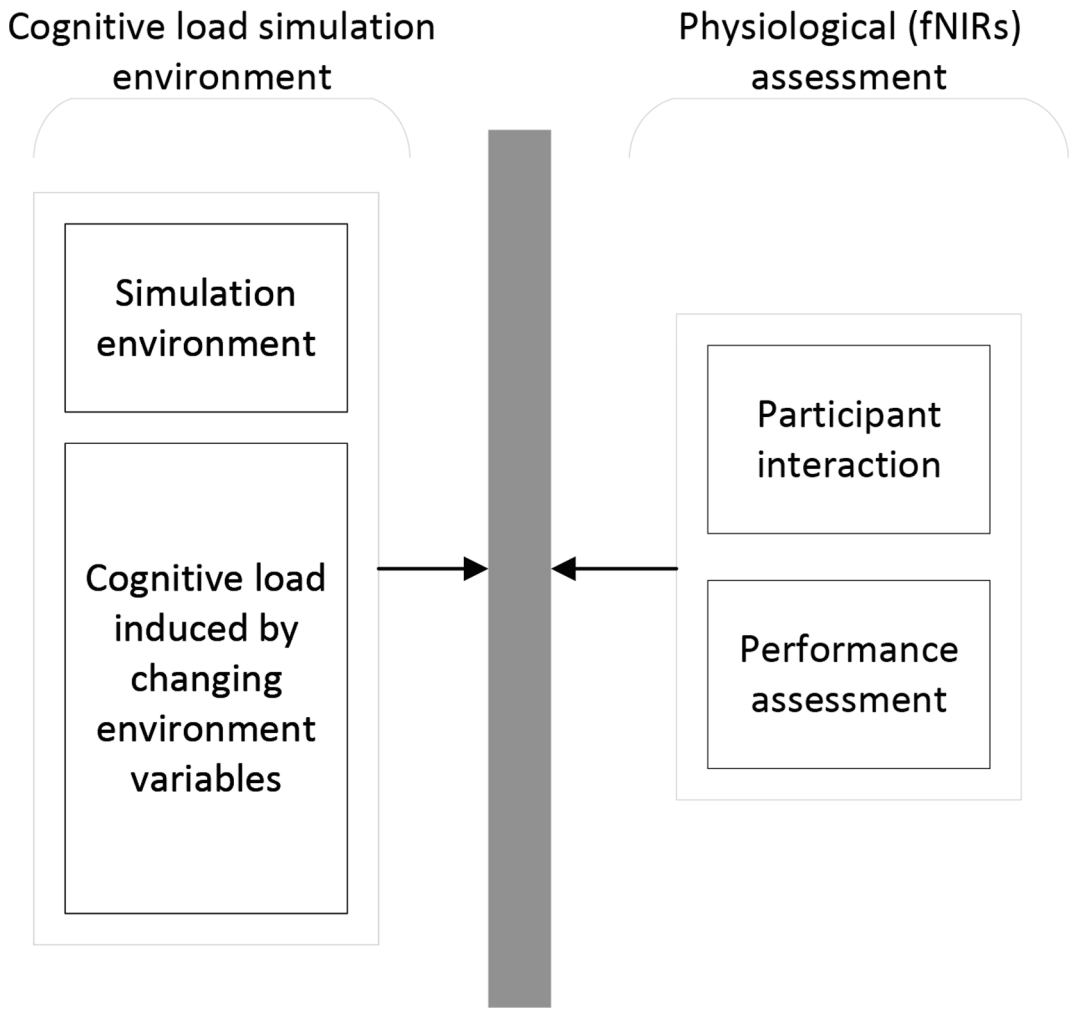

Fig. 6. Cognitive load simulation environment.

5. Artificial intelligence in fNIRS analysis

AI, which includes ML and DL, leverages computational algorithms with learning capabilities to recognize patterns from data. Sometimes, it is difficult to interpret exact information from data samples (Mehta & Shukla, 2022). In this respect, DL and ML offer the underlying algorithms to learn from data without being specifically programmed to do so. AI-based models suffer from the requirement of a lengthy computation time and the problem of vanishing gradients (Khademi, Ebrahimi, & Kordy, 2022), causing researchers to use statistical and other methods for data analysis. However, the recent advancements in AI and the availability of graphic processing units (GPUs) enable neuroscientists to decode and classify fNIRS signals with unprecedented details.

In neuroimaging, ML/DL models takes fNIRS signals as training data

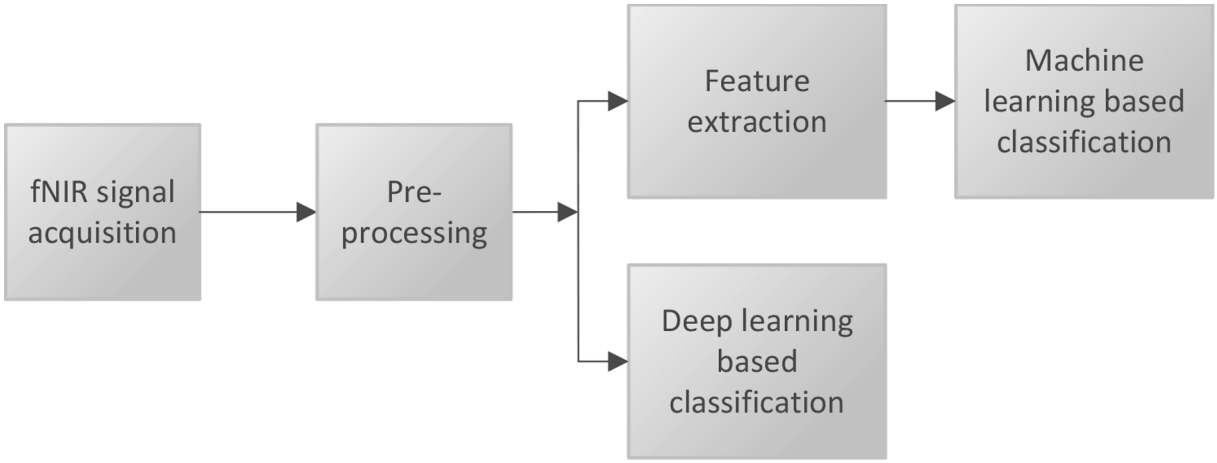

Fig. 7. The overall steps of fNIRS analysis using ML and DL include signal acquisition, pre-processing, feature extraction and classification.

to learn and predict the associated class labels. In the training phase, the ML/DL algorithm optimally configures the hyper-parameters in a way that the trained model can be generalized to produce the desired outcome when it is presented with an unseen data sample. Fig. 7 depicts a general flow of DL/ML model implementation. In the first step, the raw fNIRS signals are captured. These signals typically contain noise caused by changes in the heart rate, blood pressure, etc. In the pre-processing phase, the signal artifacts and other outliers in the data set are removed. Most of the studies presented in this review adopt bandpass and Butterworth filters as well as other methods for this purpose. The input spectrum and its correspondence are determined during an optional feature extraction process. Features selection improves the classification performance by reducing the data dimension and computational complexity. It is generally used with ML algorithms, and sometimes with DL algorithms, to increase robustness of the model. There have been few papers that use feature extraction along with DL algorithms, but most of the studies apply raw fNIRS signals as the model input. Most of the studies reported in the literature used the summary based statistical features (e.g., mean, variance, maxima, minima, slope, skewness, kurtosis, and normalization) or parameterization techniques (e.g., Wigner-Ville Distribution, continuous wavelet transform and Hough transform) to extract useful features from the data. A well-trained model can provide predictions pertaining to different levels of mental workload. To further enhance the model generalizability capability, most of the studies either utilise the n-fold Cross Validation (CV) or leave-one-out method.

Classifications tasks vary widely, and they can be categorized into three main groups: (a) supervised learning (b) unsupervised learning, and (c) reinforcement learning. In supervised learning, labels (target outputs) are generally determined by humans, and a supervised algorithm maps the input features to a desired output (label). Supervised learning algorithms need external assistance in the form of handcrafted labeled data for training and test phases. So, the algorithm learns pattern from training data and validate the model on test data for classification and prediction purposes. Classification methods, e.g. CNNs (Albawi, Mohammed, & Al-Zawi, 2017), ANNs (Abiodun et al., 2018), Support Vector Machine (SVM) (Vapnik, 1999), Decision trees (Kotsiantis, 2013), Random forests (Breiman, 2001), Naive Bayes (Fix & Hodges, 1951), Logistic Regression (DeMaris, 1995) and Linear Regression (Su, Yan, & Tsai, 2012) are common supervised learning algorithms. In contrast, unsupervised learning algorithms use unlabeled data for inference. These algorithms learn features from the raw data and develop a predictive model to categorise the input data into different clusters with dimensionality reduction. Example of unsupervised learning algorithms are K-Means clustering (Hartigan & Wong, 1979), Principal Component Analysis (PCA) (Maćkiewicz & Ratajczak, 1993), and Independent Component Analysis (ICA) (Stone, 2002). Reinforcement learning (RL) works on the principle of sequential decision making. It uses learning agents to interact with a dynamic environment, which maximizes the rewards when a task is successfully achieved. The main factors contribute to RL are the environment model, policy, raw signals, and reward function. Traditional RL models can only solve the problems having a low dimensional space. However, the recent introduction of deep neural networks (DNNs) in terms of reinforcement agents gives the model ability to learn from multi-dimensional inputs (Ibarz et al., 2021). Over time, more and more DNNs combined with RL give the power to solve problems in a high dimensional space, leading to various new RL research domains such as robotics (Bhagat, Banerjee, Ho Tse, & Ren, 2019) and autonomous driving (Kiran et al., 2021). Among all the three learning methods, supervised learning is mainly used to predict and classify cognitive load pertaining to fNIRS signals.

6. Bibliometric examination of ML and DL models applied to fNIRS data

This section discusses the trends in the formulation of ML/DL models performed on fNIRS data. A comprehensive summary of DL design, architecture and experimental paradigm is presented in Table 2.

Most articles in neuroscience employ fNIRS data sets that are not publicly available. The performance measure, e.g. a simple accuracy measure or other metrices such as the mean squared error (MSE), root mean squared error (RMSE), F1-score, true positives, or false positive, cannot be generalised since each study has different test subjects, data procurement protocols and different cognitively demanding tasks. Studies on fNIRS indices in mental workload can be categorized into three categories, as illustrated in Fig. 8: (1) ML-based fNIRS analysis; (2) DL-based fNIRS analysis; and (3) hybrid AI-based models for fNIRS analysis.

6.1. Machine learning trends in fNIRS analysis

ML, which is a subset of AI, is capable of processing patient data and imitating the ability of humans in recognizing patterns. This section covers ML approaches to analyze fNIRS data. A total of 25 studies are reviewed, which apply ML to objectively evaluate the mental workload. A summary of ML based algorithms in the literature are as follows. Fig. 9 displays the distribution of studies utilizing ML classifiers for the analysis of fNIRS data. SVM emerge as a prominent choice within the fNIRS research community, followed by Random Forests, which are recognized for their efficacy in handling high-dimensional data. LDA and k-NN are also noted in the distribution as applied methods in fNIRS-based machine learning studies.

6.1.1. Support vector machines (SVM)

According to our investigation, SVM has been widely used in fNIRS signal analysis because of its ease of implementation and high accuracy. The idea of SVM is based on the structural minimization principle. It is mainly used for pattern recognition and regression analysis. While classifying the data samples in high dimensional classification space, it tries to find the optimal hyperplane with highest margin between classes. These hyperplanes are trained with algorithms so that different categories of input data points are separated. Several researchers like Gateau et al. (Gateau et al., 2015), Asgher et al. (Asgher et al., 2019), Keles et al. (Keles et al., 2021), Derosiere et al. (Derosiere et al., 2014), Dong et al. (Dong & Jeong, 2018) Abibullaev et al. (Abibullaev & An, 2012), and Kurihara et al. (Kurihara et al., 2020) used the SVM to classify mental workload from the fNIRS signals.

Khanam et al. (Khanam et al., 2022) applied the ANVOA test on the conventional mean, minimum, maximum, standard deviation (SD), slope, and skewness features from all 36 channels. The ANOVA analysis signified that only two channels in the frontal and motor area indicated statistical interference among different levels of workload. SVM was trained on the features obtained from two significant channels and achieved an accuracy rate of 71.48 %.

Zhu et al. (Q. Zhu et al., 2021) employed conventional feature extraction methods to explore the relationship between fNIRS signals and cognitive load based on a Sternberg experiment (Sternberg, 1969). Experimental results highlight the fact that the significant features for the prediction cognitive load using SVM varied across participants because each person processes information differently. So, instead of generalized models, personalized models are required to predict cognitive from fNIRS signals. A further pipeline to filter, clean and model fNIRS data has also been presented in this study.

To reduce the number of false positive, Lim et al. (Lim et al., 2020) introduced the feature extraction method named deep contribution ratio, which uses the k-means clustering method as well as Euclidean distance method to identify activated and non-activated channels. Experimental results showed that deep contribution ratio achieved better accuracy (80 %) in comparison with those obtained from conventional slope-based features (59.8 %).

Asgher et al. (Asgher et al., 2019) processed fNIRS data using a proposed Fixed-Value Modified Beer-Lambert law (FV-MBLL) and

Table 2
Description of data collected from the included studies.

