REVIEW

Deep learning in fNIRS: a review

Condell Eastmond,*,† Aseem Subedi,† Suvranu De, and Xavier Intes
Center for Modeling, Simulation and Imaging for Medicine, Rensselaer Polytechnic, Department of Biomedical Engineering, Troy, New York, United States

Abstract

Significance: Optical neuroimaging has become a well-established clinical and research tool to monitor cortical activations in the human brain. It is notable that outcomes of functional near-infrared spectroscopy (fNIRS) studies depend heavily on the data processing pipeline and classification model employed. Recently, deep learning (DL) methodologies have demonstrated fast and accurate performances in data processing and classification tasks across many biomedical fields.

Aim: We aim to review the emerging DL applications in fNIRS studies.

Approach: We first introduce some of the commonly used DL techniques. Then, the review summarizes current DL work in some of the most active areas of this field, including brain–computer interface, neuro-impairment diagnosis, and neuroscience discovery.

Results: Of the 63 papers considered in this review, 32 report a comparative study of DL techniques to traditional machine learning techniques where 26 have been shown outperforming the latter in terms of the classification accuracy. In addition, eight studies also utilize DL to reduce the amount of preprocessing typically done with fNIRS data or increase the amount of data via data augmentation.

Conclusions: The application of DL techniques to fNIRS studies has shown to mitigate many of the hurdles present in fNIRS studies such as lengthy data preprocessing or small sample sizes while achieving comparable or improved classification accuracy.

 [DOI: 10.1117/1.NPh.9.4.041411]

Keywords: functional near-infrared spectroscopy; brain–machine interface; data processing; biophotonics; real-time imaging.

Paper 22009SSVRR received Jan. 26, 2022; accepted for publication Jun. 22, 2022; published online Jul. 20, 2022.

1 Introduction

Over the last two decades, functional near-infrared spectroscopy (fNIRS) has become a well-established neuroimaging modality to monitor brain activity.[1] The ability of fNIRS to quantify cortical tissue hemodynamics over a long time, with relative high-spatial sampling and temporal resolution, has enabled its adoption in numerous clinical settings.[2,3] fNIRS offers the unique advantage to be employed in freely mobile subjects with less restrictions than electroencephalography (EEG) or functional magnetic resonance imaging (fMRI). This permits the deployment of fNIRS in naturalistic scenarios and in patient populations that are typically not considered suitable for EEG or fMRI imaging.[4] Still, fNIRS faces numerous challenges for increased clinical adoption due to experimental settings,[5] variations in statistical results,[6] etc. Of importance, current trends in fNIRS aim to improve spatial resolution via increased spatial sampling, improve cortical sensitivity using data processing to remove unwanted physiological noise, improve quantification by anatomical coregistration, and increase robustness via artifact identification and removal.[7] Current algorithmic implementations, however, require a high level of expertise

*Address all correspondence to Condell Eastmond, eastmc2@rpi.edu
†These authors contributed equally to this work.

to set up parameters that can be system- and/or application-specific but also greatly impact the interpretability of the processed data. Moreover, the computational cost of these methods does not lend itself to bedside implementations. Following a ubiquitous trend in the field of data processing and analysis, new approaches leveraging developments in deep learning (DL) have been recently proposed to help overcome these caveats to a large extent.

The successes of DL methodologies across all biomedical engineering fields promise the development of dedicated data-driven, model-free data processing tools with robust performances, user-friendly employability, and real-time capabilities. DL methods are increasingly utilized across the biomedical imaging field, including biomedical optics[8] and neuroimaging modalities including fMRI, magnetoencephalography (MEG), and EEG.[9] Following this trend, DL methodologies have also been recently used for fNIRS applications. In this review, we provide a summary of these current efforts. First, we introduce the basic concepts of DL including training and design considerations. Second, we provide a synthetic summary of the different studies reporting DL models in fNIRS applications. This section is divided into subfield, namely brain–computer interface (BCI), clinical diagnostic, and analysis of cortical activity. We then provide a short discussion and future outlook section.

2 Deep Learning Methodology

DL can be viewed as black-box version of parametric machine learning techniques. Traditional machine learning techniques might make several assumptions about raw data distributions. Most notable is that data can be mapped to distinct classes (categorical data) or a regression line (continuous data) by a suitable transformation of input data. In parametric methods, weights are used to reduce multidimensional input data into separable space. These weights are learned by minimizing the objective function, trying to reduce error in prediction of score. With the introduction of hidden layers between input and output space, it is argued[10] that several abstractions can be learned that help in better distinction. These models are termed artificial neural networks (ANNs).

Inspired by biological neurons, ANNs generate a map between the input training data (x) and the output (y) using simple nonlinear operations performed at nodes, or "neurons," that form a computational graph.[11] The weights (Θ) of the edges of the graph are updated, i.e., "trained" by minimizing a loss function $L(y, \hat{y})$ that measures the difference between the model output ($\hat{y}$) and the true output (y). Network training is accomplished efficiently using the chain rule of differentiation in an algorithm known as backpropagation using gradient descent.[12] The number of nodes in each layer of the graph defines the width of the network, whereas the number of layers defines its depth.

A deep network is one with sufficient depth, though there is no consensus on how deep the network has to be considered a "deep neural network." (DNN) The entire network can be viewed as a differentiable function that learns a relationship between the input and the output using multiple levels of function composition, with the initial layers learning low-level features of the input, and the deeper layers extracting higher-level features. The advent of high-performance computing and the availability of large-scale data are fueling the current rapid advances in DL.

2.1 *Deep Learning Architectures*

DL derives its versatility from the available cell/nodal operations. These operations include linear transformations, filters, and gates that have been inspired by other domain-specific tools, which compute abstract features in the hidden layers. In general, the choice of these operations defines the application of a network. The three most commonly used types of networks used in fNIRS studies are shown in Fig. 1.

Dense networks such as multilayer perceptrons (MLPs) use linear transformation as cell operations and are synonymous to ANNs. Convolutional neural networks (CNNs) use convolution operations, where a fixed-size filter is used for convolution over the input image or feature map. These are shown to be highly capable of recognizing digits,[13] objects,[14] and images in general. Similarly, long-short-term-memory (LSTM) networks can be unfurled in the time

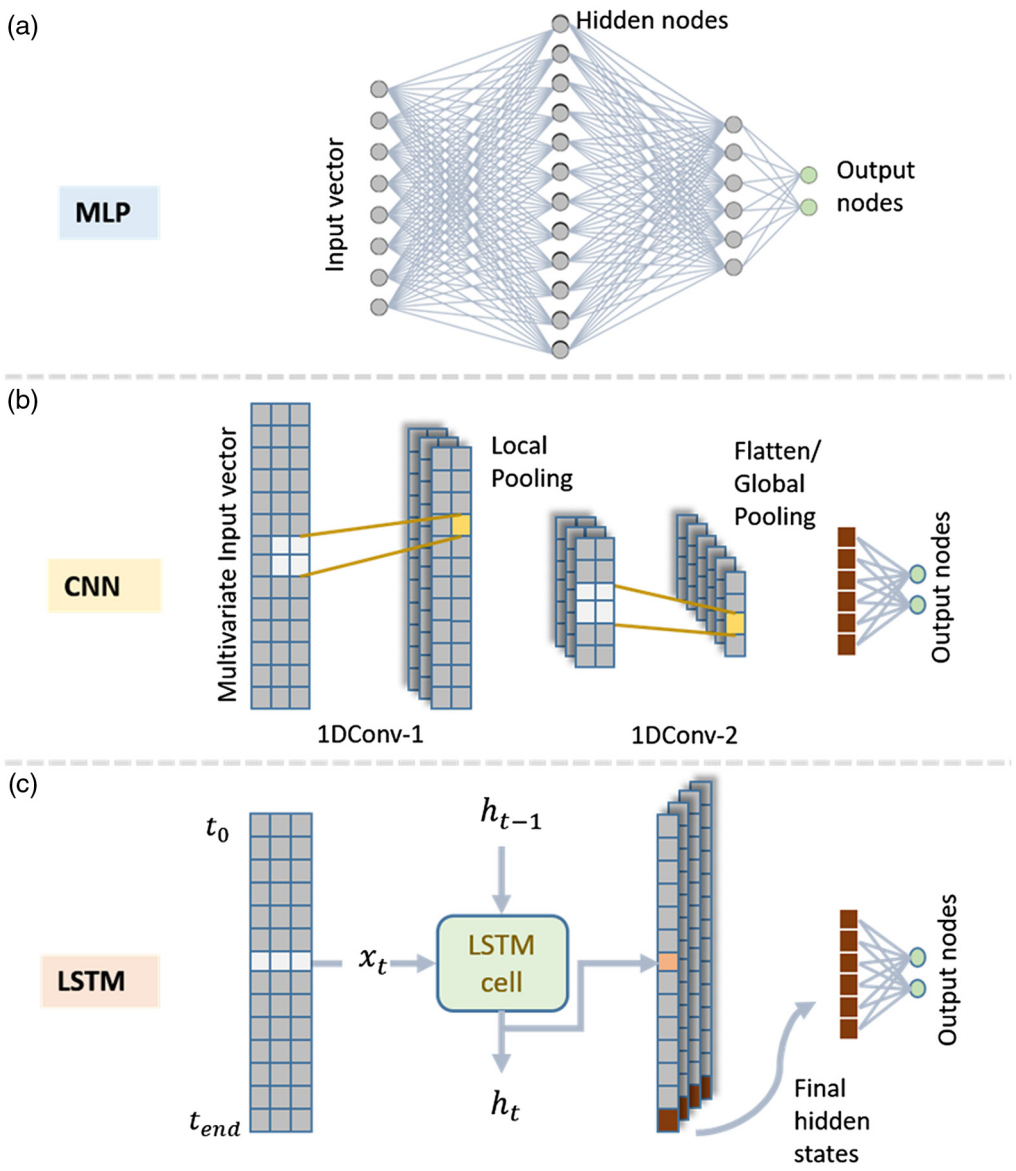

Fig. 1 Illustrations of the three most common classes of network architectures used in the reviewed articles are shown in the figure above. (a) An MLP has all nodes fully connected. (b) A convolutional NN (CNN) with a kernel size 2×2, with subsequent pooling layers. (c) An LSTM architecture where final hidden states are used. For illustrative purposes, the output layer is constructed for binary classification problems.

domain to learn time series data, including handwriting[15] and semantics for language translation.[16] LSTM cells use gates to maintain a cell-state memory.[17] Appropriate data are passed through the cells in successive timesteps, avoiding the vanishing gradient problem (see Sec. 1.2.1) for long time series data.[18]

A fundamental attribute of neural networks is introducing nonlinearities using so-called "activation functions"—based on the idea of the firing of biological neurons. The most common types of activation functions are listed in Table 1. The sigmoid function compresses values in the range of 0 and 1, activating large positive values while zeroing out large negative ones. The sigmoid function can also be used at the output for binary classifiers, for example, in Refs. 19 and 20. Dolmans et al.[21] used the sigmoid output for seven-class classification, although the softmax activation is widely used for multiclass classification. The softmax activation uses exponentiation followed by normalization to assign probabilities to the outputs. Higher softmax outputs at the output layer may be interpreted as confidence in the prediction.[22]

Table 1 Activated outputs of input "x" from each of the activation functions is shown in the table. Also shown are the bounds/range of activated values.

Activation function	Output values	Bounds
Sigmoid	$\frac{1}{1+e^{-x}}$	(0,1)
Softmax	$\frac{e^{x_i}}{\sum_j e^{x_j}}$	(0,1)
ReLU	$\max(0, x)$	$[0, \infty)$
Leaky-ReLU	if $x < 0$, αx else x	$(-\infty, \infty)$
ELU	if $x < 0$, $\alpha(e^x - 1)$ else x	$(-\alpha, \infty)$

The rectified linear unit (ReLU) activation is most widely used for the hidden nodes since it is computationally inexpensive and helps resolve the vanishing gradient problem.[22] It has proven to be efficient and effective for CNNs. It deactivates all nodes with negative outputs, which, although effective, deactivates those nodes throughout the training. This issue, also known as dead-ReLU problem, is solved by a recent activation function called exponential linear unit (ELU), which has also been used by Mirbagheri et al.,[19] Saadati et al.,[23] Ortega and Faisal,[24] whereas some also used another variant called leaky ReLU.[25,26]

The final ingredient of neural networks is the loss function, which depends upon the output type. For classification tasks, the cross-entropy loss function[27] measures the difference in the probability distributions between ground truths and network predictions. Multiclass classifications[28–31] use categorical cross-entropy as the loss function, whereas binary classification problems use binary cross-entropy. For regression[32] or reconstruction[33] problems, the mean squared error loss function is used. These loss functions may be modified to implement constraints or regularization, e.g., a linear combination of MSE, variance, and two other metrics for a denoising autoencoder (DAE) in Ref. 34.

2.2 *Practical Considerations in Deep Learning*

2.2.1 *Training deep networks*

After deciding on functions and architecture suitable to the problem, it is important to understand the underlying considerations involved to help the network learn the input–output map. It is common to normalize or scale the input to keep the parameters within tractable bounds. Two common methods include minmax scaling, where data are scaled between 0 and 1, and standardization, where data are scaled to have zero mean and unit variance. Though some networks have been shown to learn well without normalization[35] if the input data do not have a vast range, scaling is recommended to help convergence. Weight initialization also aids in convergence.[36] Techniques include the He and Glorot[36] initializations, where weights are drawn randomly from a normal or uniform distribution with predefined statistical moments. Moreover, dropout[37] is usually introduced between layers to randomly switch off a certain fraction of nodes so the network does not overfit the training data.

Gradient updates during backpropagation can either explode[38] or vanish if depth is too large;[22] ReLU and LSTM cells were developed primarily to overcome this problem. In addition, skip connections between deep layers help in propagating gradients backward, avoiding vanishing gradients.[39] These can be additive, e.g., ResNet,[39] or augmentative, e.g., UNet.[40] On the other hand, gradient clipping and use of the SELU[38] activation function have been proposed to solve the exploding gradient problem.

