CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	EEG datasets for healthcare: A scoping review
Authors	Peres da Silva, Caroline;Tedesco, Salvatore;O'Flynn, Brendan
Publication date	2024-03-11
Original Citation	Peres da Silva, C., Tedesco, S. and O'Flynn, B. (2024) 'EEG datasets for healthcare: A scoping review', IEEE Access, 12, pp. 39186-39203. https://doi.org/10.1109/ACCESS.2024.3376254
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1109/ACCESS.2024.3376254
Rights	© 2024, the Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0 - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-06 11:41:15
Item downloaded from	https://hdl.handle.net/10468/15714



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 December 2023, accepted 26 February 2024, date of publication 11 March 2024, date of current version 19 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376254

SURVEY

EEG Datasets for Healthcare: A Scoping Review

CAROLINE PERES DA SILVA[1,2], (Member, IEEE), SALVATORE TEDESCO[1], (Member, IEEE), AND BRENDAN O'FLYNN[1], (Senior Member, IEEE)

[1]Tyndall National Institute, University College Cork, Cork 21, T12 R5CP Ireland
[2]SFI Centre for Research Training in Artificial Intelligence, University College Cork, Cork 21, T12 R5CP Ireland

Corresponding author: Caroline Peres da Silva (caroline.peres@tyndall.ie)

This work was supported in part by the Science Foundation Ireland (SFI) Centre for Research Training in Artificial Intelligence under Grant 18/CRT/6223; and in part by SFI under Grant 12/RC/2289-P2 (INSIGHT), Grant 13/RC/2077 (CONNECT), and Grant 16/RC/3918 (CONFIRM), which was co-funded by European Regional Development Fund (ERDF).

ABSTRACT The rapidly evolving landscape of artificial intelligence (AI) and machine learning has placed data at the forefront of healthcare innovation. Electroencephalography (EEG) has gained significant attention for its potential to revolutionize healthcare applications. However, the effective utilization of EEG data in advancing medical diagnoses and treatment hinges on the availability and quality of relevant datasets. In this context, we conducted a scoping review to explore the wealth of EEG datasets designed for healthcare applications. This review serves as a critical exploration of the current landscape, aiming to identify datasets related to healthcare conditions while assessing their reusability. Our findings highlight both the opportunities and limitations in the wealth of open access EEG datasets. Available. As AI increasingly relies on high-quality, well labelled data, barriers impeding the sharing and utilization of EEG data for healthcare (such as lack of comprehensive documentation or adherence to FAIR principles) must be addressed so as to leverage the potential of advanced deep learning models to unlock new possibilities for diagnosis and analysis of a wide array of medical conditions.

INDEX TERMS Electroencephalography, EEG, deep learning, open access, data sets.

I. INTRODUCTION

Electroencephalography (EEG) is a method of recording the electric field generated by postsynaptic activity of large groups of neurons in the cerebral cortex. This is generally accomplished by placing electrodes on the subject's scalp. The EEG signal recorded is an oscillating continuous signal whose characteristics change depending on the consciousness and general neural activity of the subject [1].

In general, there are two types of EEG recording: rest and task. EEG recordings carried out during rest usually involve participants being seated on a chair and instructed to relax and keep their eyes open or closed. Task recordings, instead, are those where the participants are instructed to perform a specific task (i.e., cognitive, motor, etc). Both rest and task recordings are the direct measurement of EEG without the need for stimulations or markers.

On the other hand, Evoked Potentials – EP (also known as event-related potentials, ERP) are the category of EEG recordings where a stimulus, usually audio or visual, is shown to the subject and a marker is sent to the EEG measurement device to record when the stimuli presentation started. These stimuli cause a specific known reaction in the recorded EEG, such as the P300 wave for visual recognition [1] or a driving response for oscillating stimuli where the EEG wave synchronizes with the stimulus frequency [2]. The EEG signals are time-locked to the stimuli, and appear as somatosensory, visual, or auditory potentials. For example, the P300 wave is an ERP that indicates that the subject has detected the target item in a sequence of non-target stimuli items.

From a medical point-of-view, EEG has a long history of being used to study and detect abnormalities in participants. One of the first documented experiments of EEG recording in humans was detecting EEG changes during sleep [3], [4], pointing to the existence of different phases of sleep. Today, it is used in a clinical setting for epilepsy diagnosis [5], Alzheimer's Disease diagnosis [6], [7], sleep studies [8], [9], [10], and monitoring depth of anaesthesia, coma, and other brain states [11]. It is also used in research for other neurological and psychiatric conditions [12], [13], [14]. Brain-Computer Interface (BCI) is an active area of research,

The associate editor coordinating the review of this manuscript and approving it for publication was Berdakh Abibullaev.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

and it uses EEG to allow people to control a computer even when paralyzed or with neuromotor difficulties [15]. BCI also studies how to identify emotional states in human participants [16]. A comprehensive overview of BCI methods of EEG analysis is given in [17].