Authors	Functional near-infrared spectroscopy (fNIRS) properties		fNIRS Features		Artificial intelligence-based approaches		Performance evaluation		Participants
	Brain area	**Environment**	**Feature methods**	**Derived features**	**Strategy**	**Architecture**	**Accuracy**	**Other metrics**	
(Gateau, Durantin, Lancelot, Scannella, & Dehais, 2015)	Prefrontal cortex and dorsolateral prefrontal cortex (DLPFC)	ISAE (French Aeronautical University in Toulouse, France) flight simulator	ANOVA	Mean, kurtosis and skewness	5-fold CV	SVM	62 %	Specificity = 58 % Sensitivity = 72 %	19 (13 males and 6 females)
(Oku & Sato, 2021)	Prefrontal cortex	Watching video lecture of astronomy for 27 min and answer 10 questions	N/A	Mean values of HbO2 and dHb	Leave one out	Random Forest and GLMNET	Random forest = 66 % GLMNET = 63 %	Random forest (sensitivity = 0.63 ± 0.066, specificity = 0.66 ± 0.0420, and Cohen's kappa coefficient = 0.26) GLMNET (sensitivity = 0.62 ± 0.067, specificity = 0.64 ± 0.042, and Cohen's kappa Coefficient = 0.22)	18 (8 males and 10 female)
(Kornev et al., 2022)	Left and right brain hemispheres	Iowa Gambling Task (IGT)	Pearson coefficient	Mean, variance and standard deviation	10-fold CV	Multiple regression, decision trees, ANN, SVM and random forest	Best accuracy is achieved by SVM with radial basis function	SVM RMSE = 3.37 to 7.84 SVM R-squared = 0.29 – 0.96	30 (5 males and 25 females)
(X. Zhou, Hu, Liao, & Zhang, 2021)	Prefrontal cortex	Civil engineering lab (identification of hazards)	Fisher criterion	Mean	10-fold CV	LDA	70 %	N/A	48 (35 males and 13 women)
(Lamb, Neumann, & Linder, 2022)	Prefrontal cortex	VR based questions about presented content	ANOVA	N/A	2-fold CV	Random Forest	83.9 %	Sensitivity = 0.73 ± 0.071 Specificity = 0.71 ± 0.044 Cohen's kappa coefficient = 0.41	40 (21 males and 19 females)
(Khalil, Asgher, & Ayaz, 2022)	Prefrontal cortex	n-back task	Shapiro–Wilk test	N/A	Leave one out and 10-fold CV	CNN based model	94.52 %	N/A	26
(Zaman & Islam, 2021)	Prefrontal cortex	n-back tasks	Wigner-Ville Distribution	N/A	N/A	ResNet50	98 %	N/A	10 (6 males and 4 females)
(Le, Xuan, & Aoki, 2022)	N/A	Driving in simulation-based environment	N/A	N/A	N/A	Random forests	98.24 %	PPV = 97.02 % TPR = 97.17 % TNR = 98.71 % F1-score = 97.10 NPV = 98.77 FPR = 1.29	17 (5 males and females)
(Asgher et al., 2019)	Prefrontal cortex	Mental arithmetic	N/A	Mean-Variance, Mean-Peak, Mean Slope, Peak Slope, Peak and Variance	10-fold CV	SVM	94 %	N/A	20 (10 males and 10 females)
(E. Q. Wu et al., 2021)	Medial prefrontal cortex, left and right	Physical flight simulator (cognitive states during simulated	Hough Transform features	N/A	5-fold CV	Scalable gamma non-negative matrix	92 %	N/A	40 pilots

(*continued on next page*)

Table 2 (*continued*)

Authors	Functional near-infrared spectroscopy (fNIRS) properties		fNIRS Features		Artificial intelligence-based approaches		Performance evaluation		Participants
	Brain area	**Environment**	**Feature methods**	**Derived features**	**Strategy**	**Architecture**	**Accuracy**	**Other metrics**	
	dorsolateral prefrontal cortex, left and right ventrolateral prefrontal cortex, and left and right temporal cortex	failure of the aircraft)				network (SGNMN)			
(Keles et al., 2021)	Prefrontal cortex	Laparoscopic trainer box (simulated surgery)	Wilcoxon signed-rank test	Mean, skewness and kurtosis	5-fold CV	SVM	90 %	N/A	11 surgeons and 17 medical students
(Kwon & Im, 2021)	Prefrontal cortex	Mental arithmetic and idle state tasks	N/A	N/A	Leave-one-subject out CV	CNN-based model	71.20 % ± 8.74 %	N/A	18 (10 males and 8 females)
(Derosiere, Dalhoumi, Perrey, Dray, & Ward, 2014)	Prefrontal cortex and the right parietal areas	Thumb abduction tasks	*t*-test	N/A	N/A	SVM	90 %	N/A	7 (male)
(Dong & Jeong, 2018)	Prefrontal cortex	Simple arithmetic (SA) and 1-back and 2-back tasks	Wilcoxon signed-rank test and PCA	N/A	Nested CV	SVM	77 %	N/A	22 (18 males and 7 females)
(Asgher et al., 2020)	Prefrontal cortex	Logic and arithmetic task with four difficulty levels	*t*-test	Normalization, signal mean, maxima, variance, minima, slope, variance, skewness, kurtosis and signal peak	10-fold CV	CNN and LSTM	CNN = 87.45 % LSTM = 89.3 %	N/A	7 (2 males and 5 females)
(Le, Aoki, Murase, & Ishida, 2018)	N/A	Real car different driving task along with digit recalling n-back task	PCA	N/A	5-fold CV	Random forests	96.08 %	N/A	5 (4 males and 1 female)
(Ho, Gwak, Park, & Song, 2019)	Prefrontal cortex	Stroop task experiment	PCA	N/A	N/A	SVM, Adaboost, Deep Belief Network and Convolution Neural Network	SVM = 64.74 % ± 1.57 % AdaBoost = 71.13 % ± 2.96 % DBN = 84.26 % ± 2.58 % CNN = 72.77 % ± 1.92 %	N/A	16 (8 males and 8 females)
(Abibullaev & An, 2012)	Frontal cortex	n-back task	Continuous wavelet transforms features	N/A	5-fold CV	BPNN, LDA and SVM	N/A	AUC BPNN = 0.7672 AUC SVM = 0.9404 AUC LDA = 0.8902	9 (8 males and 1 female)
(L. M. Wang et al., 2022)	Frontal cortex	Verbal fluency test	N/A	N/A	N/A	CNN (VGG-16 based)	100 %	TPR = 100 FNR = 100	13 (6 males and 7 females)
(Naseer, Qureshi, Noori, & Hong, 2016)	Prefrontal cortex	Mental arithmetic task vs rest signals	N/A	Mean, peak, slope, variance, kurtosis, and skewness and feature normalization between 0 and 1	10-fold CV	LDA, QDA, k-NN, Naive Bayes, SVM and ANN	LDA = 71.6 ± 1.1 % QDA = 90.1 ± 1.3 % k-NN = 69.8 ± 0.5 %	LDA (Precision = 72.8 ± 6.2, Recall = 73.5 ± 9.2) QDA (Precision = 90.0 ± 4.4, Recall 91.2 ±	7

(*continued on next page*)

Table 2 (*continued*)

Authors	Functional near-infrared spectroscopy (fNIRS) properties		fNIRS Features		Artificial intelligence-based approaches		Performance evaluation		Participants
	Brain area	**Environment**	**Feature methods**	**Derived features**	**Strategy**	**Architecture**	**Accuracy**	**Other metrics**	
							Naive Bayes = 89.8 ± 1.4 % SVM = 89.5 ± 1.0 % ANN = 91.4 ± 0.3 %	5.5) k-NN (Precision = 69.1 ± 1.3, Recall = 70.4 ± 2.6) Naive Bayes (Precision = 91.5 ± 5.1, Recall 88.5 ± 5.0) SVM (Precision = 89.1 ± 4.2, Recall = 91.8 ± 5.5) ANN (Precision = 90.1 ± 2.7, Recall = 91.5 ± 4.4)	
(J. Wang, Grant, Velipasalar, Geng, & Hirshfield, 2021)	Frontal cortex	n-back task	N/A	N/A	10-fold CV	CNN-BiGRU-SLA	77.53 %	Precision = 77.41 Recall = 77.65 F1-score = 77.42	22
(Khanam, Hossain, & Ahmad, 2022)	Frontal area, motor part, parietal area, and occipital area	n-back task	ANOVA	Mean, minimum, maximum, standard deviation, slope and skewness	N/A	SVM	73.40 ± 0.076 %	N/A	26 (9 males and 17 females)
(Q. Zhu, Shi, & Du, 2021)	Prefrontal cortex	Sternberg test	N/A	Mean, peak, standard deviation, kurtosis and skewness	10-fold CV	SVM (Gaussian radial basis function)	70.02 ± 4.41 %	N/A	15 (14 males and 1 female)
(R. Liu, Reimer, Song, Mehler, & Solovey, 2021)	Prefrontal cortex	Fixed-base, full-cab Volkswagen New Beetle, Verbal, and n-back task	ANOVA	Multilayer perceptron features	10-fold CV	ESN	80.61 %	Precision = 79.08 Recall = 81.67 F1-Score = 80.38	18
(Varandas, Lima, Bermúdez i Badia, Silva, & Gamboa, 2022)	Dorsolateral prefrontal cortex	Corsi-Block task	N/A	Maximum, minimum, polarity, mean, variance, Standard deviation, kurtosis and skewness	10-fold CV	Random Forest	70.91 ± 13.67 %.	Precision = 72.86 ± 15.32 Recall = 69.09 ± 14.77 F1-Score = 70.27 ± 14.30 AUC-ROC = 72.50 ± 17.26	10 (6 males and 4 females)
(Lim et al., 2020)	Prefrontal Cortex	n-back task	Deep contribution ratios	N/A	10-fold CV	SVM	80.6 %	Sensitivity = 78.1 % Specificity = 85.5 % AUC = 85.1 %	25 (21 males and 4 females)
(Saikia, Kuanar, Borthakur, Vinti, & Tendhar, 2021)	Prefrontal cortex	n-back task	N/A	Gradient value, mean, variance, number of peaks, kurtosis, skewness, maximum and minimum value	N/A	k-NN	75 %	N/A	12
(Berivanlou, Setarehdan, & Noubari, 2016)	Prefrontal cortex	n-back task	ANOVA	Mean, variance, skewness and kurtosis	10-fold CV	Linear regression	63.7 %	N/A	10 (6 males and 4 females)

(*continued on next page*)

Table 2 (*continued*)

Authors	Functional near-infrared spectroscopy (fNIRS) properties		fNIRS Features		Artificial intelligence-based approaches		Performance evaluation		Participants
	Brain area	**Environment**	**Feature methods**	**Derived features**	**Strategy**	**Architecture**	**Accuracy**	**Other metrics**	
(L. Wang et al., 2021)	N/A	n-back task	N/A	N/A	5-fold CV	CNN	71.63 %	N/A	27
(Saadati, Nelson, Curtin, Wang, & Ayaz, 2021)	N/A	n-back task	N/A	N/A	N/A	CNN and RNN based model	98.3 %	N/A	N/A
(Izzetoglu, Jiao, & Park, 2021)	Left and right hemispheres	Driving simulator	N/A	Normalization	N/A	Logistic regression	97.5 %	N/A	10 (4 males and 6 females)
(Lu, Yan, Chang, & Wang, 2020)	N/A	Mental arithmetic	N/A	N/A	N/A	LSTM-FCN	97.1 %	N/A	8
(Çakır, Vural, Koç, & Toktaş, 2016)	Prefrontal cortex	Thales Airbus 320 Simulator	N/A	Mean, standard deviation and slope	N/A	LDA	91 %	N/A	8
(Benerradi, A. Maior, Marinescu, Clos, & L. Wilson, 2019)	Prefrontal cortex	Customized task (Game based task target color balls using joystick to induced different levels of workload)	N/A	Normalization	N/A	Logistic regression, SVM and CNN	LR = 50.99 % SVM = 53.90 % CNN = 49.53 %	N/A	11 (6 males and 5 females)
(Ho, Gwak, Park, Khare, & Song, 2019)	Prefrontal cortex	Stroop tasks	N/A	N/A	N/A	DBN and CNN	DBN = 84.26 ± 9.10 % CNN = 65.42 ± 1.58 %	N/A	16 (8 males and 8 females)
(Kurihara et al., 2020)	Prefrontal lobe	Verbal memory retrieval and visuospatial memory retrieval	N/A	N/A	20-fold CV	k-NN and SVM	SVM = 100 k-NN = 100	Positive Predictive values (PPV) = 1 Negative Predictive Values (NPV) = 1	20 (13 males and 7 females)
(Durantin, Scannella, Gateau, Delorme, & Dehais, 2016)	N/A	Flight simulator with complex scenarios	ANOVA	N/A	10-fold CV	SVM	77.8 %	Sensitivity = 79.4 % Specificity = 76 %.	9 (8 males and 1 female)
(Qing, Huang, & Hong, 2021)	Cerebral prefrontal cortex	Visualization of product videos	N/A	N/A	8-fold CV	CNN	86.2 % to 86.3 %	N/A	8 (4 males and 4 females)
(Y. Zhang et al., 2022)	N/A	Mental arithmatic and mental singing	GLM	Kalman filter based features	10-fold CV	Kalman filter and adaptive Gaussian Mixture model	97.89 %	N/A	8 (3 males and 5 females)
(Bak, Yeu, & Jeong, 2022)	Ventrolateral prefrontal cortex, medial prefrontal cortex, and orbitofrontal cortex	Buying behavior related task	*t*-test	Mean, variance, kurtosis, skewness, slope and area	10-fold CV	SVM	94 %	AUC = 0.97	33 (12 males and 21 females)
(Touhid, Anam, Alam, Foysal, & Shaiham, 2023)	N/A	Mental arithmatic	Haar wavelet-based features	Mean, Root mean square value and variance	8-fold CV	Gentle Boost	95.54 %	N/A	N/A
(Hasan, Mahmud, Poudel, Donthula, & Poudel, 2023)	N/A	n-back task	*t*-test	N/A	N/A	Random forests	96.7 %	AUC = 96.7, Precision = 97.0, Recall = 97.0, F1-Score = 97.0	68
(Cakar & Yavuz, 2023)	N/A	n-back task	N/A	N/A	N/A	Generalized Linear Mixed-	N/A	RMSE = 5.6×10^{-4}	26 (9 males and 17 females)

(*continued on next page*)

Table 2 (*continued*)

Authors	Functional near-infrared spectroscopy (fNIRS) properties		fNIRS Features		Artificial intelligence-based approaches		Performance evaluation		Participants
	Brain area	**Environment**	**Feature methods**	**Derived features**	**Strategy**	**Architecture**	**Accuracy**	**Other metrics**	
(Howell-Munson et al., 2023)	N/A	Rule Learning Task	ANOVA	N/A	N/A	Effects Model Tree Logistic regression	N/A	MSE = 3.2 × 10^{-7} F1-score = 0.76	22 (5 males, 13 females and 4 others)
(Y. Zhang et al., 2023)	N/A	Mental arithmetic and mental singing	N/A	Mean, slope and normalization	N/A	CGAN-rIRN	92.19 %	N/A	8 (2 males and 6 females)

conventional MBLL. The results highlighted the fact that a combination of mean and peak values yielded better results in mental arithmetic tasks when the data samples were processed with either FV-MBLL or conventional MBLL. Low classification scores could also be improved through oversampling by balancing the number of features for cognitive tasks.