Deep networks have multiple hyperparameters that are not actively learned or updated during training but set based on the literature or using a systematic search process. Weights are successively updated backward from the output layer by computing the mean gradients from a batch of samples. This batch size is a hyperparameter, commonly set to 32, although it depends on

the architecture and hardware capabilities. Larger batch sizes demand more computation and memory for a single iteration, whereas smaller sizes imply slower convergence, e.g., Fernandez Rojas et al.[41] used 64 batches in each update, whereas Wickramaratne and Mahmud[42] used a batch size of 8. Another critical hyperparameter that controls the convergence rate is the learning rate that multiplies the loss-gradient in the gradient descent algorithm. Learning rate is usually set to values of the order of 10^{-2} at the start, and it can be decreased further for fine-tuning. For example, Ortega and Faisal[24] used an initial learning rate of 0.03, which decays at a factor of 0.9 after each epoch. An epoch is a point at which all nonoverlapping batches in the dataset have been exhausted for training. The number of epochs, also another hyperparameter, is decided strictly based on when the network begins to converge. The most common optimizer algorithm for gradient descent is Adam,[43] which uses an adaptive learning rate aided by momentum that helps network parameters to converge efficiently. Other algorithms also used in fNIRS applications are SGD[44] and RMSprop.[42]

A notorious problem with training DNNs is the change in the distribution of inputs in every layer, which calls for careful initialization of weights, learning rate schedule, and dropout. Ioffe and Szegedy[45] introduced a technique called batch-normalization that helps mitigate this problem termed as "internal covariance shift." Batch normalization reduced training steps by a factor of 14 times in the original study while requiring less stringent conditions of initialization and learning rates. Hence, during batch training in CNN, it is often recommended to use batch normalization layers after each convolution.[19,24]

Most biomedical applications have a limited number of subjects and limited training data (see Table 2 for a summary of the data set characteristics, including number of participants, number of channels, and cortical areas monitored for all studies summarized herein). Hence, it is essential to take proper measures to avoid using a large network with a small dataset to prevent overfitting. The network will reduce the training error but lose the ability to generalize to unseen datasets. To mitigate this, it is good to start with the simplest network possible, have fewer weights, and iteratively improve the networks by adding the number of layers/nodes. Furthermore, overfitting can be avoided by regularizing weights (which adds a regularization loss to overall loss function), adding dropout layers, etc.

Since fNIRS data are sequential time series data, studies where segments of obtained data, such as resting state,[46,47] are sufficient for analysis utilize a sliding window approach. Here, a fixed-length data segment is extracted at fixed intervals, allowing the overlap between subsequent windows. This is not applicable if the entire trial duration has to be analyzed, in which case each trial will have to be a single sample.

2.2.2 *Model evaluation*

Although various modeling tools are available for the analysis of collected data in psychological and behavioral science, there is a growing concern about reproducibility of results using the settings reported in studies.[91] Although direct replication or even conceptual replication[92] might not be possible for many neuroimaging studies involving human subjects, studies now rely on simulated replication as the next best approach.[93] This entails partitioning the data into a number of subsets and assessing model performance on each of the subsets after it is trained with the rest of the data. Performance metrics on n-subsets, a.k.a test sets, are averaged to get the mean performance metric, which is representative of the expected model performance on any unseen subset of data. Since the test set is "held out" from training, this method, also known as "cross-validation" (CV), can be deemed as a crude measure of the generalizability of the model.[93]

The type of subsets chosen depends on experimental settings, research question, or the limitations of the dataset. The subsets selected can either be a shuffled subset of trials, one single trial, or, if available, trials of each participant subject. These CV schemes are called k-fold CV, leave-one-user-out (LOUO) CV, and leave-one-subject-out (LOSO) CV. Leave-one-super-trial-out is another available rigorous CV technique.[94] Many of the reviewed papers carry out a 10-fold CV, whereas a few execute LOSO CV.[48,49] The most common metrics used for evaluation are accuracy, specificity, and sensitivity.

Table 2 Description of the data collected in each paper that was considered.

Citation no.	Author	Year	Number of participants	Approximate time recorded per participant	Number of channels	Sampling rate	Region of brain	Dataset
	Name of the author	*Published year*	i.e.	i.e., 15 to 30 min	i.e.	Sampling rate in Hz (resampled rate in Hz)	*Part of the brain being observed*	*Name of the public dataset*
19	Mirbagheri et al.	2020	10	10 min	23	10	Prefrontal cortex	N/A
20	Tanveer et al.	2019	13	30 min	28	1.81	Prefrontal and dorsolateral prefrontal cortex	N/A
21	Dolmans et al.	2021	22	77 to 345 min	27	10	Not specified	N/A
23	Saadati et al.	2019	26 and 29	60 to 180 min	24 to 36	10	Not specified	Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset
24	Ortega and Faisal	2021	12	13 min	24	12.6	Sensorimotor cortex	N/A
25	Dargazany et al.	2019	10	12.5 min	80	7.8125	Motor cortex	N/A
26	Nagasawa et al.	2020	9	40 min	41	10	Sensorimotor regions	N/A
28	Ghonchi et al.	2020	29	<30 min?	36	10 (128)	Not specified	Open access dataset for EEG+ NIRS single-trial classification
29	Trakoolwilaiwan et al.	2017	8	1000 s	34	25.7	Motor cortex	N/A
30	Janani et al.	2020	10	60 min	20	15.625	Motor cortex	N/A
31	Yoo et al.	2021	18	19 min	44	1.81	Auditory cortex	N/A
32	Gao et al.	2020	13	3 to 10 h	12	Not specified	Prefrontal cortex	N/A
33	Ortega et al.	2021	10	25 minutes	24	12.5	bilateral sensorimotor cortex	N/A

Table 2 (*Continued*).

Citation no.	Author	Year	Number of participants	Approximate time recorded per participant	Number of channels	Sampling rate	Region of brain	Dataset
	Name of the author	*Published year*	i.e.	i.e., 15 to 30 min	i.e.	Sampling rate in Hz (resampled rate in Hz)	*Part of the brain being observed*	*Name of the public dataset*
34	Gao et al.	2020	30	<25 min	32	Not specified	Prefrontal and primary motor cortex and supplementary motor area	N/A
35	Ma et al.	2021	36	200 s	31	7.14	Not specified	N/A
41	Fernandez Rojas et al.	2021	18	3 to 5 min?	24	10	Somatosensory cortex	N/A
42	Wickramaratne Mahmud	2021	29	<30 min?	36	13.3	Not specified	Open access dataset for EEG+ NIRS single-trial classification
44	Sirpal et al.	2019	40	30 to 180 min	Not specified	19.5	Bilateral anterior, middle and posterior temporal regions and frontopolar, frontocentral and dorsolateral frontal regions	N/A
46	Xu et al.	2019	47	8 min	44	14.29	Bilateral inferior frontal gyrus and temporal cortex	N/A
47	Yang et al.	2020	24	27 min	48	8.138	Prefrontal cortex	N/A
48	Takagi et al.	2020	15	2 min	22	10	Prefrontal cortex	N/A
49	Lu et al.	2020	8	12 to 16 min	52	10	Prefrontal cortex	Single-trial classification of antagonistic oxyhemoglobin responses during mental arithmetic
50	Saadati et al.	2019	26	33 min	36	10.4	Not specified	Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset

Table 2 (*Continued*).

Citation no.	Author	Year	Number of participants	Approximate time recorded per participant	Number of channels	Sampling rate	Region of brain	Dataset
	Name of the author	*Published year*	i.e.	i.e., 15 to 30 min	i.e.	Sampling rate in Hz (resampled rate in Hz)	*Part of the brain being observed*	*Name of the public dataset*
51	Benerradi et al.	2019	11	30 min	16	2	Prefrontal cortex	N/A
52	Liu et al.	2021	18	6 min	8	11.8	Anterior prefontal cortex	N/A
53	Lee et al.	2018	6	30 min	40	Not specified	Supplementary motor area and primary motor cortex	N/A
54	Kim et al.	2022	42	540 s	8	10	Bilateral prefrontal areas	N/A
55	Wickramaratne and Mahmud	2021	30	50 min	20	Not specified	Motor regions	Open-access fNIRS dataset for classification of unilateral finger- and foot-tapping
56	Woo et al.	2020	11	7 min	36	11	Left motor cortex	N/A
57	Hennrich et al.	2015	10	37 min	8	10	Prefrontal cortex	N/A
58	Kwon and Im	2021	18	12 min	16	13.3	Prefrontal cortex	N/A
59	Ho et al.	2019	16	90 min	7	18	Prefrontal cortex	N/A
60	Asgher et al.	2020	15	22 min	12	8	Prefrontal cortex	N/A
61	Naseer et al.	2016	7	440 s	16	1.81	Prefrontal cortex	N/A
62	Hakimi et al.	2020	20	10 min	23	10	Prefrontal cortex	N/A
63	Erdoğan et al.	2019	11	10 min	48	3.91	Frontal cortex, primary motor cortex and somatosensory motor cortex	N/A

Table 2 (*Continued*).

Citation no.	Author	Year	Number of participants	Approximate time recorded per participant	Number of channels	Sampling rate	Region of brain	Dataset
	Name of the author	*Published year*	i.e.	i.e., 15 to 30 min	i.e.	Sampling rate in Hz (resampled rate in Hz)	*Part of the brain being observed*	*Name of the public dataset*
64	Hamid et al.	2022	9	6 min	12	1.81	Left hemisphere of M1	N/A
65	Khan et al.	2021	28	350 s	48	3.9	Frontal, frontal-central and central sulcus and central and temporal-parietal lobes	N/A
66	Ortega and Faisal	2021	9	5 min	24	12.5(80)	Bilateral sensorimotor cortex	N/A
67	Zhao	2019	47	—	24	Not specified	Primary motor and prefrontal region	N/A
68	Ghonchi et al.	2015	29	<30 min?	36	10(128)	Not specified	Open access dataset for EEG+ NIRS single-trial classification
69	Chiarelli et al.	2018	15	10 min	16	10	Sensorimotor regions	N/A
70	Cooney et al.	2021	19	2 h	8	10(250)	Bihemispheric motor regions	N/A
71	Sun et al.	2020	29	9 min	36	12.5	Not specified	Open access dataset for EEG+ NIRS single-trial classification
72	Kwak et al.	2022	29	9 min	36	12.5	Not specified	Open access dataset for EEG+ NIRS single-trial classification
73	Khalil et al.	2022	26	62 s	36	10.4	Frontal, motor cortex, parietal, and occipital regions	Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset
74	Xu et al.	2020	47	8 min	52	14.3	Bilateral temporal lobe	N/A
75	Xu et al.	2020	47	8 min	44	14.3	Bilateral temporal lobe	N/A

Table 2 (*Continued*).

Citation no.	Author	Year	Number of participants	Approximate time recorded per participant	Number of channels	Sampling rate	Region of brain	Dataset
	Name of the author	*Published year*	i.e.	i.e., 15 to 30 min	i.e.	Sampling rate in Hz (resampled rate in Hz)	*Part of the brain being observed*	*Name of the public dataset*
76	Ma et al.	2020	84	2 min	52	10	Bilateral frontal and temporal cortices	N/A
77	Wang et al.	2021	96	150 min	53	100	Prefrontal cortex	N/A
78	Chao et al.	2021	32	15 min	22	7.81	Prefrontal cortex	N/A
79	Chou et al.	2021	67	160 s	52	10	Bilateral frontotemporal regions	N/A
80	Rosas-Romero et al.	2019	5	30 to 180 min	104 to 146	19.5	Full scalp recording	N/A
81	Yang et al.	2019	24	15 min	48	8.138	Prefrontal cortex	N/A
82	Yang and Hong	2021	24	5 min	48	8.138	Prefrontal cortex	N/A
83	Ho et al.	2022	140	30 min	6	8	Prefrontal cortex	N/A
84	Behboodi et al.	2019	10	9.5 min	52	18.51	Sensorimotor and motor areas	N/A
85	Sirpal et al.	2021	40	75 min	138	19.5	Full scalp recording	N/A
86	Bandara et al.	2019	20	17 min	52	10	Frontal region	N/A
87	Qing et al.	2020	8	18 min	12	15.625	Prefrontal cortex	N/A
88	Hiwa	2016	22	6.5 min	24	10	Left hemisphere	N/A
89	Andreu-Perez et al.	2021	30	7.5 min	16	Not specified	Prefrontal cortex	N/A
90	Ramirez et al.	2022	5	14 min	16	Not specified	Left and right frontal lobes	N/A

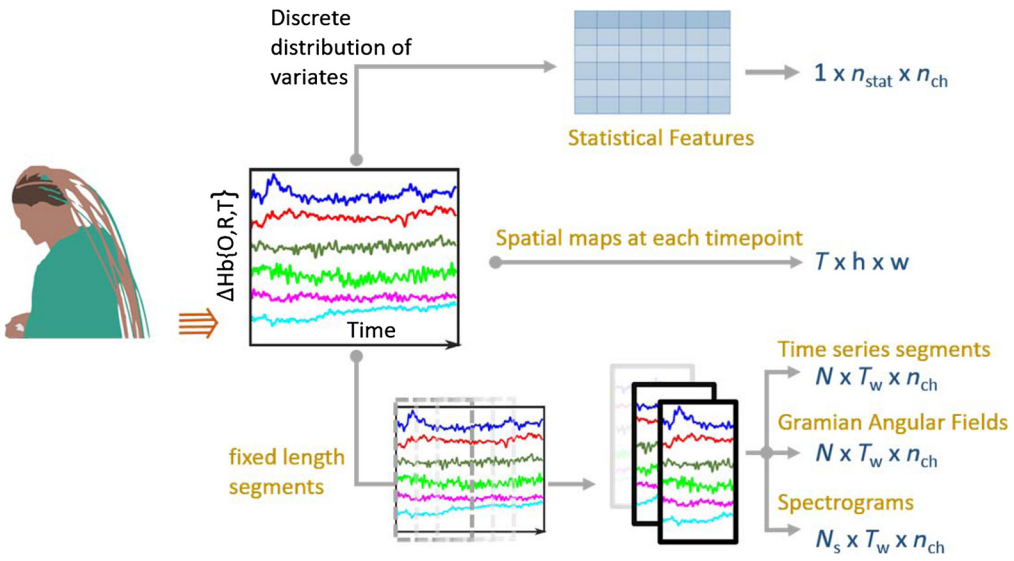

Fig. 2 Extraction of samples to be used for training NNs. Time-series data denoting hemodynamic concentration changes are obtained from the raw data after initial analysis and changed to appropriate formats (shown in right in the form of dimensions of input data samples), based on chosen architecture. *N*, *Ns*, number of samples; *T*, *Tw*, number of timepoints; n_{ch}, number of channels; *h*, *w*, height and width of spatial map images; n_{stat}, number of statistical moments/features.