With the development of mobile wearable devices, EEG headsets are now more available and accessible for both consumers and researchers. At the same time, the development of novel deep learning (DL) models has made possible to automatically classify and diagnose diseases and conditions. For example, DL models have achieved lower error-rate in cancer diagnosis than humans [18]. Specifically for EEG analyses, DL models have a better performance that traditional non-DL methods [19].

High-quality datasets are the cornerstone of successful DL algorithms [20]. Although many scientists are in favour of data sharing, there are still barriers to the practice being standard [21], [22]. The main reasons identified include lack of time, funding, and skills. Researchers are also wary of the quality of the data collected by others, as there is no way to be certain of the skill level of the data collectors, or that the correct methodology was followed during the data capture. Also, concerns regarding data documentation exist when discussing the reusability of datasets. Contextual information on the data is often necessary to fully understand and replicate analysis, but this creates a higher burden to researchers when publishing data [21].

However, sharing data used in research studies is one of the fundamental mechanisms for combating the reproducibility crisis which impacts on the use of AI in the health domain [23], [24], [25]. This is particularly important in the field of machine learning, as the models developed depend on the data used to train it. Even if one has access to the original data, it is not always possible to reproduce ML results [26], [27]. Indeed, when training models, there are various settings that can be passed to the software packages that alter how the model is trained. Added to that, deep learning models are trained via stochastic non-deterministic methods. As Summers and Dinneen identified [28], if the initialization parameters are even slightly different, the same model with the same data produces vastly different results. Therefore, to fully reproduce a DL model one needs not only the data, but the training methodology used to develop the model which includes the software code, initializations, and all parameters involved.

Availability of research data decreases with publication age [29], which means that older publications are less likely to have their data shared. The costs of hosting and maintaining data over time are barriers to making data available for years after publication. Also, Miyakawa [30] suggests that some results are not supported by raw data – 97% of 41 manuscripts that had been marked for review and asked by the editor to provide raw data were retracted or rejected. In some cases, raw data might be restricted due to what the ethics approval obtained to carry out any study allows to be shared.

There has been an increased emphasis in recent years to make science more open and reproducible. Efforts like the FAIR data [31] and the increased requirement of data management plans by funding organizations has increased the visibility of best practices in data sharing.

Considering this, to determine which datasets are available that contain EEG signals, and to understand which medical conditions are contemplated in these datasets, and how reusable these datasets are, the following scoping review was carried out by the authors.

The research questions being addressed in this scoping review are as follows:

RQ1 - Which datasets containing EEG signals focused on healthcare applications (non-epilepsy related) are available?

RQ2 - Which health conditions are covered in these datasets?

RQ3 - How reusable are these datasets?

Epilepsy diagnosis and seizure detection are thoroughly discussed in [32] and [33], thus datasets containing epileptic patients or seizure signals were excluded from this review. Also, one of the largest EEG datasets is the TUH-EEG dataset [34], which contains 26,846 clinical EEG recordings and annotations. Since it contains seizure events, it has been excluded from this review. It is available on https://isip.piconepress.com/projects/tuh_eeg/.

A significant portion of the available EEG datasets are not healthcare related; instead, those datasets are the result of work in BCI and Affective / Emotion detection. The authors suggest any interested readers to read the publication [19] for information regarding general EEG datasets, [16] and [35] for data on emotion recognition, [36] for motor imagery, and [37] for depression, while [19] and [38] provide an in-depth review of deep learning and EEG.

II. METHODS

This review follows the PRISMA guidelines for Scoping Reviews [39]. The methodology for the review is detailed below.

A. SEARCH STRATEGY

The search was conducted using the following publication databases: PubMed, Web of Science, and Scopus. On these databases, the search focused on published articles that describe and/or analyse an EEG dataset. The search was also conducted on the following data portals: Data Cite (which include results from IEEE DataPort, FigShare, and Zenodo) and Mendeley Data. Further datasets were screened from known EEG dataset repositories, listed in Table 1. The search was conducted on the week of 14/Nov/2022.