Durantin et al. (Durantin et al., 2016) optimized the Kalman filter to remove noise and other artifacts from the fNIRS signals. To estimate a pilot's mental state in a simulated flight environment, SVM was trained on the fNIRS signals filtered from Kalman filter, IIR filters and Moving Average Convergence Divergence (MACD) filter (Durantin, Scannella, Gateau, Delorme, & Dehais, 2014). Experimental results show that the predicted accuracy on Kalman filtered data was 77.8 % which was higher when compared with data filtered from IIR filters and MACD filter.

Studies presented so far do not present comparison between SVMs and ML techniques. The study conducted by Kornev et al. (Kornev et al., 2022) not only used the SVM radial basis function for classification but also compared the results with Multiple regression, artificial neural network, random forests and classification and regression trees (CART). Although this study failed to report average accuracy of each algorithm, instead it demonstrated the high performance of SVM in terms of Root Mean Square Error (RMSE) and correlation coefficient (R^2 error).

Despite the promising results offered by SVM for fNIRS signals analysis, most of the studies that used SVM for classification used the normal sized and balanced dataset. The time of the training also increases as the number of sample increases. Secondly, it is difficult to find an appropriate kernel function when the non-linearity in the data increases, so it is always recommended to use an appropriate noise removal technique before using SVM for fNIRS signal classification.

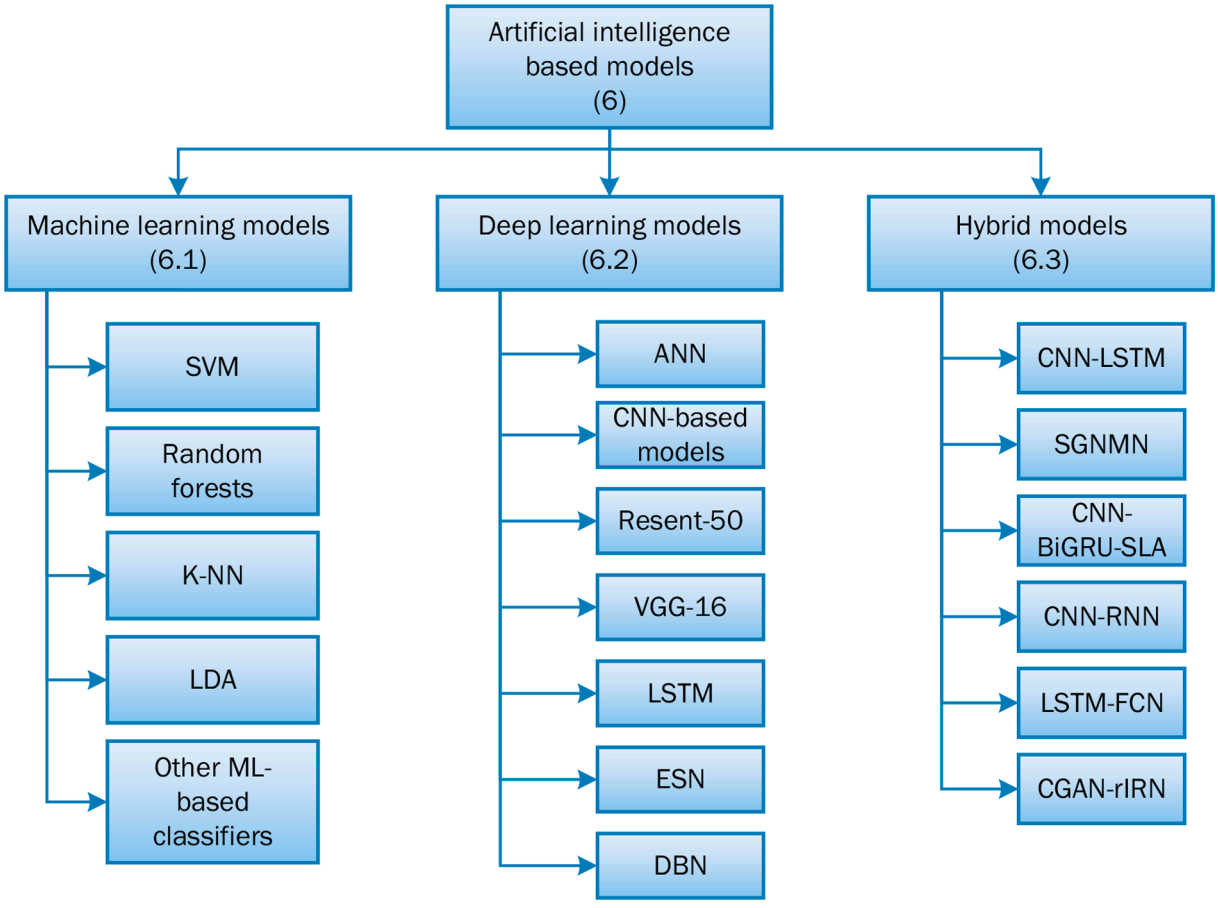

Fig. 8. Taxonomy of AI-based models applied on cognitive load fNIRS data.

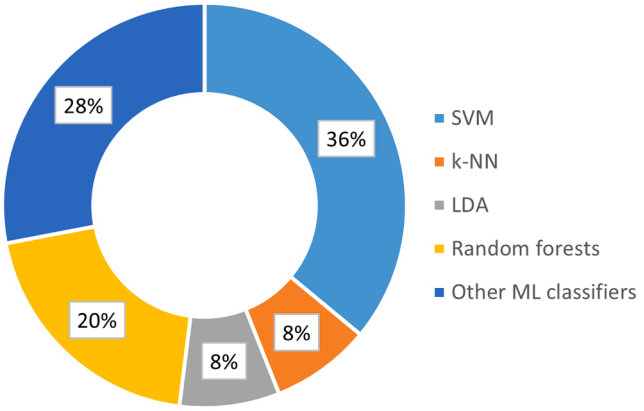

Fig. 9. Distribution of ML studies employed for the classification of fNIRS data.

6.1.2. *k-nearest Neighbor (k-NN) classifier*

The k-Nearest Neighbors (k-NN) algorithm is a widely used ML method for classification and regression tasks. It is based on the principle of close instances, which means that it relies on the similarity between a new data point and the existing data points to classify or predict its label or value. It stores all the training samples, and each input instance is represented as a vector. In k-NN, the "k" refers to the number of nearest neighbors to consider when classifying a new data point. The algorithm works by calculating the distance between the new data point and all the existing data points in the dataset. The k-nearest neighbors are then selected as the data points with the closest distances to the new data point. The classification or prediction of the new data point is based on the labels or values of these k-nearest neighbors. The distance metric used in k-NN can vary depending on the type of data and the problem at hand. The most commonly used distance metrics are Euclidean distance (Durtschi, Mahat, Mashal, & Chrysler, 2021), Manhattan distance (Ehsani & Drabløs, 2020), and Minkowski distance (Iswanto, Tulus, & Sihombing, 2021). The choice of distance metric can have a significant impact on the performance of the algorithm (Shalika & Kumar, 2021). k-NN also requires less training as compared to the other algorithms. It is suitable for the data in which relation between input and output is complex to be expressed as linear models. To classify five different levels of workload during n-back tasks, Saikia et al. (Saikia et al., 2021) evaluated the training time and accuracy of Fine k-NN, Medium k-NN, Coarse k-NN, Cosine k-NN, Cubic k-NN, and Weighted k-NN. In the classification task, both Fine k-NN and Weighted k-NN were able to achieve 75 % accuracy, while Weighted k-NN took a shorter training time (4.93 s) than that of Fine k-NN (5.59 s).

Although k-NN requires less training times as compared to those of other training algorithms, but it requires more computational times during the classification process and determining the results. A study conducted by Naseer et al. (Naseer et al., 2016) reported that classification results produced by k-NN classifier were less accurate as compared to those of other ML algorithms.

6.1.3. *Linear Discriminant analysis (LDA)*

LDA is a well-known dimensionality reduction and feature extraction technique. It is used to identify the linear combination of classes by reducing the dimensionality of vectors belonging to different classes to lower dimensional feature space in a way that features vectors of each class are separated from other classes. This technique is simple to implement and has less computational requirements. Some researchers such as Zhou et al. (X. Zhou et al., 2021) and Cakır et al. (Çakır et al., 2016) used the LDA to classify different levels of mental workload. The main limitation of LDA is its linear nature which prevents the generation of competitive results on non-linear fNIRS signals.

Cakır et al. (Çakır et al., 2016) evaluated the 3 levels mental workload of 8 pilots. The results showed that when the LDA was trained on the data of only single pilot, this model could be generalised to evaluate the mental workload of the remaining pilots. The proposed model also has a high accuracy in predicting low levels of workload but low accuracy in predicting high levels of workload due to frequent head movements. Zhou et al. (X. Zhou et al., 2021) studied hazard perception tasks in a lab environment and indicated that the LDA could achieve an accuracy rate of 70 %, when the model was trained on the features obtained from left prefrontal cortex. Fisher criteria were used to select the top five optimal features from the data and the results indicated that the left prefrontal cortex was involved more in hazard perception tasks as compared to the other regions of the brain.

6.1.4. *Random forests*

Random forest is a tree-based ensemble learning method. It builds a classifier by constructing a number of randomized decision trees (Khan, Asadi, Hoang, Lim, & Nahavandi, 2023). Each decision tree in ensemble classifier casts vote for the predicted class and then the predicted class is determined with most votes on a particular class label. The ensemble nature of model helps random forests to deal with high dimension data and complex feature spaces and make the perfect candidate to handle non-linear fNIRS signals. One of the main advantages of random forests over individual decision trees is that it is less likely to overfit the data. Overfitting occurs when a model is too complex and captures noise or irrelevant patterns in the training data, resulting in poor performance on new, unseen data. By combining multiple decision trees, random forests can reduce the variance of the model and prevent overfitting (Balyan et al., 2022). The random selection of features for each tree also helps to reduce the correlation between the trees and increase their diversity, leading to a better overall performance. The study of (Z. Khan et al., 2020) also found that as compared to other ML algorithms such as SVM and k-NN, it is easier to determine hyperparameter in random forests. The example studies that use random forests for fNIRS signal analysis are Oku et al. (Oku & Sato, 2021), Lamb et al. (Lamb et al., 2022), M. Hasan et al. (Hasan et al., 2023), Le et al. (Le et al., 2022), Le et al. (Le et al., 2018), and Varandas et al. (Varandas et al., 2022).

Varandas et al. (Varandas et al., 2022) used the Corsi-Block task (Milner, 1971) and Lamb et al. (Lamb et al., 2022) used the virtual reality-based environment to induce cognitive load. Both studies reported more than 70 % accuracy when using random forest to classify different levels of mental workload. Le et al. (Le et al., 2018) used the auditory n-back task to classify the different levels of mental workload while driving a car at around 40 km/h. The experimental results (Le et al., 2018) show that the random forests performed better when the data from all the channels were used for classification and the position of channel does not have any significant effect on accuracy. In another study, Le et al. (Le et al., 2022) analysed senior drivers' mental state and indicated that the significant changes were observed while driving a car in relaxed environment, trail driving and parking bay. The results indicated random forests performed better in terms of accuracy, true positive rate and F1-scrore as compared with those from Naive Bayes, Discriminant Analysis, SVM, Decision Trees, and K-NN methods.

Regardless of the ability of random forests in handling fNIRS high dimensional and non-linear data by using large number of decision tress, they still have some limitations. Depending on the nature and complexity of data, a large number of trees are required to overcome the problem of large variance. Random forests may produce spurious results if their parameters are optimally selected. Therefore, it is always advisable to use the cross-validation method to optimize the parameters of random forests model (Sundararajan et al., 2021).

6.1.5. *Diverse approaches in Machine learning for cognitive load analysis in fNIRS studies*

In addition to widely utilized ML classifiers such as SVM, k-NN, LDA, and random forests, logistic regression and gentle boost has also been used in fNIRS cognitive load analysis. While these methods may not be as prevalent, recent studies have shown their effectiveness in enhancing

the understanding of cognitive load dynamics. For example, A. Howell-Munson et al. (Howell-Munson et al., 2023), incorporated behavioral data, including reaction time and task difficulty, in conjunction with fNIRS to comprehensively analyze cognitive load. Their approach employed logistic regression, demonstrating superior results compared to other classifiers. Similarly, the study conducted by T. I. Touhid et al. (Touhid et al., 2023) delved into the comparative analysis of Gentle Boost algorithms alongside established classifiers such as LDA, SVM, and random forests. The experimental findings reveal that Gentle Boost, particularly when utilizing Haar wavelet-based features, exhibited superior performance in comparison with other methods. This suggests that the unique features of Gentle Boost, combined with innovative signal processing techniques like Haar wavelet transformation, contribute to a more enhanced understanding of cognitive load dynamics as captured by fNIRS data.