2.2.3 *Inputs to deep networks*

After the preprocessing steps (see Sec. 2.1), changes in oxy (ΔHbO) and deoxy (ΔHbR) hemoglobin, as well as total change (ΔHbT) are obtained in time-series format. From these, data samples are segmented based on task settings. Data are segmented trialwise for task-based experiments, and for resting state data or long trials, data are segmented using sliding windows as mentioned previously.

Many studies opt for the discrete probability distribution of concentration changes and extract statistical features such as mean, slope, variance, skewness, kurtosis, max, and range in the form of manual features. In other cases, where the network is allowed to learn and extract features itself, the data are fed in the forms of either spatial maps[20,28,50] or time series themselves. In some studies, segments of data are converted to other forms such as Gramian angular fields[32] or spectrogram maps.[30] A schematic in Fig. 2 summarizes the sample extraction procedure employed in the studies. While the general trends and techniques used for DL fNIRS studies have been examined, the applications and some application-dependent techniques are further discussed in the next section.

3 DL Applications in fNIRS

In this section of the review paper, we summarize and discuss the key findings of the literature search as well as how the techniques discussed above are currently being used. To properly understand the scope of this review, the literature search methodology is first described. The primary method of searching for relevant articles was via the PubMed search engine. Papers published between 2015 and May 2022 in which DL was used in conjunction with fNIRS data were considered for this review. Due to very few fNIRS papers being published using DL prior to 2015, the search was restricted to this time frame. Using the terms "deep learning" and "fNIRS" resulted in 35 results, 29 of which were relevant, while the terms "deep learning" and "NIRS" only produced 12 articles, none of which were both relevant and absent from the previous search. When searching the terms "neural network" and "fNIRS," 146 results were found, with many using the term neural network to refer to the neuroscience phenomenon being studied through the use of DL. As a result only three additional relevant papers were found via this search. In addition to this, using the same search terms in Google Scholar produced tens of thousands of

results, many of which were not relevant. Due to the impracticality of a comprehensive search of these results, the search was truncated after multiple pages of results yielded no relevant articles. From this Google Scholar search, an additional 25 articles were found, yielding a total of 63 articles that were considered in this review.

In recent years, the use of DL techniques in fNIRS studies has increased, and due to the versatility of fNIRS, DL has been applied to many different applications of fNIRS. While some of the studies that used DL used it for feature extraction or data augmentation, in most of the papers considered, DL was used as a classifier. As a result, those studies in which DL were used as a classifier are further subdivided into categories based on the application of fNIRS in the study. A comprehensive summary of the applications and DL architecture employed is Table 3, while the details of the experimental setups for the fNIRS studies are summarized in Table 2. Because fNIRS is of interest in studies on BCI, many of the studies found used DL classifiers for the classification of tasks for BCI applications. Other studies have used DL techniques as diagnostic tools, to detect various physiological and mental pathologies based on cortical activity. Finally, some studies, such as those using DL techniques to assess skill level and functional connectivity, were not common enough to be placed into a category of their own; however, these papers all focused on the analyses of cortical activity using DL techniques, and as such are grouped together. Figure 3 shows provided for visualization of the number of papers collected from each category for the given year. While some of the papers mentioned may fall within multiple categories, additional context from the paper such as the primary focus of the paper assisted in determining how the paper was categorized. This did not stop relevant papers from being discussed in more than one subcategory.

3.1 *Preprocessing*

Like most noninvasive neuroimaging modalities, raw fNIRS signals typically contain confounding physiological signals and other noise that originate from outside of the cerebral cortex, such as the hemodynamics of the scalp and changes in blood pressure and heart rate.[95] Most studies use some method of preprocessing to try to address this. While many of the papers presented here use band pass filtering or butterworth filtering, various other methods are employed throughout the literature. Independent component analysis (ICA) denoising,[12] wavelet filtering,[29] and correlation-based signal improvement filtering[51] are all methods used to remove some of the undesired physiological trends in fNIRS signals. Another recommended method is short separation regression, a method in which signals with only confounding physiological signals are simultaneously collected alongside fNIRS signals and the trends in these signals are removed from the fNIRS data.[96] Other algorithms such as Savitsky–Golay filtering[97] and temporal derivative distribution repair[98] are methods used to correct motion artifacts and baseline drifts in fNIRS signals. Many of these techniques are commonly used in fNIRS studies to improve the quality of the signal as well as ensure that the signal being assessed originates from the brain. Other recommended techniques involve prewhitening the data or decorrelating via PCA to remove any correlation between fNIRS signals prior to analysis,[96] further facilitating the analysis of signals originating in a region of interest.

3.2 *Feature Extraction and Data Augmentation*

One of the most notable problems with fNIRS data is the extensive manual feature extraction and artifact removal typically required for data to be analyzed, preventing many fNIRS studies from being applied in real time. As a result, more effective methods of preprocessing fNIRS data are being explored. One of the most promising benefits of DL is the ability to quickly and automatically learn and extract relevant features in fNIRS data. Some studies have already explored this benefit of DL. Tanveer et al.[20] reported the use of a DNN to extract the features that were fed to a K-nearest neighbors (KNN) classifier to detect the drowsiness of subjects during a virtual driving task. Using the features extracted from the DNN, the KNN was able to achieve a classification accuracy of 83.3%. Despite feature extraction typically being computationally expensive, even taking hours with a powerful GPU, the DNN exhibited a mean computation time of 0.024 s for 10 s time windows, a speed that would allow for feature extraction of a 30-min signal

Table 3 The applications and key findings of each paper considered.

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
19	Mirbagheri et al.	2020	CNN	BCI	Statistical features	Stress versus relaxation	Fivefold CV	Task labels for each trial	88.52 ± 0.77% accuracy
20	Tanveer et al.	2019	MLP	Cortical Analysis	segmented time series	Extracted Features to Feed to Classifier	10-Fold CV	Drowsiness detected via change in facial expression	83.3 ± 7.4% Accuracy with KNN Classifier
21	Dolmans et al.	2021	CNN+LSTM+ MLP	BCI	segmented time series	Mental Workload Level	5-Fold CV	Participant Reported Difficulty Ratings	32% Accuracy
23	Saadati et al.	2019	DNN	BCI	segmented time series	n-back, WG, DSR and MI vs relaxation	LOOCV		89% Accuracy
24	Ortega and Faisal	2021	HEMCNN	BCI	Statistical features	Left-hand gripping or right-hand gripping	Fivefold CV	Task labels for each trial	78% accuracy
25	Dargazany et al.	2019	MLP	BCI	Raw fNIRS data	Right hand, left hand, left leg, right leg, and both hand ME		Task labels for each trial	77% to 80% accuracy
26	Nagasawa et al.	2020	WGAN	BCI	Raw fNIRS data	Left-hand ME, right-hand ME, bimanual ME or rest	10-fold CV	N/A for data generation/ task labels for each trial	73.3% accuracy for augmented SVM
28	Ghonchi et al.	2020	RCNN	BCI	Three-rank tensors of upsampled fNIRS time series	Mental arithmetic or rest/ motor imagery or rest	k-fold CV	Task labels for each trial	99.63% accuracy
29	Trakoolwilaiwan et al.	2017	MLP/CNN	BCI	Segmented time series	Left-hand MI right-hand MI or Rest	10-fold CV	Task labels for each trial	89.35% accuracy for MLP and 92.68% accuracy for CNN

Table 3 (*Continued*).

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
30	Janani et al.	2020	MLP/CNN	BCI	Spectrograms+sample-point images/ segmented time series	Left- or right-hand MI/ME	Fivefold CV	Task labels for each trial	80.49 ± 6.66% accuracy (MI) and 85.66 ± 8.25% accuracy (ME)
31	Yoo et al.	2021	LSTM	BCI	Time series	English, non-English, annoyance, natural sound, music, or gunshot	Sixfold CV	Task labels for each trial	20.38 ± 4.63% accuracy
32	Gao et al.	2020	CNN	Cortical analysis	Statistical features	Expert or novice	LOUO+10-fold CV	Reported skill level of participants	91% accuracy, 95% sensitivity and 67% specificity
33	Ortega et al.	2021	CNNATT	BCI	Upsampled fNIRS data	Discrete force profiles		Simultaneously recorded grip force	55.2 FFAV%
34	Gao et al.	2020	CNNIRS	Preprocessing/ augmentation	Simulated/real HRF	Denoised HRF		Simulated HRF signals	3.03 MSE
35	Ma et al.	2021	CNN	BCI	Time series	Left-hand MI or right-hand MI	LOSO CV	Task labels for each trial	98.6% accuracy with FCN and ResNet
41	Fernandez Rojas et al.	2021	LSTM	Diagnosis	Raw HbO data	Low-cold, low-heat, high-cold or high-heat	10-fold CV	Task labels for each trial	90.6% accuracy 84.6% sensitivity 90.4% specificity
42	Wickramaratne Mahmud	2021	CNN	BCI	Statistical features	MI MA or rest	10-fold CV	Task labels for each trial	87.14 ± 3.20% accuracy
44	Sirpal et al.	2019	LSTM	Diagnosis	Time series	Epileptic siezure versus normal	10-fold CV	Labeled siezure and nonseizure segments	98.3 ± 0.4% accuracy, 89.7 ± 0.5% recall, 87.3 ± 0.0.8% precision

Table 3 (*Continued*).

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
46	Xu et al.	2019	CNN+GRU	Diagnosis	Segmented time series	ASD or TD		Diagnosis of subject	92.2% accuracy, 85.0% sensitivity, and 99.4% specificity
47	Yang et al.	2020	CNN	Diagnosis	Spatial maps from ΔHbO signals and spatiotemporal maps from statistical features	Cognitive impairment or healthy	Five-fold CV	Diagnosis of subject	90.37 ± 5.30% accuracy, 86.98 ± 7.25% recall, 82.19 ± 9.93% precision for VFT
48	Takagi et al.	2020	CNN	Cortical analysis	Oxy, deoxy and OD images	Teeth clenching or relaxed	Fivefold CV	Task labels for each trial	90.3 ± 6.5% accuracy, 88.1 ± 10.8% recall, 92.4 ± 7.8% specificity, 92.5 ± 7.1% precision
49	Lu et al.	2020	LSTM+CNN	BCI	Statistical features	Mental arithmetic or Rest	Fivefold CV	Task labels for each trial	95.3% accuracy
50	Saadati et al.	2019	CNN	BCI	Topographical activity Maps	0-back 2-back or 3-back task	10-fold CV	Task labels for each trial	97 ± 1% accuracy
51	Benerradi et al.	2019	CNN	BCI	statistical features	Mental workload level	LOUO CV	Task labels for each trial	49.53% accuracy for three classes and 72.77% Accuracy for two classes
52	Liu et al.	2021	ESN/CAE	Preprocessing/ augmentation	Segmented time series	0-back, 1-back, 2-back, or 3-back task	10x10 CV	Task labels for each trial	52.45 (ESN) and 47.21% (CAE) accuracy for four classes
53	Lee et al.	2018	MLP	Preprocessing/ Augmentation	Time series	Denoised time Series		Wavelet denoised methods	CNR of 0.63

Table 3 (*Continued*).

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
54	Kim et al.	2022	CNN	Preprocessing/ augmentation	Time series + HRF	Denoised time series	Ablation	Simulated HRF Signals	MSE of approx 0.004-0.005
55	Wickramaratne and Mahmud	2021	GAN+CNN	Preprocessing/ augmentation	GASF/ kernel PCA GASF	GASF/motor task	LOOCV	N/A for data generation/ task labels for each trial	96.67% accuracy and 0.98 AUROC for CNN+ 110% generated data
56	Woo et al.	2020	DCGAN+ CNN	Preprocessing/ augmentation	HbO t-maps	ME or rest	—	Task labels for each trial	92.42% for unaugmented data and 97.17% for Augmented Data
57	Hennrich et al.	2015	MLP	BCI	Not specified	Mental arithmeetics, word generation, mental Rotation or Rest	10-fold CV	Task labels for each trial	64.1% accuracy
58	Kwon and Im	2021	CNN	BCI	Segmented time series	MA versus RELAXATION	LOSO CV	Task labels for each trial	71.20 ± 8.74% accuracy
59	Ho et al.	2019	DBN/CNN	BCI	Statistical features	Mental workload level	*n*-fold CV	Task labels for each trial	84.26 ± 2.58% accuracy for DBN without PCA and 75.59 ± 3.4% for DBN with PCA inputs and 72.77 ± 1.92% accuracy for CNN without PCA and 68.12 ± 3.26% with PCA inputs
60	Asgher et al.	2020	LSTM	BCI	Statistical features	Mental workload level	10-fold CV	Task labels for each trial	89.31 ± 3.95% accuracy 87.51 ± 3.90% precision 86.76 ± 4.38% recall

Table 3 (*Continued*).