The following search string was used for publications:

(((dataset[Title]) OR (database[Title]) OR (repository[Title]) OR (data[Title])) AND (available) AND (health*) AND (EEG or electroencephalography) NOT (epilepsy OR seizure OR "brain-computer interface" OR emotion))

TABLE 1. List of EEG repositories screened for suitable datasets.

Repository	URL
Publication databases	
PubMed	https://pubmed.ncbi.nlm.nih.gov/advanced/
Web of Science	https://www.webofscience.com/wos/woscc/advanced-search
Scopus	https://www.scopus.com/search/form.uri?display=advanced
Databases of data-sharing portals	
DataCite	https://commons.datacite.org/
Mendeley Data	https://data.mendeley.com/
Neurology datasets repositories	
OpenNeuro	https://openneuro.org/
PhysioNet	https://physionet.org/
DANDI Archive	https://dandiarchive.org/
PREDICT	http://predict.cs.unm.edu/

The following search string was used for data portals:

health TITLE(EEG OR electroencephalogram OR electroencephalography) DATA_TYPE (DATASET) NOT epilepsy NOT emotion NOT (BCI OR "Brain-Computer Interface") NOT seizure

B. INCLUSION AND EXCLUSION CRITERIA

Publications were deemed eligible if they were describing a dataset. The datasets which were deemed eligible were the ones with a valid and unique DOI reference.

The inclusion criteria for datasets were:

- dataset has a valid and unique DOI;
- the dataset description (in the case of a publication) has a link or direct reference to the dataset;
- dataset can have any type of access (open, restricted, request);
- the dataset must include EEG signals;
- dataset may include other types of data, but it is not required;
- dataset must have been collected to study a health condition, disease, or diagnostic;
- dataset must include information about each recording, such as to which group (i.e., diagnosis, or patients versus healthy controls) the recording belongs to.

The exclusion criteria for datasets were:

- in the case of publications, if the publication uses the dataset but does not describe it;
- data collected for interventional studies, i.e., the study is analysing the results of an intervention (medication or treatment) to a group of participants;
- datasets investigating epilepsy and/or seizures;
- non-health-related dataset, such as general BCI datasets, motor BCI, emotion recognition, meditation, etc.;
- datasets with no reference to ethical approval for the study (either in the dataset description or in its related publication);

The datasets were found via three methods: (i) searching publication databases for articles that described a dataset; (ii) searching for datasets through data citation and data aggregation portals (e.g., DataCite); and (iii) by searching published datasets in repositories dedicated to neurological and physiological datasets. These specific repositories (such as OpenNeuro, PhysioNet, etc.) were identified from searching for EEG repositories in a general search engine (Google).

The results were first screened by title, keywords, and description, if these were available for the datasets, or alternatively from the abstract in the case of publications. In this step, datasets containing no mention of EEG signals were excluded. From the keywords and description, datasets were excluded if there was no mention of a medical condition or state, or if there was mention of epilepsy and/or seizures, according to the exclusion criteria. After this first screening, the full text of articles describing datasets were retrieved and surveyed. If no reference to a dataset was found (either a link or citation with DOI), it was rejected. The other inclusion and exclusion criteria discussed above were also assessed. For datasets, the repository page with metadata was accessed and screened for suitability. Datasets were also rejected in the case that their filenames were not in English, or not descriptive enough to allow understanding of the contents of the file and to which class (patients or controls) the file belonged. One of the datasets identified in the first screening was not yet published, which made it impossible to retrieve its full metadata, and therefore it was also excluded. This whole process is shown in Figure 1, where the number of records retrieved, screened, and rejected or included are shown.

For each dataset included in this review, the related publication was found usually linked in the dataset description or indicated as the preferred method of citation for the dataset. In the cases that the publication was not linked, the corresponding articles were searched via the authors listed in the dataset metadata. This was carried out because the data collection protocol, for most of the datasets, was usually not described in the dataset description or files. Instead, it was usually described in the methodology section of the related publication. For all datasets included in this review, the information of where the data collection protocol was described is included in the data extracted.