Beyond widely recognized ML classifiers, there are studies where researchers have proposed their own ML-based classification methods for analyzing cognitive load dynamics in fNIRS data. For instance, Y. Zhang et al. (Y. Zhang et al., 2022), introduced a novel classification method incorporating Kalman filtering and an adaptive Gaussian Mixture model. This approach aimed to identify intricate patterns within fNIRS signals. The results of their study demonstrated a significant improvement in classification accuracy, showing an 87 % improvement compared to conventional classifiers such as GMM, SVM, and LDA. This suggests that the integration of Kalman filtering, and the adaptive Gaussian Mixture model provides a robust framework for extracting meaningful information from fNIRS data and, enhances the efficacy of cognitive load analysis. Similarly, S. Cakar et al. (Cakar & Yavuz, 2023) proposed the Generalized Linear Mixed-Effects Model Tree, which combines Linear Mixed Models (LMM) with ML-based models specifically designed for the analysis of repeated data in fNIRS. By leveraging the strengths of LMM and ML approaches, this study aimed to address the complexities associated with repeated measures in fNIRS experiments.

6.1.6. *Functional connectivity in fNIRS using machine learning algorithms*

The exact working mechanism of the brain is yet to be fully known. Several studies investigate cognitive tasks based on fNIRS responses to critical areas of brain activation. Derosiere et al.(Derosiere et al., 2014) analysed the oxyhemoglobin (HbO_2) features from the right parietal area of the brain, and found that they are more sensitive for classification of cognitive loads as comspared with those from other parts of the brain. Meanwhile, the study conducted by Keles et al.(Keles et al., 2021) on students and surgeons during simulated surgery tasks suggested that the neural activation in the left pre-frontal cortex near the dorso and ventrolateral areas is sufficiently higher than those from other regions. The relationship between HbO_2 features and prefrontal cortex regions was also evaluated by Izzetoglu et al. (Izzetoglu et al., 2021) in simulated driving tasks. During slow-driving tasks, a high level of negative correlation was observed between HbO_2 features and the right prefrontal cortex activations. The logistic regression model was trained on these features, and it yielded an accuracy rate of 97.5 %.

6.2. *Deep learning trends in fNIRS analysis*

Different from ML, the architecture of an DNN contains many hidden layers. Multilayered networks have a finite number of non-linear elements (i.e., activation functions and neurons), which makes them more flexible and robust than ML algorithms. The first and last layers are defined as the input and output, while those in between are defined as hidden layers. Depending on the number of neurons and hidden layers, these models can easily go up to thousands or sometimes up to millions of trainable hyper-parameters. DL is prone to overfitting when dealing with smaller data sets; hence they are better in dealing with massive data sets (J. Wang et al., 2021). Nonetheless, DL can automatically learn useful features from data with less handcrafting effort. We have identified 11 studies on DL for classification of fNIRS signals. Nearly half of these studies have used the CNN models, while four studies leveraged Deep Belief Networks (DBN), Long Short-term Memory (LSTM), ANNs and Echo State Network (ESN). According to our presented taxonomy, use of algorithms other than CNN and LSTM in fNIRS signals is less prevalent. A summary of studies that utilized the DL algorithms for classification is as follows:

CNNs are designed in a way to specifically take images as the input. Numerous CNN variants have been proposed so far, which have shown excellent results in the field of computer vison (Balasundaram et al., 2023), Natural Language Processing (NLP) (Ahmed & Wang, 2023), image segmentation (M. A. Khan et al., 2020), remote sensing (Boulila, Ghandorh, Khan, Ahmed, & Ahmad, 2021), and signal processing (Ghandorh et al., 2021). In fNIRS signal classifications, the input formulation strategies, feature extraction, and feature selection methods vary significantly as a function of the architecture. DL model layers hierarchically extract features from the data samples. Performance of any CNN architecture is depending on the number of convolutional layers, pooling layers and fully connected layers. Convolution layers give the model ability to learn complex features from the data, Pooling layers not only improve the performance of the model but also reduce the dimensionality of feature maps and finally fully connected layers map the complex features to the output. During training CNN continuously optimizes weights and other parameters which will take time and once the model is trained it will take less time for classification.

Khalil et al. (Khalil et al., 2022) proposed a 6-layer CNN to classify four levels of n-back tasks. First, the data of few participants were used to train a CNN model, then the same pre-trained model was used to extract the features from data and employed transfer learning to re-train the model. Although this work does not provide the comparison with other ML/DL methods, but it compared the training time of proposed method with conventional method of training. The results suggest that their method helped in reducing the training time.

Wang et al. (L. M. Wang et al., 2022) used VGG-16 model to study the hemodynamics changes in the brain. Instead of using conventional features, authors converted the fNIRS signals of 52 channels into images which are then used to train CNN model. It was reported that their work with the proposed feature extraction models achieved 100 % accuracy. This work does not provide comparison with other ML/DL models, but it evaluates the model in terms of accuracy, True Positive Rate (TPR) and False Positive Rate (FPR).

Liu et al. (R. Liu et al., 2021) evaluated the performance of Autoencoders for analyzing the fNIRS data. This study demonstrates the significance of features extracted from Echo State Network (ESN) by training model on hand crafted features and feature obtained from convolutional autoencoders. The experimental results show that the features extracted from ESN autoencoders yield better results with an accuracy of 80.61 %.

Benerradi et al. (Benerradi et al., 2019) used a 7-layered CNN to classify the mental workload of two and three levels. The results of classification have also been compared with those from SVM and logistic regression. In 3 class modalities classification, the SVM outperformed other models but in two class modalities, CNN achieved the highest accuracy. The reason for low accuracy could be the small data (9 participants) and the sample size of only 9 s. Secondly, their model architecture has only two convolutional layers which limits the capability to extract features from the data and causes the lower performance of CNN on the three class classification tasks.

Kwon et al. (Kwon & Im, 2021) adopted the CNN model to classify fNIRS signals in mental arithmetic tasks and idle states. The evolving normalization-activation layer (H. Liu, Brock, Simonyan, & Le, 2020) was used, instead of the traditional normalization layer, in the architecture. The dropout probability was set to 0.5. Without using any feature extraction method, the proposed CNN architecture outperformed EEGNet and other ML classifiers.

Qing et al. (Qing et al., 2021) utilized the CNN input layer as a

decoding data matrix to process the conventional features from fNIRS signal lengths of 15 s, 30 s and 60 s. The method achieved 86.3 % accuracy. Zaman et al. (Zaman & Islam, 2021) used Wigner-ville Distribution to transform the fNIRS signals of different window sizes into 2-D images and evaluated the results using ResNet50 (He, Zhang, Ren, & Sun, 2016). The proposed feature extraction method improved the accuracy from 89 % to 98 %. Similarly, in their study, Ho et al. (Ho, Gwak, Park, Khare, et al., 2019) compared the performance of a 9-layer CNN with a 5 layer DBN. PCA was applied on the dataset to reduce the dimensionality. The results indicated that both models exhibited better performance when trained with hemoglobin difference (HbT) features. However, lower accuracy was observed when using oxy-hemoglobin (HbO) and deoxy-hemoglobin (HbR) features. Despite giving outstanding classification accuracy, CNN comes with its disadvantages. CNN requires a large amount of data for training, but the research studies (Cascianelli et al., 2018) used limited number of test subjects. Therefore, more test subjects need to be recruited to increase the size of dataset while incorporating CNN. CNN may give high accuracy on smaller dataset, but it may cause overfitting (Ma et al., 2020). As the fNIRS signals are highly dependent on time, with the signal changes occurring over a range of temporal scales. However, CNNs are designed to capture local features of the data without explicitly modeling the temporal dynamics. This mismatch between the inherent nature of fNIRS signals and the limited capabilities of CNNs to capture temporal dependencies can limit the performance of CNN models on fNIRS datasets. Moreover, due to involvement of a large number of parameters, it is infeasible to express the logic and actual mechanisms involved in the reasoning process of the classification procedures.

To overcome the time series classification in CNN based models, LSTM and RNN models were proposed. Generally, LSTM models are commonly used in neuroergonomic studies because of vanish gradient descent problem in RNN. LSTM models possess input gate, forget gate and output gate which give the model capability to handle sequential data and hence more suitable for fNIRS signals as compared with other models. These models predict the future information by considering past and future, which are not possible using the CNN and other models. Asgher et al. (Asgher et al., 2020) used the model with 4 LSTM layers and 4 dense layers to classify the mental workload of four different layers. The model was trained on mean and slope features extracted from the hemodynamics response while doing mental arithmetic task. The results of classification were compared with those of SVM, k-NN, ANN with a 3-layer network topology and a CNN with 2 convolutional layers, 1 max pooling layer and 4 dense layers. The developed LSTM outperformed other models and achieved the accuracy of 89.01 % followed by CNN with 87.45 %. The CNN model used in their work contains practically very few layers in comparison with those of well-known CNN architectures for example VGG-16 or ResNet. CNN with complex layers could be used in this study to yield better results. In their work, the LSTM model outperformed CNNs but due to lack of studies focusing on transforming time series data to classification tasks, CNN models could perform better as compared with other DL methods.

6.3. Hybrid models trends in fNIRS analysis

Generally, ML methods are reliable when they are used for analysing smaller data sets or hand-crafted features. Similarly, DL techniques tend to function as black boxes and perform more efficiently in terms of feature extraction through trainable hyper-parameters (E. Q. Wu et al., 2021). An increase in the performance cannot be made possible by solely improving the mathematical model of ML or by increasing the number of neurons or hidden layers in DL models. The approach of combining two methods by analyzing the information of a data set leads to a hybrid model. We identify four studies on hybrid models for classification of fNIRS signals. Most hybrid models in this review combine the convolutional operator of the CNN layers with RNN, LSTM or GRU (Lu et al., 2020; Saadati et al., 2021; J. Wang et al., 2021). The main purpose of CNNs is to extract features, while RNN, LSTM, or GRU can be used to handle data dependencies. A combination of both make a perfect fit to extract features from fNIRS signals and, at the same time, leverage the present and past data samples to learn the nature of workload patterns. These hybrid models can have an ability to classify the mental workload of subjects in the presence of noisy data and improve model efficiency from 10 % to 15 %, as indicated in the literature. Additionally, we identified a study that utilizes GANs for the analysis. Gt et al. [5] proposed a GAN-based network to classify fNIRS signals, specifically using Convolutional-based Generative Adversarial Networks (CGAN) to generate synthetic fNIRS signals. They also proposed the revised Inception Net (rIRN) to classify fNIRS signals. The model was trained on the real and synthetic features of size 160x10. The quality of generated signals has been evaluated through Maximum Mean Discrepancy (MMD), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR). Experiments revealed that increasing the dataset up to two times increased the accuracy of the model, and further increases in the dataset size decreased the accuracy. They also compared the performance of rIRN with IRN and CNN with different layers, noting that each model had a similar effect on accuracy, but rIRN yielded the highest accuracy. For the distribution of the dataset, neither k-fold nor LOOC cross-validation has been applied.

Most of the studies presented so far have tested the collected data set using various classification algorithms. It is inappropriate to highlight the best algorithm by comparing the accuracy metrice or feature extracted methods. Each study has its own architecture design, input processing method, and unique feature selection technique. Identifying the best algorithm for classification is a challenging task because researchers evaluate the validity of an algorithm by utilisng data samples with more than one ML or DL methods and find the most suitable one. Nevertheless, the analysis provided in this review helps reveal future research directions of ML/D- based algorithms for cognitive load analysis.

7. Discussion and challenges

This paper presents a comprehensive overview of research methodologies employing ML and DL approaches for the classification of cognitive load. We identified 45 experimental studies that utilized fNIRS signals to discern varying levels of cognitive load. We conducted a preliminary analysis in our systematic review to identify the cognitive tasks used in each of the sampled research. A spectrum of cognitive tasks was observed, with some studies incorporating traditional paradigms such as n-back tasks, stroop tasks, and mental arithmetic tasks. Additionally, a noteworthy aspect of the investigated literature revealed a divergence, with certain studies devising unique tasks related to activities like flying, driving, and game-based scenarios. A consistent finding across all these studies pertains to the observed correlation between increased cognitive load and heightened cortical activations in the brain. This aligns with the conceptual framework of CLT, substantiating the premise that cognitive load escalates proportionally with the demands imposed by the task at hand. The results underscore the robustness of fNIRS signals as indicators of cognitive load.

One prevalent issue encountered in the application of ML methods is the inherent challenge associated with data requirements, often necessitating larger datasets compared to traditional methods to attain comparable performance levels. has become a paradigm shift that simplifies and turns the fNIRS signal processing pipeline into an end-to-end task. This paradigm shift holds significant promise, simplifying the intricacies associated with data processing and analysis. The integration of deep learning techniques has the potential to revolutionize cognitive load classification by not only mitigating the challenges posed by fNIRS data but also by offering a more efficient approach to signal processing.

To move beyond the competition among various methodologies, and to provide a comprehensive framework for directing future endeavors in the field of automated cognitive load inference, as well as addressing the

certain peculiarities associated with fNIRS data are shown in Fig. 10, it becomes imperative to elucidate a distinct approach for the construction of cognitive load inference pipelines. These considerations imply a specific set of guidelines and methodologies that should be incorporated into the design and implementation of AI-based algorithms for cognitive load inference.

7.1. fNIRS features

The classification of various brain activities relies heavily on specific features extracted from hemodynamic signals. Currently, many studies utilize ML techniques to classify varying levels of mental workload effectively using fNIRS data. By isolating features that closely align with the characteristics of a particular class and significantly differ from those of other classes, the classification process becomes more effective in capturing distinctions in hemodynamic signals (L. Wu, Liu, Ward, Wang, & Chen, 2023). However, the substantial dimensionality of fNIRS data presents a significant challenge due to the numerous fNIRS channels, introducing the well-known issue in machine learning known as the curse of dimensionality. In the fNIRS signals domain, researchers often lack comprehensive knowledge about relevant features, leading to the inclusion of numerous candidates features to better represent the domain. Hemodynamic signals, such as HbO2, dHb, and HbT, provide a wide array of choices for feature selection due to their capacity to encompass pertinent information regarding brain activities (Z. Wang, Fang, & Zhang, 2023). Different combinations of such features provide the necessary discriminatory information for classification. Feature selection is also dependent on individual activities, the mean, peak, variance, skewness, kurtosis and slope values of HbO2, dHb, and HbT frequently have been used in fNIRS studies. In the initial stages of fNIRS studies, researchers typically compute the concentration changes of hemoglobin oxygenation throughout the task period (Murata, Sakatani, Katayama, & Fukaya, 2002). This method involves presenting time-series data illustrating cerebral oxygenation alterations for visual inspection. However, these approaches are susceptible to error, especially with increasing noise and interference levels. To address this, various statistical analysis methods have been applied to enhance the accuracy and reliability of feature extraction from fNIRS signals.