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
61	Naseer et al.	2016	MLP	BCI	Statistical features	MA or Rest	Ablation	Task labels for each trial	96.3 ± 0.3% accuracy for MLP
62	Hakimi et al.	2020	CNN	BCI	Statistical features	Stress versus relaxation	Fivefold CV	Task labels for each trial	98.69 ± 0.45% accuracy for HRF feature set and 88.60 ± 1.15% accuracy for fNIRS feature set
63	Erdoğan et al.	2019	MLP	BCI	Statistical features	MI, ME or rest	Ablation	Task labels for each trial	96.3% ± 1.3% accuracy for ME versus rest, 95.8% ± 1.2% accuracy for MI versus rest and 80.1% ± 2.6% accuracy for ME versus MI
64	Hamid et al.	2022	CNN/LSTM	BCI	Time series	Motor execution or rest	10-fold CV	Task labels for each trial	79.73% accuracy with CNN, 77.21% accuracy with LSTM and 78.97% accuracy with bi-LSTM
65	Khan et al.	2021	MLP	BCI	Statistical features	Specific finger tapping or Rest	LOSO CV	Task labels for each trial	60 ± 2% accuracy
66	Ortega and Faisal	2021	CNNATT	BCI	Segmented time series	Force profiles	FiveFold CV	Simultaneously recorded grip Force	55% FVAF
67	Zhao	2019	BiLSTM	BCI	Statistical features	Goal execution versus completion	—	Task labels for each trial	71.70% accuracy

Table 3 (*Continued*).

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
68	Ghonchi et al.	2015	LSTM/CNN	BCI	Upsampled fNIRS data	MI classes	10 x 5-Fold CV	Task labels for each trial	99.6% accuracy
69	Chiarelli et al.	2018	MLP	BCI	Segmented time series	Left-hand MI or right-hand MI	10-fold CV	Task labels for each trial	83.28 ± 2.36% accuracy
70	Cooney et al.	2021	CNN	BCI	Filtered frequency bands of segmented time series	One of four action words + one of four word combinations	Nested fivefold CV	Task labels for each trial	46.31% accuracy for overt speech and 34.29% accuracy for imagined speech
71	Sun et al.	2020	CNN	BCI	Tensors of fused EEG and fNIRS data	Mental arithmetic/motor Imagery or Rest	Fivefold CV	Task labels for each trial	77.53% accuracy for MI and 90.19% accuracy for MA
72	Kwak et al.	2022	CNN+ Attention	BCI	Three-rank tensors of upsampled fNIRS time series	Mental arithmetic/motor Imagery or Rest	Ablation	Task labels for each trial	91.96 ± 5.82% accuracy for MA and 78.59 ± 5.82% accuracy for MI
73	Khalil et al.	2022	CNN	BCI	Time series	Mental workload level	10-fold CV	Task labels for each trial	94.52% accuracy
74	Xu et al.	2020	LSTM	Diagnosis	Time series	ASD or TD	10-fold CV	Diagnosis of subject	95.7 ± 4.99% accuracy
75	Xu et al.	2020	CNN+ Attention	Diagnosis	segmented time series	ASD or TD	10-fold CV	Diagnosis of subject	93.3% accuracy 90.6% sensitivity and 97.5% specificity
76	Ma et al.	2020	Attention LSTM+CNN	Diagnosis	Normalized HbO, HbR, and HbT matrices	BD or MDD	k-fold CV	Diagnosis of subject	96.2% accuracy

Table 3 (*Continued*).

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
77	Wang et al.	2021	CNN	Diagnosis	Time series/statistical features	Depressed or nondepressed	Ablation	Diagnosis of subject	83% Accuracy 79% Precision and 83% recall (manually extracted features), 72% accuracy, 80% precision and 75% recall (raw data)
78	Chao et al.	2021	CFNN/RNN	Diagnosis	Statistical features	Fear stimulus or rest	LOSO CV	Task labels for each trial	99.94% accuracy with CFNN and 99.94% with RNN
79	Chou et al.	2021	MLP	Diagnosis	Statistical features	FES or Healthy	Sevenfold CV	Diagnosis of subject	79.7% accuracy, 88.8% specificity and 74.9% specificity
80	Rosas-Romero et al.	2019	CNN	Diagnosis	3-dimensioal tensors of HbO and HbR	Pre-ictal versus inter-ictal	Fivefold CV	Labeled siezure and nonseizure segments	99.67 ± 0.75% accuracy for CNN
81	Yang et al.	2019	CNN	Diagnosis	Activation t-map / channel correlation map	Cognitive impairment or Healthy	Sixfold CV	Diagnosis of subject	90.62% accuracy with t-maps and 85.58% with correlation maps
82	Yang and Hong	2021	CNN	Diagnosis	Connectivity map	Mean STD and variance of Δ HbO and Δ HbR		Diagnosis of subject	97.01% accuracy
83	Ho et al.	2022	LSTM/LSTM+ CNNN	Diagnosis	Segmented time series	Healthy, asymptomatic, prodromal or dementia AD	Fivefold CV	Diagnosis of subject	86.8% accuracy for CNN-LSTM and 84.4% accuracy for LSTM

Table 3 (*Continued*).

Citation no.	Author	Year	DL architecture	General task	Input	Output	Validation metrics	Ground truth	Results
	Name of the author	*Published year*	i.e., CNN, LSTM, etc.	i.e., diagnosis	i.e., Hbr, Hbo	*What is being output by the network*	i.e., LOSO, LOUO	*What was being used to assess the models*	*Put important numbers here.*
84	Behboodi et al.	2019	MLP/CNN	Cortical analysis	Time series	RSFC estimation	—	Channels anatomically located over the motor and sensorimotor cortex	0.89 AUC for MLP and 0.92 AUC for CNN
85	Sirpal et al.	2021	LSTM AE	Cortical analysis	Full spectrum EEG	ΔHbO concentration	k-fold CV	Real fNIRS data	6.52 x 10^-2 mean reconstruction error (Euclidean)
86	Bandara et al.	2019	CNN+LSTM	Cortical analysis	segmented time series	High/low valence+ high/low arousal or neutral	Fivefold CV	Emotional valence and arousal scores of DEAP Dataset	70.18% accuracy for 1s windows and 77.29% accuracy for 10 s windows
87	Qing et al.	2020	CNN	Cortical analysis	Segmented time series	Like/dislike, like/so-so, or dislike/so-so	Eightfold CV	Subject ratings for each trial	84.3, 87.9, and 86.4% accuracy for 15s, 30s, and 60s, respectively
88	Hiwa	2016	CNN	Cortical analysis	Time series	Male or female	LOOCV	Gender of subject	Approx. 60% accuracy in five best channels
89	Andreu-Perez et al.	2021	DCAE/MLP	Cortical analysis	Statistical features	Novice, Intermediate, or Expert	10 repeated stratified k-fold CV with 5 splits	Reported skill level of participants	91.43 ± 6.32% accuracy for DCAE, 91.44 ± 9.97% accuracy for MLP
90	Ramirez et al.	2022	CNN	Cortical analysis	fNIRS/fNIRS+EEG images	Discrete preference ratings		Subject ratings for each trial	66.86% accuracy with fNIRS and 91.83% accuracy with fNIRS+EEG

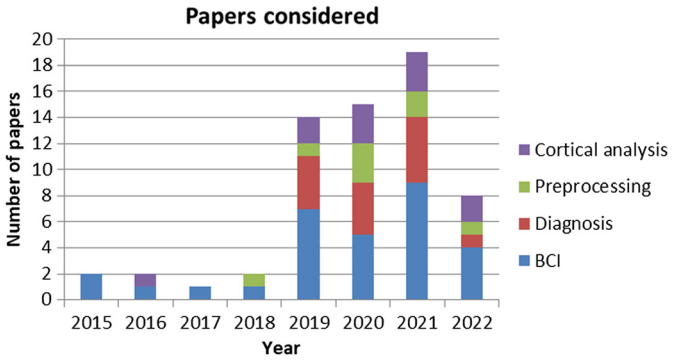

Fig. 3 Distribution of the 63 papers reviewed in this article by year and color coded by application field.

to take <5 s with a NVidia 1060 GTX GPU. On top of computational speed, the ability of DL to automatically learn and extract features may reduce bias and errors during feature extraction, allowing for an increase in classification accuracy. One study by Liu et al.[52] used an echo state network autoencoder (ESN AE) to extract the features that were fed to an MLP, achieving a four-class classification accuracy of 52.45%, outperforming the accuracy of the convolutional autoencoder (CAE) + CNN and manually extracted features fed to an MLP, which achieved accuracies of 47.21% and 37.94%, respectively.

While there have been other papers interested in using DL to extract features to feed into another classifier, there have also been papers that take raw fNIRS data and use the same neural network for feature extraction and classification. Despite end-to-end neural networks being seen as a more ideal solution than manual feature extraction, difficulties with low generalizability make them less commonly used. One study by Dargazany et al.[25] used an MLP (with two hidden layers) with raw EEG, body motion and fNIRS data to achieve a reported classification accuracy of 78% to 80% on a motor task with four classes, despite no denoising or preprocessing of data being done. Another study by Fernandez Rojas et al.[41] used raw fNIRS data as the input for an LSTM network, achieving a classification accuracy of 90.6%. To assess generalizability, a 10-fold CV was used, with the classifier achieving an accuracy of 93.1%. Both studies have provided evidence toward the claim that DL techniques are capable of removing the need for manually extracted features from fNIRS data, further progressing toward the real-time end-to-end BCI. Despite the fact that the previously mentioned studies were able to achieve high accuracies without denoising or motion artifact removal, some studies are performed when a subject's head is in motion. In such studies, fNIRS data can be heavily compromised by specific artifacts in the raw data that could bias any classification task. While many algorithms are used to try to remove motion artifacts, Lee et al. attempted to use a CNN to recognize and remove motion artifacts without relying on the parameters that must be defined to use many of the popular motion artifact removal algorithms.[53] In this study, the raw fNIRS time series and the estimated canonical response were used as inputs to the network and the resulting CNR of the DL output was 0.63, outperforming wavelet denoising, which achieved a mean CNR of 0.36. Another study done by Gao et al.[34] used subjects who were performing a precision cutting surgical task based on the fundamentals of laparoscopic surgery (FLS) program, which required a large range of motion. With a DAE and the process shown in Fig. 4, 93% motion artifact removal in simulated data and 100% artifact removal in real data were reported, outperforming all comparable artifact removal techniques, including wavelet filtering and principal component analysis. Another study to look at DL for the removal of motion artifacts, Kim et al.[54] compared a CNN to the performance of wavelet denoising and an autoregressive denoising method. Using simulated data in a manner similar to Gao et al. to determine the ground truth, Kim et al. found that the CNN resulted in a mean square error (MSE) of ~0.004 to 0.005, while the next best method, the combination of wavelet and autoregressive denoising, resulted in an MSE of ~0.009. Despite these studies all using different metrics to measure the effectiveness of the network, there is clearly an interest in finding an effective method of removing motion artifacts from fNIRS data.

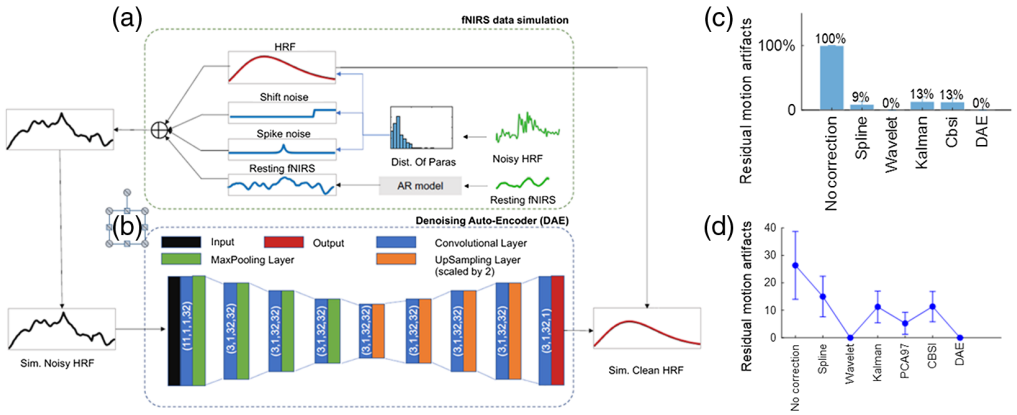

Fig. 4 The illustration of the fNIRS data simulation process and the designed DAE model. (a) The green lines are the experimental fNIRS data, including noisy HRF and resting fNIRS data, while the blue and red lines are simulated ones. (b) DAE model: The input data of the DAE model are the simulated noisy HRF, and the output is the corresponding clean HRF without noise. The DAE model incorporates nine convolutional layers, followed by max-pooling layers in the first four layers and upsampling layers in the next four layers, with one convolutional layer before the output. The parameters are labeled in parentheses for each convolutional layer, in the order of kernel size, stride, input channel size, and kernel number. (c), (d) number of residual motion artifacts for the simulated and experimental data sets, respectively. Adapted from Ref. 34.

It is clear that DL techniques have shown promise for the ability to process and extract features from fNIRS data, but due to a lack of large variety of open-source fNIRS datasets, there has been a recent interest in using DL to generate more fNIRS data for training models. Data augmentation is a technique in which new data are generated to reduce the need for large labeled datasets for many machine learning and DL algorithms that require a lot of labeled data. To be useful, this generated data must not be identical to any of the training data, but must also be realistic, i.e., it must remain within the distribution of the original dataset.[55] While this is a challenge, some DL techniques such as generative adversarial networks (GANs), have been used to accomplish this. Wickramaratne and Mahmud[55] have used a GAN to augment their fNIRS dataset to increase the classification accuracy of finger- and foot-tapping tasks. Without augmented data, a classification of 70.4% was achieved with an SVM classifier, and a CNN classifier achieved an accuracy of 80% when trained only on real data. Using a GAN to generate training data, the accuracy of the CNN classifier increased to 96.67% when trained on real data as well as 110% generated data. Similarly, Woo et al.[56] used a GAN to produce activation t-maps, which, when used to augment the training dataset of a CNN, increased the classification accuracy of a finger tapping task from 92% to 97%, showing that a network that is already performing well may benefit from data augmentation. While the previous studies have used GANs to generate an image representation of fNIRS data, only one study directly used a GAN to augment the dataset with raw time series fNIRS data. Nagasawa et al.[26] used a GAN to generate fNIRS time series data to increase the classification accuracy of motor tasks, as shown in Fig. 5. They reported that when augmenting the 16 original datasets with 100 generated datasets, the accuracy of the SVM classifier increased from around 0.4 to 0.733 while the accuracy of the neural network classifier increased from around 0.4 to 0.746. While still a relatively recent development in fNIRS studies, generated datasets using GANS have demonstrated the ability to increase the classification accuracy of commonly used classifiers such as CNN or SVM classifiers, once again demonstrating the versatility of DL techniques in fNIRS research.

3.3 *Brain–Computer Interface*

One of the most promising applications for fNIRS research is BCI. Many studies on BCI use traditional machine learning or DL techniques to identify a certain task based on cortical activation. Hennrich et al.[57] used a DNN to classify when subjects were performing mental arithmetic, word generation, mental rotation of an object and relaxation. The reported accuracy of

the DNN with two hidden layers was 64.1%, which is comparable to the 65.7% accuracy of the shrinkage LDA despite the LDA using custom-built features while the input for the DNN was denoised and normalized fNIRS data. Kwon and Im[58] used similar inputs for their CNN, denoised and baseline corrected ΔHbR and ΔHbO_2 data, to classify mental arithmetic from an idle fixation task. This CNN achieved a classification accuracy of 71.20%, which surpassed the 65.74% accuracy of the shrinkage LDA classifier that used feature vectors as inputs. Wickramaratne and Mahmud[42] also used a CNN to classify mental arithmetic from an idle fixation task in 2021. Like Kwon and Im, a shrinkage LDA with feature vectors was used as the input. Unlike the previous study however, the inputs for the CNN were Gramian angular summation fields (GASF), which are a type of image that is constructed from time series data that maintains some temporal correlation between points. With this CNN and GASF inputs, a classification accuracy of 87.14% was achieved, once again outperforming the shrinkage LDA, which achieved an accuracy of 66.08%. Similarly, Ho et al.[59] also used a two-dimensional (2D) representation of fNIRS data to try and discriminate between differing levels of mental workload. In this study, Ho et al. found that using a CNN with spectrograms generated from fNIRS data achieved a classification accuracy of 82.77%. This was outperformed by a deep belief network, which is a type of network similar to an MLP. The deep belief network, using manually extracted features from the HbR and HbO_2 signals achieved an accuracy of 84.26%. In a similar mental workload task, Asgher et al.[60] found that using an LSTM with similarly extracted features from HbR and HbO_2 time signals resulted in a mental workload classification accuracy of 89.31%. While many of the studies presented compared the performance of DL techniques to that of traditional machine learning or other algorithms, Naseer et al.[61] compared the classification accuracy of an MLP to that of a kNN, Naive Bayes, SVM, LDA, and QDA algorithm. On a two-class mental workload task, the MLP achieved an accuracy of 96% while the QDA, Naive Bayes, and SVM classifiers all achieved similar accuracies of about 95% and the LDA and kNN algorithms performed much worse, with accuracies of 80% and 65%, respectively. All classifiers in this study used manually extracted features such as skewness and kurtosis of the fNIRS as inputs

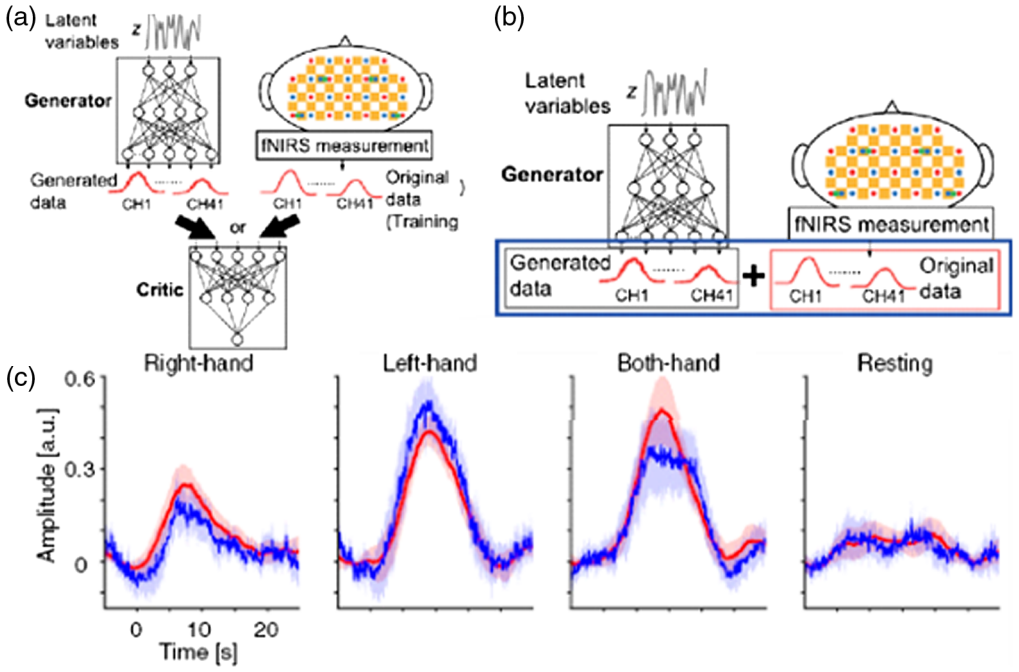

Fig. 5 Framework of GAN and data augmentation. (a) The generator creates the data from the random variables *z*, and the critic evaluates the generated and original (measured) data. (b) After the training process, the data generated by the generator (referred to as generated data) are combined with the original fNIRS data as augmented data. (c) Trial-averaged waveforms for the four tasks considered in a CV hold. The red lines denote measured original data and the blue lines denote generated data using WGANs. The shaded area represents 95% confidence intervals. Adapted from Ref. 26.

into the algorithms. Hakimi et al.[62] also used manually extracted features from fNIRS time series to perform two-class classification between stress and relaxation states and with a CNN, achieved an accuracy of 98.69%, further reinforcing that DL can give very high classification accuracies with mental workload tasks.

While the classification of mental tasks such as arithmetic or word generation is commonly used, many forms of BCI are designed to help those who have restricted or reduced motor function. Because of this, another popular type of experiment found in BCI studies involves executing motor tasks. Trakoolwilaiwan et al.[29] were able to achieve an accuracy of 92.68% with a CNN in a three-class test to distinguish the finger tapping of the left hand, right hand, and both hands at rest despite the CNN being used as both a feature extractor and classifier. The CNN outperformed the SVM and ANN classifiers, which reported accuracies of 86.19% and 89.35%, respectively, which were given the extracted feature inputs commonly used in BCI fNIRS studies (signal mean, variance, kurtosis, skewness, peak, and slope). Since much of the interest in BCI applications involve aiding those who lack the ability to move, some studies focus not on the cortical activations of movement but rather on the cortical activations of imagining movement. In one of these motor imagery studies, Janani et al.[30] had subjects perform a hand clenching or foot tapping task, and shortly after, imagine themselves performing the same task. Two different types of input images were used to see from which method a CNN classifier would more effectively extract features. The first type of input image turned all of the data points within a 20-s window into an $M \times N$ matrix, where M is the number of data points and N is the number of channels. The second type of input used a short-time Fourier transform to turn the one-dimensional data into 2D time-frequency maps of each channel that were stacked on top of each other to form an input image. One study by Erdoğan et al.[63] similarly performed classification between motor imagery versus motor execution using an MLP with a classification accuracy of 96.3% between finger tapping and rest and 80.1% accuracy between finger tapping and imagined finger tapping when using manually extracted features from fNIRS data. Hamid et al.[64] attempted to distinguish between a treadmill walking task and rest using bandpass filtered fNIRS data and an LSTM. The LSTM achieved an accuracy of 78.97%. When compared with traditional classifiers that had statistical features manually extracted, the kNN achieved the next best performance of 68.38% accuracy. The SVM and LDA were also outperformed by DL, achieving accuracies of 66.63% and 65.96%, respectively, despite using manually extracted features as inputs. This demonstrates the ability for DL to perform well on fNIRS data without requiring the extensive processing or feature extraction typically used in conjunction with other classifiers for fNIRS data. For many of these BCI motor tasks, being used for prosthetics would be more applicable. In these situations, more fine motor control using fNIRS would need to be assessed. Khan et al.[65] addressed this by performing six-class classification between rest and each finger on the right hand of the subjects, achieving an accuracy of 60%. Ortega and Faisal[24] attempted to distinguish between a left- and right-hand gripping task using a PCA to reduce dimensionality of the denoised time series data before feeding the segmented time series into a CNN-based architecture. The resulting accuracy of this study was 77%. To further investigate this, Ortega et al.[33] used a CNN with attention and simultaneously recorded EEG signals to try to reconstruct the grip force of each hand during the task. This resulted in an average fraction of variance accounted for of 55% when reconstructing the discrete grip force profiles, demonstrating that not only can the DL techniques distinguish which hand was performing a motor task, they also display progress toward using fNIRS and EEG signals to determine the amount of force exerted during that motor task. Ortega and Faisal[66] then attempted to use this architecture to determine force onset and which hand was providing more force. This resulted in a force onset detection of 85% but only a hand disentanglement accuracy of 53%, showing that there is still progress to be made toward the complex decoding and reconstruction of motor activities.

In the motor imagery tasks, the first input image method achieved an accuracy of 77.58% using HbO_2 data, while the second method achieved an accuracy of 80.49% using HbO_2 data. It could be noted that HbR and HbT were also tested but for both methods, HbO_2 showed consistently higher results. While HbO_2 is commonly used in fNIRS studies due to higher SNR, Yücel et al.[95] and Herold et al.[96] reported that trends found in HbO_2 signals but not in HbR signals may be due to higher sensitivity of HbO_2 to systemic signals not originating in the brain. For this reason, it is generally recommended that both HbR and HbO_2 signals are assessed in

fNIRS studies. Other traditional classifiers were also used and the closest results were achieved by the metacognitive radial basis function network, which achieved a classification accuracy of 80.83%. Another study that focused on motor imagery, Ma et al.[35] used a type of time series data that included a one-hot label as the input for the neural networks. Of the DL techniques used, the fully connected network (FCN) and residual network (ResNet) achieved the highest average accuracy of 98.6%. Of the traditional machine learning techniques tested, the SVM had the highest classification accuracy of 94.7%. Interestingly enough, for the DL networks, the mean classification accuracy achieved with $HbO_2 + HbR + HbT$ data, 98.3%, was the same as the accuracy when only HbR+HbT data were input, which was attributed to the feature extraction capabilities of the DL techniques. This study also evaluated the accuracy of individual channels, with single-channel classification accuracy ranging from 61.0% to 80.1% with the three highest accuracy channels being found in the somatosensory motor cortex and primary motor cortex.

While many studies have found considerable success in distinguishing between certain tasks using cortical activations, most studies use simple tasks in controlled environments and focus on distinguishing tasks from each other and resting state. For a practical BCI, more complex tasks and classification methods will need to be considered. Zhao et al.[67] addressed this by having participants perform a task in which they would pick up a table tennis ball with chopsticks and lift it about 20 cm in the air, using their nondominant hand. An LSTM was used to try to determine when the task was completed, and ΔHbO_2 data were used as the input, resulting in an accuracy of 71.70%, outperforming the 66.6% accuracy of the SVM that was given mean, variance, kurtosis, and skew features of the fNIRS data as inputs.

One commonly used technology for BCI studies is EEG, due to the high temporal resolution and portability and noninvasiveness. Because it has a high temporal resolution but low spatial resolution, it is common to combine EEG measurements with fNIRS, due to both being portable and noninvasive. In a motor imagery study, Ghonchi et al.[68] used fNIRS to augment the EEG data being collected. Three types of DL networks were used as classifiers, a CNN for its capability to extract special features, an LSTM for its ability to extract temporal features, and a recurrent CNN (RCNN) for its ability to extract both temporal and spatial features. The RCNN achieved the highest classification accuracy of 99.6% when both EEG and fNIRS data were used. Interestingly, the accuracy of both the CNN and LSTM increased when fNIRS data were added, jumping from 85% and 81% to 98.2% and 95.8%, respectively. This indicates that despite EEG data having higher temporal resolution, fNIRS data still contribute both spatial and temporal information. A 2016 study by Chiarelli et al.[69] also found an increase in classification accuracy when combining EEG and fNIRS data. When performing two-class classification on a motor imagery task with an MLP, the average accuracies of EEG and fNIRS data alone were 73.38% and 71.92%, respectively, but when using both modalities, accuracy increased to 83.28%, further reinforcing that the simultaneous acquisition of EEG and fNIRS can provide more relevant information than either modality on their own. Cooney et al.[70] found that when combining fNIRS and EEG data, they were able to distinguish between multiple combinations of overt speech with a CNN classifier, achieving an accuracy of 46.31%. When tested on imagined speech, the classifier achieved an accuracy of 34.29%, which is also higher than the random chance value of 6.25% for 16 possible combinations, which shows promise in the use of EEG and fNIRS for assisting patients who may be unable to verbally communicate. Sun et al.[71] used a CNN to try to distinguish between rest mental arithmetic and motor imagery tasks using both EEG and fNIRS data by generating tensors of the fused EEG and fNIRS data. For the motor imagery tasks, this resulted in an accuracy of 77.53% and for the mental arithmetic tasks, this resulted in an accuracy of 91.83%. Using the same dataset, Kwak et al.[72] applied a branched CNN architecture that used the fNIRS data to generate spatial feature maps, which were then fed to the EEG maps to try to obtain higher spatial resolution than ordinary EEG and higher temporal resolution than fNIRS. The resulting classification accuracies were 78.97% for motor imagery and 91.96% for mental arithmetic tasks, which are only slight improvements over the tensor fusion methods of Sun et al. Khalil et al.[73] also used a fusion of fNIRS and EEG data to distinguish between rest and a mental workload tasks. With a CNN, an accuracy of 68.94% was achieved when trained on data from 5 of the 26 participants. When training on 16 participants then performing transfer learning to an additional five participants, accuracy increased to 94.52%, exemplifying not only how important larger datasets are for DL but also how transfer learning can be used to address this. It may be of

interest to explore the use of transfer learning from other fNIRS datasets, which use different hardware systems, different tasks, or different fNIRS channel arrangements, since many studies rely on collecting fNIRS data specific to their own applications. As a result, it would be important to see if transfer learning can help extract meaningful features from fNIRS data independently of the part of the brain being recorded or the hardware being used. While many studies have only recently begun looking to use fNIRS for real-time BCI applications, current studies in the field have found use in DL techniques for increased classification accuracy and automatic feature extraction.