In the literature various methods have been proposed to extract cortical activities from fNIRS data, primarily utilizing changes in HbO2. Commonly employed statistical techniques in fNIRS studies include the Wilcoxon signed-rank test, Shapiro-Wilk test, *t*-test and ANOVA (Bak et al., 2022; Durantin et al., 2016; Keles et al., 2021; Khalil et al., 2022). These methods compare differences between conditions with respect to condition variance. To avoid assumptions about the exact shape or timing of the time course of changes in HbO2 and dHb in response to stimuli, these approaches often take average values during the task period. Features extracted from fNIRS signals typically provide a measure known as the *p*-value, which indicates the level of significance. However, it is essential to recognize a potential issue associated with interpreting *p*-values. A *p*-value of 0.05, for instance, implies that there is a 5 % chance of obtaining the observed result if the null hypothesis were true. In simpler terms, if 100 statistical tests are conducted, and the null hypothesis is true for all of them, it is expected that, by chance, 5 of them will be deemed significant at the $p < 0.05$ level. This statistical phenomenon underscores the importance of cautious interpretation of *p*-values, as the probability of obtaining significant results by chance increases with the number of statistical tests performed.

Additionally, GLM is also a popular and adaptable analytical technique for examining fNIRS signals at the individual and group levels (Y. Zhang et al., 2022). Because of its adaptability to both quantitative and qualitative independent variables, it is well-suited to capture the

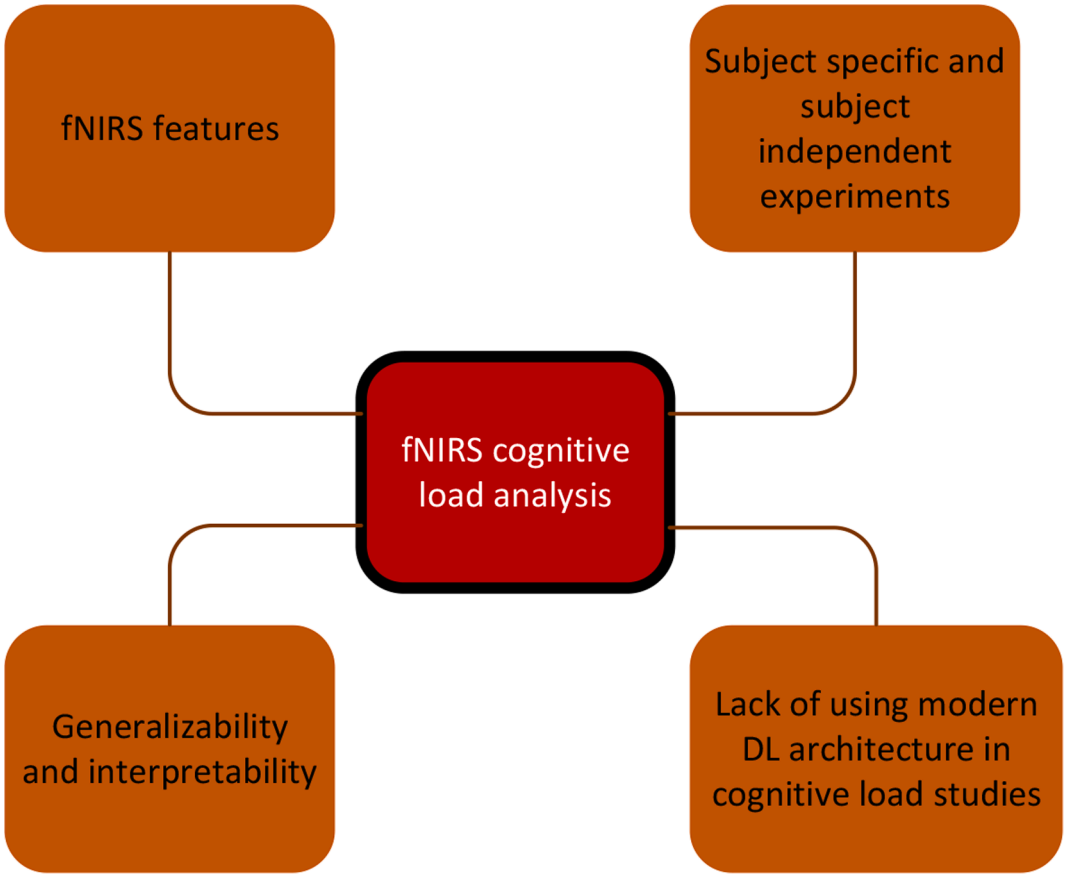

Fig. 10. Challenges associated with cognitive load analysis with fNIRS data.

complex dynamics of cognitive processes. In the fNIRS studies, GLM plays an important role in analyzing the functional timeline of data, aligning with the actual hemodynamic response observed in the brain. The functional timeline of data in GLM analyses involves tracking variations in HBO2 and dhB signals over time. The method involves multiple regression analyses, where GLM is incorporated as a linear combination of regressors to predict or explain a related variable. In the case of fNIRS studies, these regressors are carefully selected to represent various experimental conditions or cognitive states, allowing for a comprehensive examination of the underlying neural processes. In addition to the conventional feature extraction methods mentioned earlier, researchers in the field of brain activity classification have explored alternative approaches, incorporating features from the frequency domain for example, Wavelet based features, Haar wavelet and Wigner-Ville distribution to reveal distinct patterns in hemodynamic signals. Frequency domain features are commonly applied in signal processing to analyze time-series data by decomposing it into different frequency components. In the fNIRS studies, the frequency domain analyis has been employed to extract features that capture temporal variations in hemodynamic signals. This approach allows for the identification of specific frequency components associated with different cognitive processes, contributing valuable information for classification tasks. Furthermore, some researchers have proposed their own unsupervised feature extraction methods, introducing novel techniques to capture unique aspects of brain activity. These methods often aim to identify patterns or features that may not be apparent through traditional approaches, enhancing the richness of information available for classification.

Feature extraction methods, as described earlier, have found extensive application in cognitive load studies. Notably, these methods play a crucial role in the analysis of hemodynamic signals, fNIRS data. While traditional statistical techniques like the *t*-test and ANOVA have been prevalent in extracting features, advancements in ML have introduced DL methods that often bypass the need for explicit feature extraction due to their deep neural or convolutional-based architectures. Fig. 11 presents a comparative view of approaches in feature extraction strategies within both ML and DL frameworks for fNIRS studies. This figure not only illustrates the utilization of statistical feature extraction methods but also highlights the studies that opt for raw fNIRS data. Interestingly, the rise of DL methods has not eliminated the use of feature extraction in certain studies. Despite the inherent capability of deep neural networks to automatically learn hierarchical representations, there are instances where researchers have incorporated feature extraction methods into DL frameworks. This integration aims to enhance the interpretability of the model or to extract specific information from hemodynamic signals that may not be captured effectively by the neural network alone. It is noteworthy that each feature extraction method, whether traditional or novel, has its own set of advantages and limitations. The selection of a particular method depends on the study's objectives and the characteristics of the dataset under consideration. Traditional statistical techniques like the *t*-test and ANOVA are known for their simplicity and ease of interpretation. They provide insights into the average values and variance of features during specific experimental conditions, aiding in the understanding of differences in brain activities. On the other hand, frequency domain methods, including Wavelet features, Haar wavelet, and Wigner-Ville distribution, offer a more comprehensive evaluation of the temporal and frequency characteristics of hemodynamic signals. These methods, applied in signal processing, allow the decomposition of time-series data into different frequency components. In fNIRS studies, the frequency domain analysis becomes particularly valuable as it enables the identification of specific frequency components associated with different cognitive processes. The coexistence of both traditional and novel feature extraction methods highlights the versatility and adaptability required in the field of brain activity classification. Researchers continue to explore and refine these techniques to address the challenges posed by the dimensionality in fNIRS data, ensuring that the extracted features are not only relevant but also contribute meaningfully to the accurate classification of mental workload and other cognitive states.

7.2. Subject specific and subject independent experiments

The evaluation of the classification performance of fNIRS data is typically conducted in an offline manner, utilizing pre-recorded data-

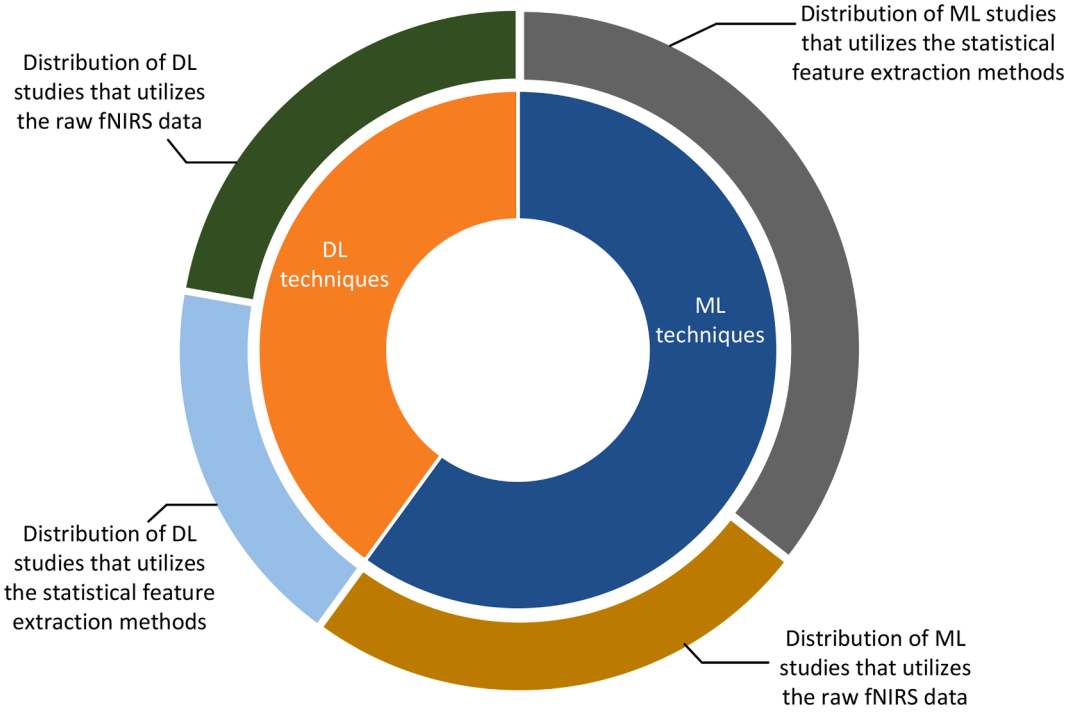

Fig. 11. A Comparative view of feature extraction strategies in ML and DL fNIRS studies.

sets. Within the existing body of literature, a predominant trend emerges wherein researchers commonly employ either *k*-fold cross-validation (*k*-fold CV) or Leave-One-Out Cross-Validation (LOOCV) methods to measuere the effectiveness of their models. In the context of *k*-fold CV, the dataset is partitioned into *k* subsets or folds. The model is trained on $k-1$ of these folds and evaluated on the remaining one. This process is repeated k times, with each fold serving as the test set exactly once. The results are then averaged to provide a comprehensive performance metric that accounts for variations in the training and testing data. On the other hand, LOOCV involves leaving out a single data point as the test set while training the model on the remaining dataset. This process is iteratively repeated for each data point in the dataset, ensuring that each instance serves as a test set exactly once. The final performance metric is derived by averaging the results across all iterations. LOOCV is particularly useful when dealing with smaller datasets, as it maximizes the use of available data for both training and testing. Examining the distribution of studies utilizing different cross-validation methods, Fig. 12 illustrates the prevalence of specific strategies within the research community. Notably, among these methods, 10-fold cross-validation has been widely accepted and frequently employed by researchers. It is followed by 5-fold, 8-fold, LOOCV, and 20-fold CV methods, each demonstrating varying degrees of adoption within the scholarly community. Despite the popularity of specific cross-validation approaches, a noteworthy finding from the analysis is that 42 % of the studies do not explicitly mention the validation methods employed.

The fNIRS data exhibits inherent subject dependence and session dependence, characterized by substantial inter-subject and inter-session variabilities (Huang et al., 2021). Consequently, when a model is trained and tested on the same subjects or sessions, the performance results may significantly differ from those obtained when testing on new subjects or sessions that were not encountered during the training phase. To tackle the challenges posed by subject dependence and session dependence in subject's data various techniques have been devised. These techniques include within-subject, subject-specific, subject-dependent, cross-subject, and subject-independent approaches. Within-subject methods or subject-specific involve training and testing on the same subject, focusing on individual variations. Cross-subject methods, on the other hand, involve training on one set of subjects and testing on a different set, aiming to generalize across individuals. Subject-independent methods are designed to create models that can be trained on one set of subjects and seamlessly applied to a completely new set, thus addressing the challenge of generalization.

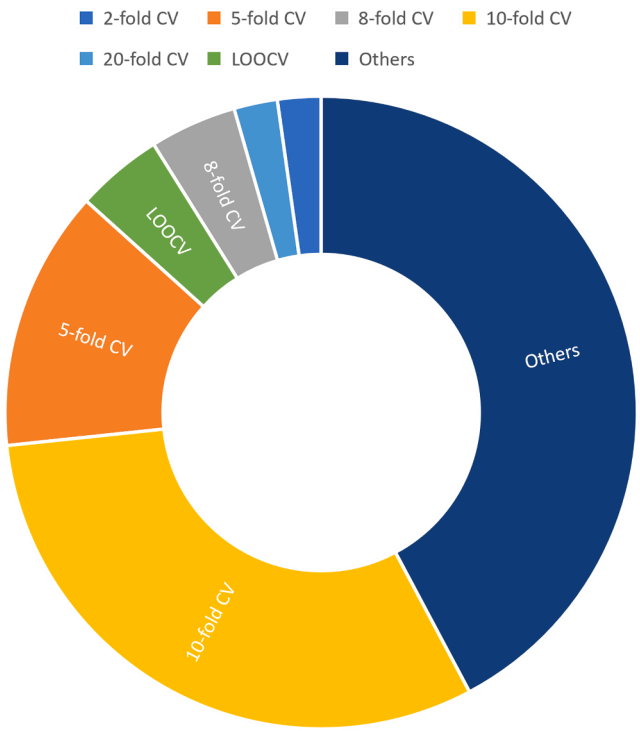

Fig. 12. Distribution of studies using different CV Methods.

Despite the existence of these methodological advancements, a notable gap exists in the current literature on cognitive load and fNIRS classification. There is a lack of specific implementation of within-subject, subject-specific, subject-dependent and cross-subject methods in studies within this domain. The presence of significant inter-subject variability poses a significant challenge in the classification of cognitive load using fNIRS data. In the majority of studies, ML/DL models for cognitive load are commonly trained and tested using k-fold or LOOCV methodologies. This training approach is favored for its ability to yield higher classification accuracy (Y. Zhou et al., 2021). However, a notable drawback is its limited generalization ability across different subjects. Despite the prevalence of k-fold and LOOCV methods in training models for cognitive load classification, there has been a lack of comparative analyses between these cross-validation techniques and subject-specific methods within the fNIRS community. Conversely, such evaluations have been undertaken in related domains, such as EEG and other physiological signal domains. To address this challenge effectively, future studies should prioritize adopting subject-specific methods that explicitly consider the individual characteristics of each subject in the training and testing phases.

7.3. Generalizability and interpretability challenges in cognitive load studies

The lack of explainability in fNIRS poses a substantial hurdle in cognitive load research. While fNIRS is a valuable tool for capturing neural activity and understanding cognitive processes, it frequently struggles to offer transparent explanations for their findings and the underlying mechanisms behind them. One common approach in fNIRS analysis involves employing traditional ML and DL techniques, treating AI as a black box without delving into the interpretability of the results. The prevalent utilization of traditional ML and DL methods without sufficient explainability limits our understanding of the cognitive load phenomena captured by fNIRS. While these approaches can yield accurate predictions or classifications based on fNIRS data, they often lack the ability to provide meaningful insights into the neural processes and features driving those predictions. Researchers in the field of of neurology have used models based on CNN, LSTM, GANs, and autoencoders to analyze fNIRS data. However, a noticeable gap exists in the literature as there is a lack of studies specifically dedicated to investigating the generalizability and interpretability of DL models in the cognitive load domain using fNIRS data.

To address this gap, it is essential to leverage layer-wise model explanation techniques in the analysis of fNIRS signals. These techniques offer valuable insights into the inner workings of deep learning models and provide a deeper understanding of the specific brain regions, functional connections, and neural patterns associated with cognitive processes. Several layer-wise model explanation techniques, such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016), Gradient-weighted Class Activation Mapping (GradCAM) (M. Han & Kim, 2019), and Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), can be utilized in the analysis of fNIRS data. By applying these layer-wise model explanation techniques to fNIRS data, researchers can gain valuable insights into the underlying neural mechanisms of cognitive processes. These techniques enable the identification of specific brain regions, functional connections, and neural patterns that contribute to cognitive load, attention, memory, or other cognitive states. Additionally, these explanations can provide interpretable evidence for the predictions made by deep learning models, enhancing the understanding and trustworthiness of the results. Furthermore, combining these layer-wise model explanation techniques

with traditional statistical analyses can lead to a comprehensive understanding of fNIRS. By integrating the strengths of both approaches, researchers can validate and interpret the findings in a more robust manner. This knowledge allows researchers to focus on the most informative regions or wavelengths in the brain, enabling a more targeted and interpretable investigation of cognitive load. Moreover, the development of hybrid models that combine traditional ML/DL approaches with XAI techniques holds promise for bridging the gap between accuracy and interpretability in fNIRS research. These models can retain the predictive power of ML/DL algorithms while providing transparent explanations for their outcomes.

7.4. Lack of using modern architecture in cognitive load studies

Various classifiers have been utilized in conjunction with ML algorithms to address the task of classification or labeling and to train systems in quantifying different levels of cognitive workloads. The classification of cognitive load using fNIRS data has been explored through diverse machine learning algorithms, including SVM, k-NN, LDA, and Random Forests. SVM, known for its simplicity and high accuracy, is extensively employed in fNIRS signal analysis, demonstrating its effectiveness in mental workload classification. The k-NN classifier is notable for its shorter training time, albeit with increased computational demands during classification. LDA's simplicity and low computational requirements are acknowledged, but its linear nature poses challenges in handling non-linear fNIRS signals. Random Forests are praised for their capacity to handle high-dimensional and non-linear data, with studies reporting success in mental workload classification.

Deep learning models, with a specific focus on CNNs and LSTM networks, are also discussed. CNNs, originally designed for image inputs, are explored for their ability to transform fNIRS signals into images and classify hemodynamic changes in the brain. The advantages of CNNs, such as high accuracy, are contrasted with their limitations, including data size requirements and potential overfitting. LSTM, addressing temporal dynamics, is highlighted for outperforming other machine learning methods in certain studies.

In the past few years, newer DL architectures such as GhostNet (K. Han et al., 2020), Densenet (Y. Zhu & Newsam, 2017), and Capsule Net (Sabour, Frosst, & Hinton, 2017) have gained attention for their improved robustness, optimization, and better generalization capabilities compared to earlier models. These architectures have shown success in various computer vision tasks, but their potential in the context of cognitive load classification using fNIRS signals remains largely unexplored. Furthermore, the recent rise of transformer-based models, originally designed for natural language processing tasks, introduces a new dimension to DL. Transformers, with their attention mechanisms, have demonstrated superior generative AI capabilities compared to traditional architectures like GANs. The attention mechanisms in transformers allow them to capture complex relationships in data, making them potentially advantageous for tasks involving intricate patterns, such as those found in cognitive load studies. It is imperative to evaluate these modern DL architectures and transformer-based models specifically in the field of cognitive load classification using fNIRS technology. Their enhanced capabilities in handling complex relationships and capturing patterns may lead to improved accuracy and interpretability in understanding neural activity associated with cognitive processes. As the field of cognitive load research continues to evolve, embracing these newer DL architectures and transformer-based models can contribute to a more comprehensive understanding of the brain's response to cognitive tasks, offering novel insights into the intricacies of cognitive load classification with fNIRS data.

8. Future implications and limitations

The main limitation of this article is that it is focused on the theme of AI and cognitive load in a relation with fNIRS only, whereas areas such as motor imagery, stress, and emotion recognition have been excluded. The main reason to exclude these areas is that either they are very wide, or they have been explained previously. Future work to improve the interpretation of AI models and clinical applicable metrics will be necessary to translate AI models in daily use.

In this review, we explore the feasibility of using fNIRS indices to quantify mental workload during various cognitively demanding tasks. The presence of open-source libraries has made it possible for the scientific community to design DL architectures with relative ease. In DL studies, the trend of using their own data set has increased. Secondly, fNIRS signals are highly affected by the age, gender, demographic and size of the data sets (Huang et al., 2021). Studies presented so far consider a limited number of participants and unequal gender distribution, as shown in Table 2. Besides that, the analysis performed on the data sets is based on the general interpretation of fNIRS signals; hence, it is difficult to compare the model performance based on various metrics used in the published studies.

Despite astonishing developments in AI, research on fNIRS is still in the early development phase. The relationships between different brain regions and across different cognitively demanding tasks still need further investigation. Few studies suggest that neural activations are higher in the left pre-frontal regions during cognitively demanding tasks, while some suggest that features from the right pre-frontal regions are best suited for DL analysis (Derosiere et al., 2014; Keles et al., 2021; Kornev et al., 2022). The list of challenges mentioned in Section 7 not only is valid in the field of neurology but also applies to other health domains. AI has become an increasingly popular topic of research in recent years, especially in relation to cognitive load. The majority of articles reviewed in this study on cognitive load focused on the emerging technology of AI, and these articles were published within the past three years. Almost all studies that compare DL with ML or with raw data instead of using handcrafted features reported a small but meaningful improvement. We observed that there is a scope of improvement in modelling and designing DL models because almost all of the studies use their own dataset to benchmark AI models. Reluctance in sharing data or model architecture limits the scope of work to small scale project.

A wide variety of both ML and DL models to analyze fNIRS signals have been proposed so far, which makes it difficult to identify the best-performing models due to a lack of comparison provided in publications. Delayed responses in fNIRS signals cause difficulty to synchronize with online analysis. Studies presented so far mostly emphasize feature selection and classification on an offline basis. The next big leap in fNIRS research could be automation using DL models. AI is likely to advance neurosciences in the near future. Research institutions should provide demographic-rich (age, gender, race) fNIRS data in a standardized format without compromising the privacy of participants. Advancement in the portable wearable fNIRS sensors will effectively reduce the errors in measurement. Availability of data will also help researchers to design optimized model architecture which can be deployed to mobile devices by using tools like TensorFlow Lite. This would enable neuroscientists to develop real-time applications by using inexpensive and portable fNIRS devices.

9. Conclusion

fNIRS is an important tool and can classify cognitive load in human performance tasks. This study has reviewed ML/DL methods used in the assessment of cognitive load by using the PRISMA protocol. In this paper, we reviewed the studies that applied DL-based classification methods on fNIRS signals collected from the participants during n-back tasks, Stroop tasks and simulated game-based tasks. The model architecture in the reviewed studies vary significantly depending on the input formulation and the task under consideration. These architectural differences can have a significant impact on the model's performance and the overall effectiveness of the AI system. This article has pointed out key strengths of ML/DL algorithms and surveyed the major

achievements and limitations of state-of-the-art ML/DL approaches for fNIRS signals. By analyzing 45 articles that utilized ML/DL models to classify cognitive load based on fNIRS data, it was concluded that more than 70 % of the studies have applied CNN directly or in the form of hybrid architecture to the fNIRS signals. It was inferred that most of the researchers have adopted feature extraction techniques to leverage the full potential of ML/DL models. Some researchers also aimed to utilize convolutional layers to analyze local features from the data. Feature extraction methods ensure that the input is readily usable for model training. Few studies also indicate that features extracted from left or right prefrontal cortex of the brain can be a factor that affects the model's accuracy. AI models can be trained using various methods, the efficiency of the model depends on the quality of preprocessing on fNIRS signals. DL algorithms are computationally expensive, they outperform ML algorithms with low pre-processing demand. We highlighted the fact that the future investigation of DL models in the domain of cognitive load not only aims at improving the accuracy of models but also inspects the aspects of practicability, such as robustness, explanation, and optimization.

We found that hybrid models generally achieve better performance compared with those of traditional models and have more potential to accurately classify different levels of mental workload. The hybrid models incorporating convolutional layers with recurrent layers are able to outperform the conventional methods. We have recommended that an in-depth investigation of hybrid models is beneficial, particularly the number of layers and arrangement of convolutional layers, fully connected layers, and recurrent layers. As cognitive studies focus merely on the objective and system paradigms, no classification technique can be declared as the best option for general use. Several challenges have been identified in the literature including model interpretability and feature engineering. We expect that AI has a potential to meet these challenges by transferring latest advances in DL technologies into massive multimodal data of fNIRS signals.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research was supported by the Australian Research Council (ARC) (Project ID: DE210101623).

References

Abibullaev, B., & An, J. (2012). Classification of frontal cortex haemodynamic responses during cognitive tasks using wavelet transforms and machine learning algorithms. *Medical engineering & physics, 34*(10), 1394–1410.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*(11), e00938.

Agbangla, N. F., Audiffren, M., Pylouster, J., & Albinet, C. T. (2022). Load-dependent prefrontal cortex activation assessed by continuous-wave Near-Infrared spectroscopy during two executive tasks with three cognitive loads in young adults. *Brain Sciences, 12*(11), 1462.

Ahmed, Z., & Wang, J. (2023). A fine-grained deep learning model using embedded-CNN with BiLSTM for exploiting product sentiments. *Alexandria Engineering Journal, 65*, 731–747.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *Paper presented at the 2017 international conference on engineering and technology (ICET)*.

Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., & Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. *Paper presented at the Computer graphics forum*.

Asadi, H., Bellmann, T., Qazani, M. C., Mohamed, S., Lim, C. P., & Nahavandi, S. (2023). A novel decoupled model predictive control-based motion cueing algorithm for driving simulators. *IEEE Transactions on Vehicular, Technology(99)*, 1–12.

Asadi, H., Mohammadi, A., Mohamed, S., Qazani, M. R. C., Lim, C. P., Khosravi, A., & Nahavandi, S. (2019). A model predictive control-based motion cueing algorithm using an optimized nonlinear scaling for driving simulators. *Paper presented at the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*.

Asgher, U., Ahmad, R., Naseer, N., Ayaz, Y., Khan, M. J., & Amjad, M. K. (2019). Assessment and classification of mental workload in the prefrontal cortex (PFC) using fixed-value modified beer-lambert law. *IEEE Access, 7*, 143250–143262.

Asgher, U., Khalil, K., Khan, M. J., Ahmad, R., Butt, S. I., Ayaz, Y., & Nazir, S. (2020). Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. *Frontiers in neuroscience, 14*, 584.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one, 10*(7), e0130140.

Bai, X., Wang, X., Liu, X., Liu, Q., Song, J., Sebe, N., & Kim, B. (2021). Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition, 120*, Article 108102.

Bak, S., Yeu, M., & Jeong, J. (2022). Forecasting Unplanned Purchase Behavior under Buy-One Get-One-Free Promotions Using Functional Near-Infrared Spectroscopy. *Computational intelligence and neuroscience, 2022*.

Baker, W. B., Parthasarathy, A. B., Busch, D. R., Mesquita, R. C., Greenberg, J. H., & Yodh, A. (2014). Modified beer-Lambert law for blood flow. *Biomedical optics express, 5*(11), 4053–4075.