3.4 *Diagnostic Tools*

One promising use of DL techniques with fNIRS is in clinical applications, most notably as a diagnostic tool. Xu et al.[46] recorded resting state fNIRS data from the bilateral frontal gyrus and bilateral temporal lobe of children. Using a single channel in the left temporal lobe, a CGRNN classifier was able to successfully classify autism spectrum disorder (ASD) in children with 92.2% accuracy, 85.0% sensitivity, and 99.4% specificity for 7 s of resting-state HbR data. As shown in Fig. 6, multiple HbO_2 and HbR channels showed statistically significant differences between the group with ASD and the control group. Xu et al.[74] used a CNN classifier with the attention layers and achieved an accuracy sensitivity and specificity of 93.3%, 90.6%, and 97.5%, respectively, using similar resting state data to classify between ASD and typically developing (TD) subjects. Xu et al.[75] further explored using fNIRS data to detect autism is subjects and found that using a CNN+LSTM classifier and HbO_2 data, they were able to achieve a single-channel accuracy, sensitivity and specificity of 95.7%, 97.1%, and 94.3%, respectively, for the detection of ASD. Aside from autism, the use of fNIRS to diagnose other psychological disorders has been studied. With an average classification accuracy of 96.2%, Ma et al.[76] were able to distinguish bipolar depression from major depressive disorder in adults during a verbal fluency task using an LSTM. Wang et al.[77] managed to distinguish between healthy subjects and those diagnosed with major depressive disorder with an accuracy of 83.3%. This was accomplished using long recording times of 150 min each from a relatively large sample size of 96 subjects. As can be seen in Table 2, this is a larger sample size than any of the other fNIRS studies presented, which lends to confidence in the generalizability of this neural network to new subjects. Chao et al.[78] used a cascade forward neural network (a network similar to an MLP) to perform and achieved an average classification accuracy of 99.94% between depressed and healthy subjects when a fear stimulus was presented across 32 subjects. Chou et al.[79] used an MLP network to classify between subjects with first episode schizophrenia and healthy subjects,

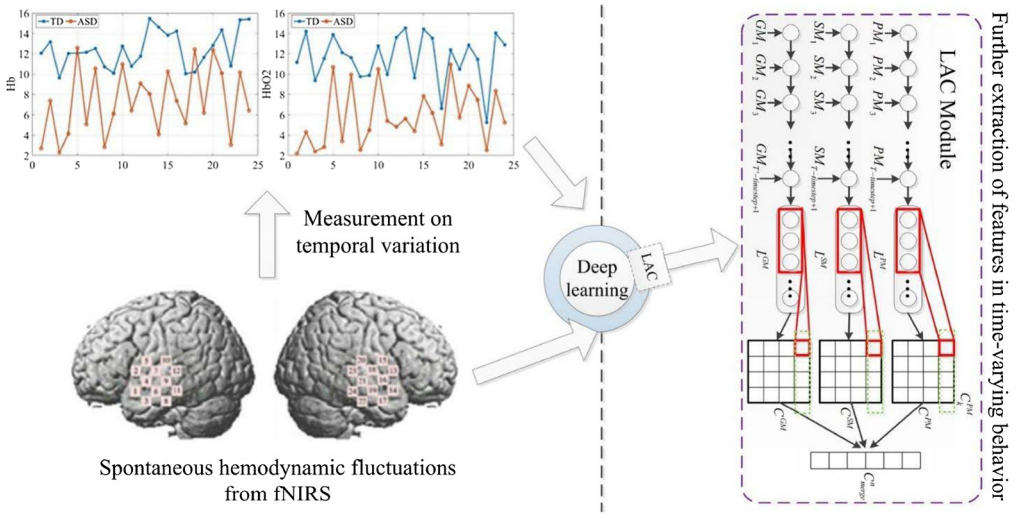

Fig. 6 A DL model combining LSTM and CNN (LAC) was used to accurately identify ASD from a TD subject based on time-varying behavior of spontaneous hemodynamic fluctuations from fNIRS. Adapted from Ref. 75.

achieving a classification accuracy of 79.7% with only about 160 s of recorded data from each subject.

EEG technology is commonly used in clinical settings, however, some studies have used EEG alongside fNIRS for diagnostic studies. Sirpal et al.[44] used an LSTM architecture and fNIRS data to detect seizures with 97.0% accuracy, and with EEG data, achieved 97.6% accuracy, but with combined EEG and fNIRS data, accuracy increased to 98.3%, once again displaying how hybrid EEG-fNIRS recordings can increase classification accuracy, even when accuracy is already high. Rosas-Romero et al.[80] also combined fNIRS and EEG signals to detect epilepsy. Despite only having recordings from five subjects, a CNN was able to detect pre-ictal segments fNIRS and EEG data with an average accuracy of 99.67% with fivefold CV.

Another use for fNIRS is to help detect mild cognitive impairment (MCI), the prodromal stage of Alzheimer's disease. Yang et al.[81] employed three different strategies using CNNs to detect MCI. Using an N-back, Stroop, and verbal fluency task (VFT), a CNN trained on concentrations changes of HbO_2 achieved accuracies ranging from 64.21% in the N-back task to 78.94% in the VFT. When using activation maps as the inputs for the CNN, the accuracy ranged from 71.59% in the VFT to 90.62% in the N-back task. The final strategy employed, using correlation maps as inputs showed lower accuracies than the activation maps, with the highest accuracy being 85.58% for the N-back task. Yang et al.[47] used temporal feature maps as inputs for the CNN classifier, resulting in average accuracies of 89.46%, 87.80%, and 90.37% with the N-back, Stroop, and VFT, respectively. In another study done in 2021 by Yang and Hong,[82] pretrained networks were used to distinguish between subjects with MCI and the healthy control group. Using resting state fNIRS data, the network with the highest accuracy, VGG19, achieved an accuracy of 97.01% when connectivity maps were used as the input, outperforming the conventional machine learning techniques, with LDA classifier reporting the highest accuracy of 67.00%. Ho et al.[83] attempted to use fNIRS and DL techniques to distinguish not only between healthy and prodromal Alzheimer's afflicted subjects but also subjects with asymptomatic Alzheimer's disease and dementia due to Alzheimer's disease. Not only did this study try to distinguish between different stages of Alzheimer's disease, the study used a notably large sample size of 140 participants, which was larger than any other study reported in this review as shown in Table 2. The 86.8% accuracy of the CNN-LSTM network when fivefold cross-validated not only shows the ability of the network to distinguish between a wide range of subjects with and without Alzheimer's disease but also shows the ability of this network to distinguish between different stages of Alzheimer's disease, making this a very promising tool for clinical use.

Yet another clinical application for fNIRS measurements with DL techniques was explored by Fernandez Rojas et al.,[41] where raw fNIRS data and an LSTM were used to distinguish between high and low levels of pain as well as whether the pain was caused by a hot or cold stimulus with an achieved accuracy of 90.6%. Being able to assess the intensity of pain as well as the cause of it could be exceptionally useful in instances where patients are unable to communicate, such as with nonverbal patients. This further displays the usefulness of fNIRS with DL techniques as a robust and accurate diagnostic tool.

3.5 *Analysis of Cortical Activations*

Outside of BCI and diagnostic tools, fNIRS data still have many uses. Understanding the functional connectivity of the brain is an essential part of understanding the mechanisms behind numerous neurological phenomena. As a result, there is an interest in using neuroimaging to understand the functional connectivity of the brain. Behboodi et al.[84] used fNIRS to record the resting-state functional connectivity (RSFC) of the sensorimotor and motor regions of the brain. In this study, four methods were used, seed-based, ICA, ANN, and CNN. Unlike the first two methods, very limited preprocessing was used for the ANN and CNN, with both using filtered HbO_2 data to form the connectivity maps. Each connectivity map was then compared with the expected activation based on the physiological location of each detector, which was used as the ground truth, using an ROC curve, as shown in Fig. 7, with the CNN achieving an area under the curve (AUC) of the receiver operating characteristic curve (ROC curve) of 0.92, which outperformed the ANN, ICA, and seed-based methods with each reporting an AUC of 0.89, 0.88, and 0.79, respectively. Another study that investigated RSFC, Sirpal et al.,[85] collected EEG and

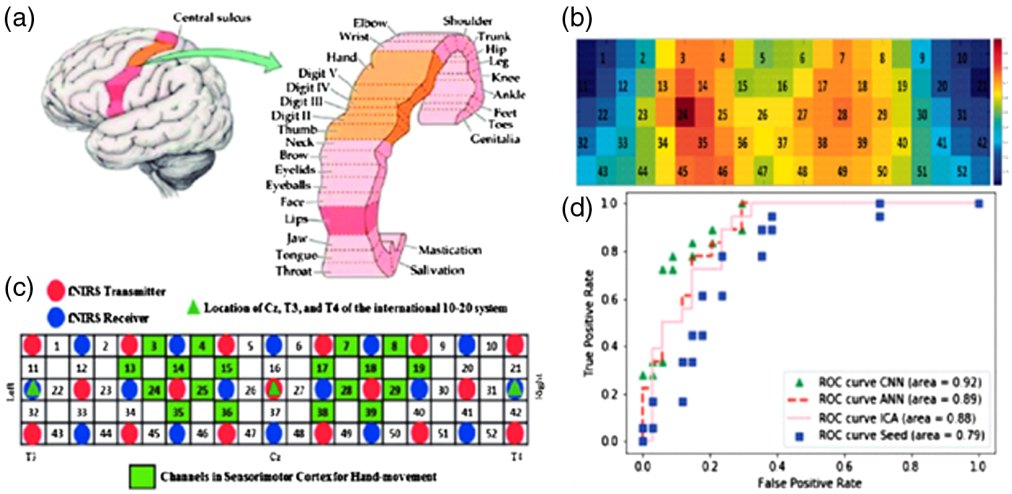

Fig. 7 (a) The cortical areas of hand-movement in sensorimotor and motor cortices. (b) Anatomical areas of sensorimotor and motor areas that are related to the hand-movement monitored by fNIRS are highlighted in green. (c) Group RSFC map derived from CNN-based resting-state connectivity detection. (d) ROC curves on different methods. Adapted from Ref. 84.

fNIRS data and attempted to use the EEG data and an LSTM to recreate the fNIRS signals. Using only the gamma bands of the EEG signal was found to have the lowest reconstruction error, below 0.25. The reconstructed fNIRS signals were further validated by comparing the functional connectivity of the signals constructed using only gamma bands and those using the full spectrum EEG signals with the functional connectivity of the experimental fNIRS data. Despite the signals formed using gamma bands only having a lower reconstruction error, the root MSE of the full spectrum EEG signals was consistently lower.

Other types of fNIRS studies have also been done that utilized DL. Some studies have analyzed the cortical activations of subjects to predict emotions. Bandara et al.[86] used music videos from the DEAP database[99] to classify the emotional valence and arousal of subjects using a CNN +LSTM architecture and fNIRS data recorded from the prefrontal cortex. Using the subjects' self-assessments as ground-truth, a classification accuracy of 77% was reported. Another study collecting fNIRS signals from the prefrontal cortex, Qing et al.[87] used a CNN to determine the preference levels of subjects toward various Pepsi and Coca-Cola ads, achieving an average three-class classification accuracy of 87.9% for 30 s videos. 15 and 60 s videos showed similar accuracies of 84.3% and 86.4%, respectively. Similarly, Ramirez et al.[90] attempted to decode consumer preference toward 14 different products. With a CNN, Ramirez et al. were able to distinguish between a strong like and strong dislike of the presented product with an accuracy of 68.6% with fNIRS data, 77.98% with just EEG data, and 91.83% with combined fNIRS and EEG data. Hiwa et al.[88] studied the use of fNIRS and CNNs for the identification of a subject's gender, achieving an accuracy of ∼60% when using the filtered fNIRS data from only five channels.