Balasundaram, A., Dilip, G., Ashokkumar, S., Manickam, M., Gurunathan, K., & Kothandaraman, D. (2023). Detecting True Medicinal Leaves Among Similar Leaves Using Computer Vision and CNN. In: Springer.

Balyan, A. K., Ahuja, S., Lilhore, U. K., Sharma, S. K., Manoharan, P., Algarni, A. D., & Raahemifar, K. (2022). A hybrid intrusion detection model using ega-pso and improved random forest method. *Sensors, 22*(16), 5986.

Banuelos-Lozoya, E., Gonzalez-Serna, G., Gonzalez-Franco, N., Fragoso-Diaz, O., & Castro-Sanchez, N. (2021). A systematic review for cognitive state-based QoE/UX evaluation. *Sensors, 21*(10), 3439.

Benerradi, J., Maior, A. H., Marinescu, A., Clos, J. L., & Wilson, M. (2019). Exploring machine learning approaches for classifying mental workload using fNIRS data from HCI tasks. *Paper presented at the Proceedings of the Halfway to the Future Symposium 2019*.

Berivanlou, N. H., Setarehdan, S. K., & Noubari, H. A. (2016). Quantifying mental workload of operators performing n-back working memory task: Toward fNIRS based passive BCI system. *Paper presented at the 2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME)*.

Bhagat, S., Banerjee, H., Ho Tse, Z. T., & Ren, H. (2019). Deep reinforcement learning for soft, flexible robots: Brief review with impending challenges. *Robotics, 8*(1), 4.

Bhangale, K. B., & Kothandaraman, M. (2022). Survey of deep Learning Paradigms for speech processing. *Wireless Personal Communications*, 1–37.

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

Boulila, W., Ghandorh, H., Khan, M. A., Ahmed, F., & Ahmad, J. (2021). A novel CNN-LSTM-based approach to predict urban expansion. *Ecological Informatics, 64*, Article 101325.

Breiman, L. (2001). *Random forests. Machine learning, 45*(1), 5–32.

Buchner, J., Buntins, K., & Kerres, M. (2021). A systematic map of research characteristics in studies on augmented reality and cognitive load. *Computers and Education Open, 2*, Article 100036.

Buckley, S., Chaput, M., Simon, J. E., Criss, C. R., Brazalovich, P., McCarren, G., & Grooms, D. R. (2022). Cognitive load impairs time to initiate and complete shooting tasks in ROTC members. *Military Medicine, 187*(7–8), e898–e905.

Cakar, S., & Yavuz, F. G. (2023). Hybrid statistical and machine learning modeling of cognitive neuroscience data. *Journal of Applied Statistics*, 1–22.

Çakır, M. P., Vural, M., Koç, S.Ö., & Toktaş, A. (2016). Real-time monitoring of cognitive workload of airline pilots in a flight simulator with fNIR optical brain imaging technology. *Paper presented at the International Conference on Augmented Cognition*.

Canário, N., Jorge, L., Martins, R., Santana, I., & Castelo-Branco, M. (2022). Dual PET-fMRI reveals a link between neuroinflammation, amyloid binding and compensatory task-related brain activity in Alzheimer's disease. *Communications biology, 5*(1), 1–7.

Cascianelli, S., Bello-Cerezo, R., Bianconi, F., Fravolini, M. L., Belal, M., Palumbo, B., & Kather, J. N. (2018). Dimensionality reduction strategies for cnn-based classification of histopathological images. *Paper presented at the Intelligent Interactive Multimedia Systems and Services, 2017*, 10.

Castro-Alonso, J. C., de Koning, B. B., Fiorella, L., & Paas, F. (2021). Five strategies for optimizing instructional materials: Instructor-and learner-managed cognitive load. *Educational Psychology Review, 33*(4), 1379–1407.

Catana, C., Drzezga, A., Heiss, W.-D., & Rosen, B. R. (2012). PET/MRI for neurologic applications. *Journal of nuclear medicine, 53*(12), 1916–1925.

Cosme, G., Tavares, V., Nobre, G., Lima, C., Sá, R., Rosa, P., & Prata, D. (2022). Cultural differences in vocal emotion recognition: A behavioural and skin conductance study in Portugal and Guinea-Bissau. *Psychological Research, 86*(2), 597–616.

Curum, B., & Khedo, K. K. (2021). Cognitive load management in mobile learning systems: Principles and theories. *Journal of Computers in Education, 8*(1), 109–136.

DeMaris, A. (1995). A tutorial in logistic regression. *Journal of Marriage and the Family*, 956–968.

Derosiere, G., Dalhoumi, S., Perrey, S., Dray, G., & Ward, T. (2014). Towards a near infrared spectroscopy-based estimation of operator attentional state. *PloS one, 9*(3), e92045.

Dong, S., & Jeong, J. (2018). Onset classification in hemodynamic signals measured during three working memory tasks using wireless functional near-infrared spectroscopy. *IEEE Journal of Selected Topics in Quantum Electronics, 25*(1), 1–11.
Durantin, G., Scannella, S., Gateau, T., Delorme, A., & Dehais, F. (2014). Moving average convergence divergence filter preprocessing for real-time event-related peak activity onset detection: Application to fNIRS signals. *Paper presented at the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.
Durantin, G., Scannella, S., Gateau, T., Delorme, A., & Dehais, F. (2016). Processing functional near infrared spectroscopy signal with a Kalman filter to assess working memory during simulated flight. *Frontiers in Human Neuroscience, 9*, 707.
Durtschi, B., Mahat, M., Mashal, M., & Chrysler, A. (2021). Preliminary analysis of RFID localization system for moving precast concrete units using multiple-tags and weighted Euclid distance k-NN algorithm. *Paper presented at the 2021 IEEE International Conference on RFID (RFID)*.
Eastmond, C., Subedi, A., De, S., & Intes, X. (2022). Deep learning in fNIRS: A review. *Neurophotonics, 9*(4), 041411.
Ehsani, R., & Drabløs, F. (2020). Robust distance measures for k NN classification of cancer data. *Cancer informatics, 19*, 1176935120965542.
Farkish, A., Bosaghzadeh, A., Amiri, S. H., & Ebrahimpour, R. (2022). Evaluating the effects of educational multimedia design principles on cognitive load using EEG signal analysis. *Education and Information Technologies*, 1–17.
Fischer-Jbali, L., Montoro, C., Montoya, P., Halder, W., & Duschek, S. (2022). Central nervous activity during an emotional stroop task in fibromyalgia syndrome. *International Journal of Psychophysiology, 177*, 133–144.
Fix, E., & Hodges, J. (1951). *Discriminatory analysis.* Nonparametric Discrimination: Consistency Properties USAF School of Aviation Medicine, Randolph Field. Retrieved from.
Fournier, É., Kilgus, D., Landry, A., Hmedan, B., Pellier, D., Fiorino, H., & Jeoffrion, C. (2022). The impacts of human-cobot collaboration on perceived cognitive load and usability during an industrial task: An exploratory experiment. *IISE Transactions on Occupational Ergonomics and Human Factors(just-accepted)*, 1–11.
Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., & Andersen, S. A. W. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: A randomized trial. *Surgical endoscopy, 34*(3), 1244–1252.
Fujikawa, S., Ohsumi, C., Ushio, R., Tamura, K., Sawai, S., Yamamoto, R., & Nakano, H. (2022). *Potential applications of motor imagery for improving standing posture balance in rehabilitation.* In Neurorehabilitation and Physical Therapy: IntechOpen.
Gateau, T., Durantin, G., Lancelot, F., Scannella, S., & Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PloS one, 10*(3), e0121279.
Gemignani, J., & Gervain, J. (2021). Comparing different pre-processing routines for infant fNIRS data. *Developmental cognitive neuroscience, 48*, Article 100943.
Ghandorh, H., Khan, M. Z., Alsufyani, R., Khan, M., Alsofayan, Y. M., Khan, A. A., & Alahmari, A. A. (2021). An ICU admission predictive model for COVID-19 patients in Saudi Arabia. *International Journal of Advanced Computer Science and Applications, 12*(7).
Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). Ghostnet: More features from cheap operations. *Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
Han, M., & Kim, J. (2019). Joint banknote recognition and counterfeit detection using explainable artificial intelligence. *Sensors, 19*(16), 3607.
Harauzov, A., Ivanova, L., Vasiliev, P., & Podvigina, D. (2022). fMRI studies of opponent interregional Interactions in the Macaca mulatta brain. *Journal of Evolutionary Biochemistry and Physiology, 58*(4), 1001–1014.
Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.
Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics), 28*(1), 100-108.
Hasan, M., Mahmud, M., Poudel, S., Donthula, K., & Poudel, K. (2023). Mental workload classification from fNIRS signals by leveraging machine Learning. *Paper presented at the 2023 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition*.
Heitmann, S., Grund, A., Fries, S., Berthold, K., & Roelle, J. (2022). The quizzing effect depends on hope of success and can be optimized by cognitive load-based adaptation. *Learning and Instruction, 77*, Article 101526.
Ho, T. K. K., Gwak, J., Park, C. M., Khare, A., & Song, J.-I. (2019). Deep leaning-based approach for mental workload discrimination from multi-channel fNIRS. In *Recent trends in communication, computing, and electronics* (pp. 431–440). Springer.
Ho, T. K. K., Gwak, J., Park, C. M., & Song, J.-I. (2019). Discrimination of mental workload levels from multi-channel fNIRS using deep leaning-based approaches. *IEEE Access, 7*, 24392–24403.
Howell-Munson, A., Unal, D. S., Mowad, T., Arrington, C., Walker, E., & Solovey, E. (2023). Classification of brain signals collected during a rule Learning Paradigm. *Paper presented at the International Conference on Artificial Intelligence in Education*.
Huang, Z., Wang, L., Blaney, G., Slaughter, C., McKeon, D., Zhou, Z., . . . Hughes, M. C. (2021). The Tufts fNIRS Mental Workload Dataset & Benchmark for Brain-Computer Interfaces that Generalize.
Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: Lessons we have learned. *The International Journal of Robotics Research, 40*(4–5), 698–721.
Iswanto, I., Tulus, T., & Sihombing, P. (2021). Comparison of distance models on K-Nearest neighbor algorithm in stroke disease detection. *Applied Technology and Computing Science Journal, 4*(1), 63–68.
Izzetoglu, M., Jiao, X., & Park, S. (2021). *Understanding driving behavior using fNIRS and machine learning.* Paper presented at the International Conference on Transportation and Development 2021.
Keles, H. O., Cengiz, C., Demiral, I., Ozmen, M. M., & Omurtag, A. (2021). High density optical neuroimaging predicts surgeons's subjective experience and skill levels. *PloS one, 16*(2), e0247117.
Khademi, Z., Ebrahimi, F., & Kordy, H. M. (2022). A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals. *Computers in Biology and Medicine, 143*, Article 105288.
Khalil, K., Asgher, U., & Ayaz, Y. (2022). Novel fNIRS study on homogeneous symmetric feature-based transfer learning for brain–computer interface. *Scientific Reports, 12*(1), 1–12.
Khan, M. A., Asadi, H., Hoang, T., Lim, C. P., & Nahavandi, S. (2023). Measuring cognitive load: Leveraging fNIRS and machine Learning for classification of workload levels. *Paper presented at the International Conference on Neural Information Processing*.
Khan, M. A., Khan, M. A., Ahmed, F., Mittal, M., Goyal, L. M., Hemanth, D. J., & Satapathy, S. C. (2020). Gastrointestinal diseases segmentation and classification based on duo-deep architectures. *Pattern Recognition Letters, 131*, 193–204.
Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2020). Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification, 14*, 97–116.
Khanam, F., Hossain, A. A., & Ahmad, M. (2022). Statistical valuation of cognitive load level hemodynamics from functional Near-Infrared spectroscopy signals. *Neuroscience Informatics, 100042*.
Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*.
Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology, 55*(4), 352.
Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior, 27*(1), 99–105.
Klein, F., Debener, S., Witt, K., & Kranczioch, C. (2022). fMRI-based validation of continuous-wave fNIRS of supplementary motor area activation during motor execution and motor imagery. *Scientific Reports, 12*(1), 1–20.
Kooijman, L., Asadi, H., Mohamed, S., & Nahavandi, S. (2022). Does a Secondary task inhibit vection in virtual reality?. *Paper presented at the 2022 IEEE International Conference on Systems*. and Cybernetics (SMC.
Kooijman, L., Asadi, H., Mohamed, S., & Nahavandi, S. (2023). A virtual reality study investigating the train illusion. *Royal Society Open Science, 10*(4), Article 221622.
Kornev, D., Nwoji, S., Sadeghian, R., Esmaili Sardari, S., Dashtestani, H., He, Q., & Aram, S. (2022). Gaming behavior and brain activation using functional near-infrared spectroscopy, Iowa gambling task, and machine learning techniques. *Brain and Behavior, 12*(4), e2536.
Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*(4), 261–283.
Krampe, C. (2022). The application of mobile functional near-infrared spectroscopy for marketing research–a guideline. *European Journal of Marketing, 56*(13), 236–260.
Kurihara, Y., Kaburagi, T., Nishio, K., Hamada, Y., Matsumoto, T., & Kumagai, S. (2020). Discrimination of verbal/visuospatial memory retrieval processes by measuring prefrontal lobe blood volume with functional Near-Infrared spectrometry. *IEEE Access, 8*, 208683–208695.
Kwon, J., & Im, C.-H. (2021). Subject-independent functional Near-Infrared spectroscopy-based brain-computer Interfaces based on convolutional neural networks. *Frontiers in Human Neuroscience, 15*, Article 646915.
Lagomarsino, M., Lorenzini, M., De Momi, E., & Ajoudani, A. (2022). An online framework for cognitive load assessment in industrial tasks. *Robotics and Computer-Integrated Manufacturing, 78*, Article 102380.
Lamb, R., Neumann, K., & Linder, K. A. (2022). Real-time prediction of science student learning outcomes using machine learning classification of hemodynamics during virtual reality and online learning sessions. *computers and education. Artificial Intelligence, 100078*.
Lamichhane, B., Westbrook, A., Cole, M. W., & Braver, T. S. (2020). Exploring brain-behavior relationships in the N-back task. *NeuroImage, 212*, Article 116683.
Le, A. S., Aoki, H., Murase, F., & Ishida, K. (2018). A novel method for classifying driver mental workload under naturalistic conditions with information from near-infrared spectroscopy. *Frontiers in Human Neuroscience, 12*, 431.
Le, A. S., Xuan, N. H., & Aoki, H. (2022). Assessment of senior drivers' internal state in the event of simulated unexpected vehicle motion based on near-infrared spectroscopy. *Traffic injury prevention*, 1–5.
Lewis, B., Garcia, C. C., Price, J. L., Schweizer, S., & Nixon, S. J. (2022). Cognitive training in recently-abstinent individuals with alcohol use disorder improves emotional stroop performance: Evidence from a randomized pilot trial. *Drug and Alcohol Dependence, 231*, Article 109239.
Li, R., Yang, D., Fang, F., Hong, K.-S., Reiss, A. L., & Zhang, Y. (2022). Concurrent fNIRS and EEG for brain function investigation: A systematic. *Methodology-Focused Review. Sensors, 22*(15), 5865.
Lim, L. G., Ung, W. C., Chan, Y. L., Lu, C.-K., Sutoko, S., Funane, T., & Tang, T. B. (2020). A unified analytical framework with multiple fNIRS features for mental workload assessment in the prefrontal cortex. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28*(11), 2367–2376.

Liu, H., Brock, A., Simonyan, K., & Le, Q. (2020). Evolving normalization-activation layers. *Advances in Neural Information Processing Systems, 33*, 13539–13550.

Liu, R., Reimer, B., Song, S., Mehler, B., & Solovey, E. (2021). Unsupervised fNIRS feature extraction with CAE and ESN autoencoder for driver cognitive load classification. *Journal of Neural Engineering, 18*(3), Article 036002.

Liu, R., Wang, L., Koszalka, T. A., & Wan, K. (2022). Effects of immersive virtual reality classrooms on students' academic achievement, motivation and cognitive load in science lessons. *Journal of Computer Assisted Learning*.

Liu, Y., Lu, S., Liu, J., Zhao, M., Chao, Y., & Kang, P. (2022). A Characterization of brain area activation in orienteers with different map-recognition memory ability task levels—Based on fNIRS evidence. *Brain Sciences, 12*(11), 1561.

Lu, J., Yan, H., Chang, C., & Wang, N. (2020). Comparison of machine learning and deep learning approaches for decoding brain computer interface: An fNIRS study. *Paper presented at the International Conference on Intelligent Information Processing*.

Ma, H., Liu, Y., Ren, Y., Wang, D., Yu, L., & Yu, J. (2020). Improved CNN classification method for groups of buildings damaged by earthquake, based on high resolution remote sensing images. *Remote Sensing, 12*(2), 260.

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences, 19*(3), 303–342.

Magnusdottir, E. H., Johannsdottir, K. R., Majumdar, A., & Gudnason, J. (2022). Assessing cognitive workload using Cardiovascular measures and voice. *Sensors, 22*(18), 6894.

Mehta, N., & Shukla, S. (2022). Pandemic analytics: How countries are leveraging big data analytics and artificial intelligence to fight COVID-19? *SN Computer Science, 3*(1), 1–20.

Mejia-Puig, L., & Chandrasekera, T. (2022). The presentation of self in virtual reality: A cognitive load study. *journal of. Interior Design.*

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British medical bulletin*.

Murata, Y., Sakatani, K., Katayama, Y., & Fukaya, C. (2002). Increase in focal concentration of deoxyhaemoglobin during neuronal activity in cerebral ischaemic patients. *Journal of Neurology, neurosurgery, and psychiatry, 73*(2), 182.

Naseer, N., Qureshi, N. K., Noori, F. M., & Hong, K.-S. (2016). Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface. *Computational intelligence and neuroscience, 2016*.

Nasirizad Moghadam, K., Chehrzad, M. M., Reza Masouleh, S., Maleki, M., Mardani, A., Atharyan, S., & Harding, C. (2021). Nursing physical workload and mental workload in intensive care units: Are they related? *Nursing Open, 8*(4), 1625–1633.

Oku, A. Y. A., & Sato, J. R. (2021). Predicting student performance using machine learning in fNIRS data. *Frontiers in Human Neuroscience, 15*, Article 622224.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Brennan, S. E. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic reviews, 10*(1), 1–11.

Patlar Akbulut, F. (2022). Hybrid deep convolutional model-based emotion recognition using multiple physiological signals. *Computer Methods in Biomechanics and Biomedical Engineering*, 1–13.

Picon, A., San-Emeterio, M. G., Bereciartua-Perez, A., Klukas, C., Eggers, T., & Navarra-Mestre, R. (2022). Deep learning-based segmentation of multiple species of weeds and corn crop using synthetic and real image datasets. *Computers and Electronics in Agriculture, 194*, Article 106719.

Qing, K., Huang, R., & Hong, K.-S. (2021). Decoding three different preference levels of consumers using convolutional neural network: A functional near-infrared spectroscopy study. *Frontiers in Human Neuroscience, 14*, Article 597864.

Reddy, G., Spencer, C. A., Durkee, K., Cox, B., Fox Cotton, O., Galbreath, S., & Severe-Valsaint, G. (2022). Estimating cognitive load and cybersickness of pilots in VR simulations via unobtrusive physiological sensors. *Paper presented at the International Conference on Human-Computer Interaction*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *" Why should i trust you?" Explaining the predictions of any classifier.* Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering, 16*(5), Article 051001.

Saadati, M., Nelson, J., Curtin, A., Wang, L., & Ayaz, H. (2021). Application of recurrent convolutional neural networks for mental workload assessment using functional near-infrared spectroscopy. *Paper presented at the International Conference on Applied Human Factors and Ergonomics*.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems, 30*.

Saha, S., Jindal, K., Shakti, D., Tewary, S., & Sardana, V. (2022). Chirplet transform-based machine-learning approach towards classification of cognitive state change using galvanic skin response and photoplethysmography signals. *Expert Systems, e12958*.

Saikia, M. J., Kuanar, S., Borthakur, D., Vinti, M., & Tendhar, T. (2021). *A machine learning approach to classify working memory load from optical neuroimaging data.* Paper presented at the Optical Techniques in Neurosurgery, Neurophotonics, and Optogenetics.

Saleem, A. A., Siddiqui, H. U. R., Raza, M. A., Rustam, F., Dudley, S., & Ashraf, I. (2023). A systematic review of physiological signals based driver drowsiness detection systems. *Cognitive neurodynamics, 17*(5), 1229–1259.

Sepp, S., Howard, S. J., Tindall-Ford, S., Agostinho, S., & Paas, F. (2019). Cognitive load theory and human movement: Towards an integrated model of working memory. *Educational Psychology Review, 31*(2), 293–317.

Shalika, M., & Kumar, V. (2021). Enriching and clustering short text using KNN. *International Research Journal on Advanced Science Hub, 3*, 111–116.

Skulmowski, A., & Xu, K. M. (2021). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, 1–26.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American scientist, 57*(4), 421–457.

Stone, J. V. (2002). Independent component analysis: An introduction. *Trends in cognitive sciences, 6*(2), 59–64.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643.

Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics, 4*(3), 275–294.

Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval, 11*(1), 19–38.

Sundararajan, K., Georgievska, S., Te Lindert, B. H., Gehrman, P. R., Ramautar, J., Mazzotti, D. R., & Ridder, L. (2021). Sleep classification from wrist-worn accelerometer data using random forests. *Scientific Reports, 11*(1), 24.

Sweller, J. (2016). Cognitive load theory, evolutionary educational psychology, and instructional design. *Evolutionary perspectives on child development and education*, 291–306.

Touhid, T. I., Anam, M., Alam, M. R., Foysal, M., & Shaiham, S. (2023). *Study on Accuracy Improvement of Mental Arithmetic Task Classification Using Different Classifiers with DWT Feature Extraction Method.* Paper presented at the 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE).

Tugtekin, U., & Odabasi, H. F. (2022). Do Interactive Learning environments have an effect on Learning outcomes, cognitive load and metacognitive judgments? *Education and Information Technologies*, 1–40.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Varandas, R., Lima, R., & Bermúdez i Badia, S., Silva, H., & Gamboa, H.. (2022). Automatic cognitive fatigue detection using wearable fNIRS and machine learning. *Sensors, 22*(11), 4010.

Wang, H., Guo, H., Zhang, K., Gao, L., & Zheng, J. (2022). Automatic sleep staging method of EEG signal based on transfer learning and fusion network. *Neurocomputing, 488*, 183–193.

Wang, J., Grant, T., Velipasalar, S., Geng, B., & Hirshfield, L. (2021). Taking a deeper look at the brain: Predicting visual perceptual and working memory load from high-density fNIRS data. *IEEE Journal of Biomedical and Health Informatics, 26*(5), 2308–2319.

Wang, L., Huang, Z., Zhou, Z., McKeon, D., Blaney, G., Hughes, M. C., & Jacob, R. J. (2021). *Taming fNIRS-based BCI Input for Better Calibration and Broader Use.* Paper presented at the The 34th Annual ACM Symposium on User Interface Software and Technology.

Wang, L. M., Huang, Y. H., Chou, P. H., Wang, Y. M., Chen, C. M., & Sun, C. W. (2022). Characteristics of brain connectivity during verbal fluency test: Convolutional neural network for functional near-infrared spectroscopy analysis. *Journal of Biophotonics, 15*(1), e202100180.

Wang, Z., Fang, J., & Zhang, J. (2023). Rethinking delayed hemodynamic responses for fNIRS classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Wani, J. A., Sharma, S., Muzamil, M., Ahmed, S., Sharma, S., & Singh, S. (2022). Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: Methodologies, applications, and challenges. *Archives of Computational Methods in Engineering, 29*(1), 641–677.

Wilson, J. C., Nair, S., Scielzo, S., & Larson, E. C. (2021). Objective measures of cognitive load using deep multi-modal learning: A use-case in aviation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5*(1), 1–35.

Wu, E. Q., Tang, Z., Yao, Y., Qiu, X.-Y., Deng, P.-Y., Xiong, P., & Zhou, M. (2021). Scalable gamma-driven multilayer network for brain workload detection through functional near-infrared spectroscopy. *IEEE transactions on. Cybernetics.*

Wu, L., Liu, A., Ward, R. K., Wang, Z. J., & Chen, X. (2023). Signal processing for brain-computer Interfaces: A review and current perspectives. *IEEE Signal Processing Magazine, 40*(5), 80–91.

Yan, Z., Wu, Y., Li, Y., Shan, Y., Li, X., & Hansen, P. (2022). Design eye-Tracking augmented reality headset to reduce cognitive load in repetitive Parcel scanning task. *IEEE Transactions on Human-Machine Systems, 52*(4), 578–590.

Žagar, D., Svetina, M., Brcko, T., Perkovič, M., Dimc, F., & Košir, A. (2022). Analysis of Marine-pilot biometric data recordings during port-approach using a full-Mission simulator. *Sensors, 22*(7), 2701.

Zaman, S., & Islam, R. (2021). Classification of FNIRS using wigner-ville distribution and CNN. *int. J. Image. Graph. Signal Process, 13*, 1–13.

Zhang, H., Zhang, Y., Xiao, Y., & Wu, C. (2022). Analyzing the influencing factors and workload variation of takeover behavior in semi-autonomous vehicles. *International journal of environmental research and public health, 19*(3), 1834.

Zhang, Y., Liu, D., Li, T., Zhang, P., Li, Z., & Gao, F. (2023). CGAN-rIRN: A data-augmented deep learning approach to accurate classification of mental tasks for a fNIRS-based brain-computer interface. *Biomedical optics express, 14*(6), 2934–2954.

Zhang, Y., Liu, D., Zhang, P., Li, T., Li, Z., & Gao, F. (2022). Combining robust level extraction and unsupervised adaptive classification for high-accuracy fNIRS-BCI: An evidence on single-trial differentiation between mentally arithmetic-and singing-tasks. *Frontiers in neuroscience, 16*, Article 938518.

Zhao, G., Zhang, L., Chu, J., Zhu, W., Hu, B., He, H., & Yang, L. (2022). An augmented reality based Mobile photography application to improve Learning gain, decrease cognitive load, and achieve better emotional state. *International Journal of Human-Computer Interaction*, 1–16.

Zhou, T., Cha, J. S., Gonzalez, G., Wachs, J. P., Sundaram, C. P., & Yu, D. (2020). Multimodal physiological signals for workload prediction in robot-assisted surgery. *ACM Transactions on Human-Robot Interaction (THRI), 9*(2), 1–26.

Zhou, X., Hu, Y., Liao, P.-C., & Zhang, D. (2021). Hazard differentiation embedded in the brain: A near-infrared spectroscopy-based study. *Automation in Construction, 122*, Article 103473.

Zhou, Y., Huang, S., Xu, Z., Wang, P., Wu, X., & Zhang, D. (2021). Cognitive workload recognition using EEG signals and machine learning: A review. *IEEE Transactions on Cognitive and Developmental Systems*.

Zhu, Q., Shi, Y., & Du, J. (2021). Wayfinding information cognitive load classification based on functional Near-Infrared spectroscopy (fNIRS). *Journal of computing in civil engineering*.

Zhu, R., Wang, Z., Ma, X., & You, X. (2022). High expectancy influences the role of cognitive load in inattentional deafness during landing decision-making. *Applied Ergonomics, 99*, Article 103629.

Zhu, Y., & Newsam, S. (2017). *Densenet for dense flow*. Paper presented at the 2017 IEEE international conference on image processing (ICIP).

Zhuang, C., Meidenbauer, K. L., Kardan, O., Stier, A. J., Choe, K. W., Cardenas-Iniguez, C., & Berman, M. G. (2022). Scale Invariance in fNIRS as a measurement of cognitive load. *Cortex*.