While most studies have focused on using cortical activations to classify when a specified task is being completed, a few studies have begun looking into predicting the skill level of the subject at a given task. Andreu-Perez et al.[89] classified the expertise of subjects watching 30 s clips of the video game League of Legends using fNIRS data and facial expressions. The fully connected deep neural network (FCDNN) and deep classifier autoencoder (DCAE) were compared against many traditional machine learning techniques, including SVM and kNN in a three-class test to determine the skill level of the subject watching. Using only fNIRS data, the DL classifiers had the two highest accuracies of 89.84% and 90.70% for the FCDNN and DCAE, respectively, while the most accurate machine learning technique, SVM, only achieved an accuracy of around 58.23%. When the predicted emotion scores were also included, the accuracy of the FCDNN and DCAE improved to 91.44% and 91.43%, respectively. Most of the machine learning techniques also saw minor or no improvements, with the SVM still achieving an accuracy of 58.69%. The XGBoost classifier saw a large increase in accuracy when emotion

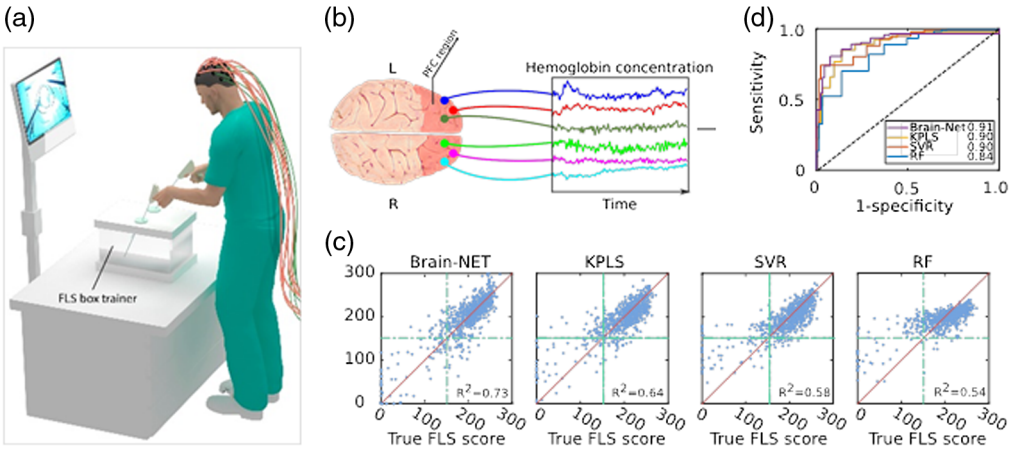

Fig. 8 (a) Schematic depicting the FLS box simulator where trainees perform the bimanual dexterity task. A continuous-wave spectrometer is used to measure functional brain activation via raw fNIRS signals in real time. (b) Examples of acquired time-series hemoglobin concentration data from six PFC locations while the subject is performing the PC surgical task. (c) The true versus predicted FLS score plots. Each blue dot represents one sample. The red line is the $y = x$ line. The dot-dashed green lines represent the certification pass/fail threshold FLS score value. (d) ROC curves for each model with corresponding AUC values in the legend. Adapted from Refs. 100 and 32.

scores were added, increasing from 50.79% to 71.55%, which was still about 20% lower than the accuracies of the DL techniques used. Another study by Gao et al.[32] looked to predict the surgical skill level of medical school students who were being assessed on tasks based on the FLS protocols. In this study, subjects would perform an FLS precision cutting task while fNIRS data were recorded from the prefrontal cortex. The subjects would be assessed and scored in accordance with FLS procedures. A brain-NET architecture was used to predict the FLS scores of each participant based on the features extracted from the recorded fNIRS data. The brain-NET model reported an ROC AUC of 0.91, outperforming the kernel partial least squares, random forest, and support vector regression methods that were also tested when the dataset was sufficiently large. Figure 8 shows the R^2 value of each methodology as the size of the dataset increases. The ROC curve can also be seen along with a graphic of the experimental setup in Fig. 8. From the wide range of studies published, there are many different applications of fNIRS currently being researched, and the use of DL techniques has become of interest in recent years.

4 Discussion and Future Outlook

Over the last two decades, machine learning has become increasingly popular for processing neuroimaging data due to its benefit over traditional analysis methods.[101] Of importance, ML methods enable fully processing spatiotemporal data sets and allow for inferencing at the single subject/trial level. Among all ML approaches, DL is becoming increasingly utilized over the last half decade with great promise.[102,103] Following similar trends, DL models have found increased utility in fNIRS applications, ranging from simplifying the data processing pipeline to performing classification or prediction tasks. DL models are expected to outperform ML methods due to their potential to directly extract features from raw data (no need to perform prior feature extraction) and learn complex features in a hierarchical manner. This seems to be further supported by the findings of this review in which, out of the 32 papers reporting on a comparative study of DL techniques to traditional machine learning techniques, 26 have been shown outperforming the latter in terms of classification accuracy. Such trends have also been reported over a large range of biomedical applications, but we cannot exclude a publication bias due to the specific nature of the review topic. Still, the application of DL to fNIRS is in its very early stages and faces many challenges.

First, the implementation of DL models is an expert field. The selection of the main architecture as well as the number of layers, the activation function, are still dependent on the user

expertise but greatly influence the performance and applicability of any DL model. If this design flexibility enables powerful implementations, it leads to a wide range of architecture and hyperparameters employed in the set of work reviewed herein. Hence, there is still not a consensus on which architecture and hyperparameters are optimal for a specific problem and building on current work requires some level of technical expertise to assess which implementation would be optimal. This may be circumvented in the future with the advent of network automated designed via neural architecture search methods,[104] but these have not been yet applied to the field of fNIRS. Moreover, following the principles of the no-free-lunch theorem, it is expected that prior knowledge on the problem at hand should guide in the selection of the ML/DL algorithm. Hence, beyond technical expertise in DL, ones need also to have a good grasp of the neurophysiology as well as instrumentation characteristics used in the application to design optimal models.

This leads to another significant challenge, which is associated with the data-driven nature of DL model training and validation. The lack of sufficient training data is a common challenge in the application of DL methods in neuroimaging. This is even more challenging for fNIRS applications that are less ubiquitous than MRI or EEG, which benefit from publicly available repository. As we are still far from being able to model the complexity of brain functions and dynamics, the DL models can be trained only on experimental data conversely to many other fields in which efficient *in silico* data generators are available.[105] This is highlighted in Table 3, which shows that almost all reviewed work depended on proprietary data and with relatively small number of subjects. Moreover, in numerous scenarios, the data quality can be poor such that a subset of the spatiotemporal data is missing or inadequate (for instance, compromised by motion artifacts, shallow physiological variations). As previously mentioned, data augmentation approaches have been implemented successfully to alleviate this challenge. Another approach is to leverage new developments in transfer learning that optimally refine well-trained networks on large data sets to smaller one. But still, these methods are expected to work well within homogeneous settings. As the field benefit from an increased number of fNIRS systems with varying optode characteristics and associated electronics, the raw characteristics of the acquired signals can greatly vary (SNR, CNR, sampling rate) and hence, limit generalizability. Another issue for generalizability is that overfitting may be especially prevalent in fNIRS studies where fNIRS data tend to be highly correlated.[95] Moreover, in many instances, the data set is imbalanced for available classes. Hence, it is crucial for eliciting confidence in the results to report on CV results. Herein, most reviewed work used k-fold CV, typically 5- or 10-fold CV. Still, in many applications, the data set is comprised of multitrials per subject. Hence, it is important to assess the potential bias associated with each subject. This can be performed using LOUO CV and LOSO CV, respectively. Still, such well-established methods were used only in 8 of the 63 papers considered in this review.

Last, despite demonstrating high performances, the DL implementations reported herein are suffering from the black-box issue. In other words, the extracted high-level features from the data inputs that lead to high task performances during training and validation are not accessible and hence, cannot be interpreted. However, in the last few years, various eXplainable AI (XAI) tools such as class activation maps,[106] Grad-CAMS,[107] and saliency maps[108] have been proposed to impart understandability and comprehensibility.[109] Such tools can, for instance, provide visual map(s) that highlight the main data features leveraged for the model decision. Such a map can then correlate the extracted features with known neurophysiology, specific application characteristics (for instance, hand switching during surgery), and/or correlate with other biomarkers, such as videos, gaze measurements, and motion tracking devices. For this reason, XAI tools are well-poised to lead to the discovery of new spatiotemporal features that will advance our neuroscience knowledge at large. This is exemplified by the recent report of differences in activation maps between expert and novice surgeons while executing a certification task, with activation maps obtained via a dot-attention method in a DL classifier model.[110]

5 Conclusion

We reviewed the most recent published work relevant to the application of DL techniques to fNIRS. This literature review indicated that DL models were mainly sued for classification tasks

based on fNIRS data and that in most of the cases, DL model prediction accuracy outperformed traditional techniques, including established ML methods. Another subset of work reported on developing DL models to reduce the amount of preprocessing typically done with fNIRS data or increase the amount of data via data augmentation. In all cases, DL models provided very fast inference computational times. These characteristics have a transformative power for the field of fNIRS at large as they pave the way to fast and accurate data processing and/or classification tasks. Of note, DL models, when validated and established, offer the unique potential for real-time processing on the bedside at minimal computational cost. While the deployment of DL models that are widely accepted by the community face numerous challenges, the findings reviewed here provide evidence that DL will play an increased role in fNIRS data processing and use for a wide range of bedside applications. Moreover, as an ever-increased number of studies are made available to the community, it is expected that the next generation of DL models will have the possibility to be tested and validated in various scenarios.

Disclosures

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

Acknowledgments

We are grateful for the funding provided by NIH/National Institute of Biomedical Imaging and Bioengineering Grant Nos. 2R01EB005807, 5R01EB010037, 1R01EB009362, 1R01EB014305, and R01EB019443, the Medical Technology Enterprise Consortium Grant No. 20-05-IMPROVE-004, and the US Army Combat Capabilities Development Command Grant No. W912CG2120001.

References

1. F. F. Jobsis, "Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters," *Science* **198**, 1264–1267 (1977).
2. H. Obrig, "NIRS in clinical neurology: a 'promising' tool?" *Neuroimage* **85**(Pt 1), 535–546 (2014).
3. K. S. Hong and M. A. Yaqub, "Application of functional near-infrared spectroscopy in the healthcare industry: a review," *J. Innov. Opt. Health Sci.* **12**, 1930012 (2019).
4. V. Quaresima and M. Ferrari, "A mini-review on functional near-infrared spectroscopy (fNIRS): where do we stand, and where should we go?" *Photonics* **6**, 87 (2019).
5. E. T. Solovey et al., "Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines," in *Proc. 22nd Annu. ACM Symp. User Interface Softw. and Technol.*, pp. 157–166 (2009).
6. M. D. Pfeifer, F. Scholkmann, and R. Labruyère, "Signal processing in functional near-infrared spectroscopy (fNIRS): methodological differences lead to different statistical results," *Front. Hum. Neurosci.* **11**, 641 (2018).
7. P. Pinti et al., "The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience," *Ann. N. Y. Acad. Sci.* **1464**(1), 5–29 (2020).
8. L. Tian et al., "Deep learning in biomedical optics," *Lasers Surg. Med.* **53**(6), 748–775 (2021).
9. K. Suzuki, "Overview of deep learning in medical imaging," *Radiol. Phys. Technol.* **10**(3), 257–273 (2017).
10. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*, California University San Diego La Jolla Inst For Cognitive Science (1985). https://apps.dtic.mil/sti/citations/ADA164453.
11. H. Abdi, "A neural network primer," *J. Biol. Syst.* **2**(3), 247–281 (2011).
12. S. Ichi Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing* **5**(4–5), 185–196 (1993).

13. Y. Lecun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**(11), 2278–2324 (1998).
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Adv. Neural Inf. Process. Syst., Vol. 25 (2012). https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.
15. A. Graves et al., "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(5), 855–868 (2009).
16. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Adv. in Neural Inf. Process. Syst.*, Vol. **27** (2014). https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.
17. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
18. R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. 30th Int. Conf. Mach. Learn.*, pp. 1310–1318 (2013).
19. M. Mirbagheri et al., "Accurate stress assessment based on functional near infrared spectroscopy using deep learning approach," in *26th Natl. and 4th Int. Iran. Conf. Biomed. Eng. (ICBME)*, pp. 4–10 (2019).
20. M. A. Tanveer et al., "Enhanced drowsiness detection using deep learning: an fNIRS study," *IEEE Access* **7**, 137920–137929 (2019).
21. T. C. Dolmans et al., "Perceived mental workload classification using intermediate fusion multimodal deep learning," *Front. Hum. Neurosci.* **14**, 609096 (2021).
22. X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Fourteenth Int. Conf. Artif. Intell. and Stat.*, pp. 315–323 (2011).
23. M. Saadati, J. Nelson, and H. Ayaz, "Multimodal fNIRS-EEG classification using deep learning algorithms for brain-computer interfaces purposes," in *Adv. Intell. Syst. and Comput.*, Vol. 953 (2020).
24. P. Ortega and A. Faisal, "HemCNN: deep learning enables decoding of fNIRS cortical signals in hand grip motor tasks," in *Int. IEEE/EMBS Conf. Neural Eng., NER*, Vol. 2021 (2021).
25. A. R. Dargazany, M. Abtahi, and K. Mankodiya, "An end-to-end (deep) neural network applied to raw EEG, fNIRs and body motion data for data fusion and BCI classification task without any pre-/post-processing," (2019). http://arxiv.org/abs/1907.09523.
26. T. Nagasawa et al., "FNIRS-GANs: data augmentation using generative adversarial networks for classifying motor tasks from functional near-infrared spectroscopy," *J. Neural Eng.* **17**(1) (2020).
27. K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *Schedae Inf.* **25**, 49–59 (2017).
28. H. Ghonchi et al., "Deep recurrent-convolutional neural network for classification of simultaneous EEG-fNIRS signals," *IET Signal Process.* **14**(3), 142–153 (2020).
29. T. Trakoolwilaiwan et al., "Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: three-class classification of rest, right-, and left-hand motor execution," *Neurophotonics* **5**(1), 011008 (2017).
30. A. Janani et al., "Investigation of deep convolutional neural network for classification of motor imagery fNIRS signals for BCI applications," *Biomed. Signal Process. Control* **62**, 102133 (2020).
31. S. H. Yoo et al., "Decoding multiple sound-categories in the auditory cortex by neural networks: an fNIRS study," *Front. Hum. Neurosci.* **15**, 636191 (2021).
32. Y. Gao et al., "Functional brain imaging reliably predicts bimanual motor skill performance in a standardized surgical task," *IEEE Trans. Biomed. Eng.* **68**(7), 2058–2066 (2021).
33. P. Ortega, T. Zhao, and A. Faisal, "CNNATT: deep EEG & fNIRS real-time decoding of bimanual forces," (2021). http://arxiv.org/abs/2103.05334.
34. Y. Gao et al., "Deep learning-based motion artifact removal in functional near-infrared spectroscopy," *Neurophotonics* **9**(4), 041406 (2022).
35. T. Ma et al., "CNN-based classification of fNIRS signals in motor imagery BCI system," *J. Neural Eng.* **18**(5), 056019 (2021).

36. S. K. Kumar, "On weight initialization in deep neural networks," (2017). https://arxiv.org/abs/1704.08863v2 (accessed 14 January 2022).
37. A. Labach, H. Salehinejad, and S. Valaee, "Survey of dropout methods for deep neural networks," (2019). https://arxiv.org/abs/1904.13310v2 (accessed 14 January 2022).
38. G. Klambauer et al., "Self-normalizing neural networks," in *Adv. Neural Inf. Process. Syst.*, Vol. **30** (2017).
39. K. He et al., "Deep residual learning for image recognition," arXiv:1512.03385 [Cs] (2015).
40. K. Cao and X. Zhang, "An improved res-UNet model for tree species classification using airborne high-resolution images," *Remote Sens.* **12**(7), 1128 (2020).
41. R. Fernandez Rojas et al., "Pain assessment based on fNIRS using bidirectional LSTMs," (2020). http://arxiv.org/abs/2012.13231
42. S. D. Wickramaratne and M. S. Mahmud, "A deep learning based ternary task classification system using gramian angular summation field in fNIRS neuroimaging data," in *IEEE Int. Conf. E-health Netw., Appl. & Services (HEALTHCOM)*, pp. 1–4 (2021).
43. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
44. P. Sirpal et al., "fNIRS improves seizure detection in multimodal EEG-fNIRS recordings," *J. Biomed. Opt.* **24**(5), 051408 (2019).
45. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *32nd Int. Conf. Mach. Learn., ICML*, Vol. 1 (2015).
46. L. Xu et al., "Prediction in autism by deep learning short-time spontaneous hemodynamic fluctuations," *Front. Neurosci.* **13**, 1120 (2019).
47. D. Yang et al., "Detection of mild cognitive impairment using convolutional neural network: temporal-feature maps of functional near-infrared spectroscopy," *Front. Aging Neurosci.* **12**, 141 (2020).
48. S. Takagi et al., "Application of deep learning in the identification of cerebral hemodynamics data obtained from functional near-infrared spectroscopy: a preliminary study of pre- and post-tooth clenching assessment," *J. Clin. Med.* **9**(11), 3475 (2020).
49. J. Lu et al., "Comparison of machine learning and deep learning approaches for decoding brain computer interface: an fNIRS study," in *IFIP Adv. Inf. and Commun. Technol., AICT*, Vol. **581** (2020).
50. M. Saadati, J. Nelson, and H. Ayaz, "Mental workload classification from spatial representation of FNIRS recordings using convolutional neural networks," in *IEEE Int. Workshop on Mach. Learn. for Signal Process., MLSP*, October, Vol. 2019 (2019).
51. J. Benerradi et al., "Exploring machine learning approaches for classifying mental workload using fNIRS data from HCI tasks," in *Proc. Halfway to the Future Symp.*, pp. 1–11 (2019).
52. R. Liu et al., "Unsupervised fNIRS feature extraction with CAE and ESN autoencoder for driver cognitive load classification," *J. Neural Eng.* **18**(3), 036002 (2021).
53. G. Lee, S. H. Jin, and J. An, "motion artifact correction of multi-measured functional near-infrared spectroscopy signals based on signal reconstruction using an artificial neural network," *Sensors (Basel, Switzerland)* **18**(9), 2957 (2018).
54. M. Kim et al., "A deep convolutional neural network for estimating hemodynamic response function with reduction of motion artifacts in fNIRS," *J. Neural Eng.* **19**(1), 016017 (2022).
55. S. D. Wickramaratne and M. S. Mahmud, "Conditional-GAN based data augmentation for deep learning task classifier improvement using fNIRS data," *Front. Big Data* **4**, 659146 (2021).
56. S.-W. Woo, M.-K. Kang, and K.-S. Hong, "Classification of finger tapping tasks using convolutional neural network based on augmented data with deep convolutional generative adversarial network," in *8th IEEE RAS/EMBS Int. Conf. for Biomed. Rob. and Biomech. (BioRob)*, pp. 328–333 (2020).
57. J. Hennrich et al., "Investigating deep learning for fNIRS based BCI," in *Proc. Annu. Int. Conf. IEEE Eng. Med. and Biol. Soc., EMBS*, November, Vol. 2015 (2015).

58. J. Kwon and C. H. Im, "Subject-independent functional near-infrared spectroscopy-based brain–computer interfaces based on convolutional neural networks," *Front. Hum. Neurosci.* **15**, 646915 (2021).
59. T. K. K. Ho et al., "Discrimination of mental workload levels from multi-channel fNIRS using deep leaning-based approaches," *IEEE Access* **7**, 24392–24403 (2019).
60. U. Asgher et al., "Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain-computer interface," *Front. Neurosci.* **14**, 584 (2020).
61. N. Naseer et al., "Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface," *Comput. Intell. Neurosci.* **2016**, 5480760 (2016).
62. N. Hakimi et al., "Proposing a convolutional neural network for stress assessment by means of derived heart rate from functional near infrared spectroscopy," *Comput. Biol. Med.* **121**, 103810 (2020).
63. S. B. Erdoğan et al., "Classification of motor imagery and execution signals with population-level feature sets: implications for probe design in fNIRS based BCI," *J. Neural Eng.* **16**(2), 026029 (2019).
64. H. Hamid et al., "Analyzing classification performance of fNIRS-BCI for gait rehabilitation using deep neural networks," *Sensors (Basel, Switzerland)* **22**(5), 1932.
65. H. Khan et al., "Classification of individual finger movements from right hand using fNIRS signals," *Sensors (Basel, Switzerland)* **21**(23), 7943 (2021).
66. P. Ortega and A. A. Faisal, "Deep learning multimodal fNIRS and EEG signals for bimanual grip force decoding," *J. Neural Eng.* **18**(4), 0460e6 (2021).
67. Q. Zhao et al., "FNIRS based brain-computer interface to determine whether motion task to achieve the ultimate goal," in *IEEE 4th Int. Conf. Adv. Rob. and Mechatronics (ICARM)*, pp. 136–140 (2019).
68. H. Ghonchi et al., "Spatio-temporal deep learning for EEG-fNIRS brain computer interface," in *Proc. Annu. Int. Conf. IEEE Eng. in Med. and Biol. Soc., EMBS*, July, Vol. 2020 (2020).
69. A. M. Chiarelli et al., "Deep learning for hybrid EEG-fNIRS brain-computer interface: application to motor imagery classification," *J. Neural Eng.* **15**(3), 036028 (2018).
70. C. Cooney, R. Folli, and D. H. Coyle, "A bimodal deep learning architecture for EEG-fNIRS decoding of overt and imagined speech," *IEEE Trans. Biomed. Eng.* **69**(6), 1983–1994 (2022).
71. Z. Sun et al., "A novel multimodal approach for hybrid brain–computer interface," *IEEE Access* **8**, 89909–89918 (2020).
72. Y. Kwak, W.-J. Song, and S.-E. Kim, "FGANet: fNIRS-guided attention network for hybrid EEG-fNIRS brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 329–339 (2022).
73. K. Khalil, U. Asgher, and Y. Ayaz, "Novel fNIRS study on homogeneous symmetric feature-based transfer learning for brain-computer interface," *Sci. Rep.* **12**(1), 3198 (2022).
74. L. Xu et al., "Identification of autism spectrum disorder based on short-term spontaneous hemodynamic fluctuations using deep learning in a multi-layer neural network," *Clin. Neurophysiol.* **132**(2), 457–468 (2021).
75. L. Xu et al., "Characterizing autism spectrum disorder by deep learning spontaneous brain activity from functional near-infrared spectroscopy," *J. Neurosci. Methods* **331**, 108538 (2020).
76. T. Ma et al., "Distinguishing bipolar depression from major depressive disorder using fnirs and deep neural network," *Progr. Electromagn. Res.* **169**, 73–86 (2020).
77. R. Wang et al., "Depression analysis and recognition based on functional near-infrared spectroscopy," *IEEE J. Biomed. Health Inf.* **25**(12), 4289–4299 (2021).
78. J. Chao et al., "fNIRS evidence for distinguishing patients with major depression and healthy controls," *IEEE Trans. Neural Syst. Rehabil. Eng.* **29**, 2211–2221 (2021).
79. P. H. Chou et al., "Deep neural network to differentiate brain activity between patients with first-episode schizophrenia and healthy individuals: a multi-channel near infrared spectroscopy study," *Front. Psychiatr.* **12**, 655292 (2021).

80. R. Rosas-Romero et al., "Prediction of epileptic seizures with convolutional neural networks and functional near-infrared spectroscopy signals," *Comput. Biol. Med.* **111**, 103355 (2019).
81. D. Yang et al., "Evaluation of neural degeneration biomarkers in the prefrontal cortex for early identification of patients with mild cognitive impairment: an fNIRS study," *Front. Hum. Neurosci.* **13**, 317 (2019).
82. D. Yang and K. S. Hong, "Quantitative assessment of resting-state for mild cognitive impairment detection: a functional near-infrared spectroscopy and deep learning approach," *J. Alzheimer's Dis.* **80**(2), 647–663 (2021).
83. T. Ho et al., "Deep learning-based multilevel classification of Alzheimer's disease using non-invasive functional near-infrared spectroscopy," *Front. Aging Neurosci.* **14**, 810125 (2022).
84. B. Behboodi et al., "Artificial and convolutional neural networks for assessing functional connectivity in resting-state functional near infrared spectroscopy," *J. Near Infrared Spectrosc.* **27**(3) (2019).
85. P. Sirpal et al., "Multimodal autoencoder predicts fNIRS resting state from EEG signals," *Neuroinformatics* **27**(3), 191–205 (2021).
86. D. Bandara, L. Hirshfield, and S. Velipasalar, "Classification of affect using deep learning on brain blood flow data," *J. Near Infrared Spectrosc.* **27**(3), 206–219 (2019).
87. K. Qing, R. Huang, and K. S. Hong, "Decoding three different preference levels of consumers using convolutional neural network: a functional near-infrared spectroscopy study," *Front. Hum. Neurosci.* **14**, 597864 (2021).
88. S. Hiwa et al., "Analyzing brain functions by subject classification of functional near-infrared spectroscopy data using convolutional neural networks analysis," *Comput. Intell. Neurosci.* **2016**, 1841945 (2016).
89. A. R. Andreu-Perez et al., "Single-trial recognition of video gamer's expertise from brain haemodynamic and facial emotion responses," *Brain Sci.* **11**(1), 106 (2021).
90. M. Ramirez et al., "Application of convolutional neural network for classification of consumer preference from hybrid EEG and FNIRS signals," in *IEEE 12th Annu. Comput. and Commun. Workshop and Conf. (CCWC)*, pp. 1024–1028 (2022).
91. Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science* **349**(6251), aac4716 (2015).
92. R. Zwaan et al., "Making replication mainstream," *Behav. Brain Sci.* **41**, E120 (2018).
93. A. Koul, C. Becchio, and A. Cavallo, "Cross-validation approaches for replicability in psychology," *Front. Psychol.* **9**, 1117 (2018).
94. E. Yanik et al., "Deep neural networks for the assessment of surgical skills: a systematic review," *J. Defense Model. Simul.* **19**(2), 159–171 (2021).
95. M. Yücel et al., "Best practices for fNIRS publications," *Neurophotonics* **8**(1), 012101 (2021).
96. F. Herold et al., "Applications of functional near-infrared spectroscopy (fNIRS) neuroimaging in exercise–cognition science: a systematic, methodology-focused review," *J. Clin. Med.* **7**(12), 466 (2018).
97. S. Jahani et al., "Motion artifact detection and correction in functional near-infrared spectroscopy: a new hybrid method based on spline interpolation method and Savitzky-Golay filtering," *Neurophotonics* **5**(1), 015003 (2018).
98. F. A. Fishburn et al., "Temporal derivative distribution repair (TDDR): a motion correction method for fNIRS," *Neuroimage* **184**, 171–179 (2019).
99. S. Koelstra et al., "DEAP: a database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.* **3**(1), 18–31 (2012).
100. C. Davatzikos, "Machine learning in neuroimaging: progress and challenges," *NeuroImage* **197**, 652–656 (2019).
101. S. Vieira, W. H. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications," *Neurosci. Biobehav. Rev.* **74**(Pt A), 58–75 (2017).
102. L. Zhang et al., "A survey on deep learning for neuroimaging-based brain disorder analysis," *Front. Neurosci.* **14**, 779 (2020).

103. X. He, K. Zhao, and X. Chu, "AutoML: a survey of the state-of-the-art," *Knowl.-Based Syst.* **212**, 106622 (2021).
104. J. T. Smith et al., "Deep learning in diffuse optical imaging," *J. Biomed. Opt.* **27**(2), 020901 (2022).
105. Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: a strong baseline," in *Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1578–1585 (2017).
106. R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 618–626 (2017).
107. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps" (2013).
108. A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf.Fusion* **58**, 82–115 (2020).
109. F. N. U. Rahul et al., "A deep learning model for a priori estimation of spatiotemporal regions for neuroimaging guided non-invasive brain stimulation," *Brain Stimul.* **14**(6), 1689 (2021).
110. A. Nemani et al., "Assessing bimanual motor skills with optical neuroimaging," *Sci. Adv.* **4**(10), eaat3807 (2018).

Aseem Subedi is a PhD student at Rensselaer Polytechnic Institute working at the CeMSIM lab. His primary research focus is to use deep learning models to simplify and optimize feature recognition from multimodal physiological biomarkers. His current work involves using cognitive responses from cortical areas to understand characteristics of skill acquisition in trainee surgeons, through EEG, fNIRS, and other auxiliary modalities.

Xavier Intes received his PhD from the Université de Bretagne Occidentale and a postdoctoral training at the University of Pennsylvania. He is a professor in the Department of Biomedical Engineering, Rensselaer Polytechnic Institute, and an AIMBE/SPIE/Optica fellow. He acted as the chief scientist of Advanced Research Technologies Inc. His research interests are on the application of diffuse functional and molecular optical techniques for biomedical imaging in preclinical and clinical settings.

Biographies of the other authors are not available.