C. EXTRACTED INFORMATION

For each included dataset, the following information was extracted (either from the dataset link or its related publication):

- dataset reference (DOI and full citation);
- related publication reference(s);
- dataset URL (unique);
- access type;
- country and data collection site;
- year of publication;
- health condition or diagnosis of focus;
- data modalities contained in the dataset (such as subject demographics, clinical history or Electronic Health Record - EHR, psychometrical or psychological test, physiological signals, etc.);
- population included and number of participants for each group;

- number of classes in the dataset (e.g., patients versus controls, sleep stages, patient subgroups);
- type of EEG data collection: rest, task, evoked potentials;
- number of recording sessions, recording time, number of trials/tasks, and duration;
- mode of EEG data annotation: manual, task/stimulation/event markers, automatic/algorithm/machine;
- EEG device used and form factor (i.e., cap, free electrodes, headset/headband, polysomnography – PSG - setup), number of channels, electrodes placement and references, sampling frequency;
- EEG data format (.eeg, .bdf, BIDS, etc);
- if the EEG data provided is raw or is pre-processed (and in this case, which pre-processing has been implemented);
- whether the recording protocol is available in the dataset description/files, the associated publication, or not available;
- whether the dataset includes an explanation of how to analyse it, files describing the data/variables/etc, or code used by the researchers to analyse the dataset.

D. ASSESSMENT OF RECORDS

For each dataset included in this review, we assessed the reusability of the dataset to answer research question 3. This was carried out using two methods: the first was an automated tool developed by the FAIRsFAIR Project [40], and the second method was via a questionnaire completed by the authors when extracting metadata from the datasets.

The FAIR guiding principles [31], Findability, Accessibility, Interoperability, and Reusability, inform researchers and data producers and publishers on the best approaches to publish data and other research objects. The FAIRsFAIR project [41] is an European project that aims to develop global standards for FAIR certification of repositories. Developed by the FAIRsFAIR project, the "F-UJI Automated FAIR Data Assessment Tool" [42] is a REST web service tool to programmatically assess a data object according to the FAIRsFAIR Data Object Assessment Metrics [43]. These metrics were developed by the FAIRsFAIR project to provide a method to systematically assess a data object. For this review, we used the version v2.0.2 of the F-UJI tool, available at https://github.com/pangaea-data-publisher/fuji/releases/tag/v.2.0.2 . The F-UJI web server was run in a computer, and programmatically queried for each dataset contained in this review. The code used for this is available in the Supplementary Materials, and at *xxxxx (permanent link to repo)*.

The F-UJI Automated tool reports a score for each of the FAIR assessment metrics [43] grouped by the principles *Findable*, *Accessible*, *Interoperable*, and *Reusable*. In this review, we report it in terms of percentage: the highest score possible in one of these metrics is shown as 100%, and the lowest score (zero) is shown as 0%.

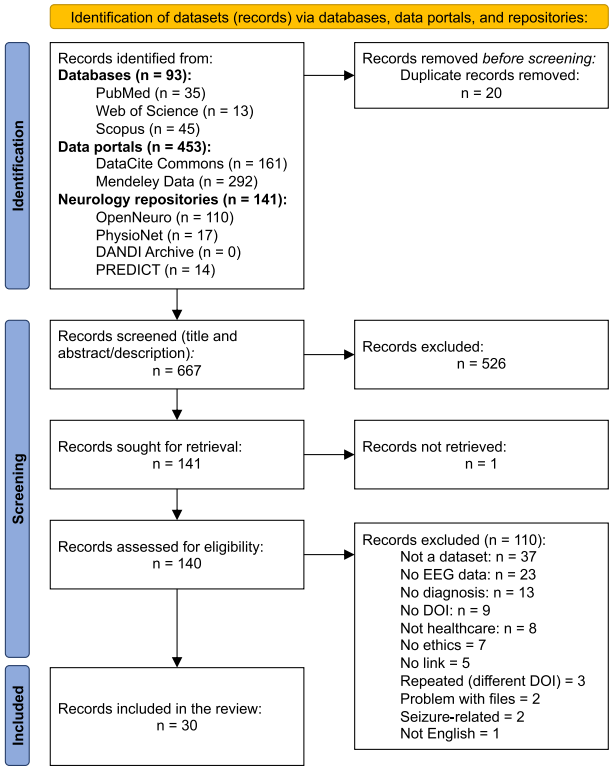


FIGURE 1. Dataset selection for inclusion using the PRISMA framework.

Considering the FAIR data requirements, information on the documentation level of the dataset and the file format provided was also collected. For this, the file format of the data files was recorded during the data extraction. A scale of 0 to 3 points was created to determine the documentation level of the dataset — this represents how well the context of the data provided is explained, i.e., if the dataset contains the data collection protocol, the methodology, and the code used for analysis. For each type of documentation present in the dataset its score would gain one point. In this way, a score of zero means the dataset has no contextual information (i.e., only the data files and a brief abstract is provided), and a score of 3 means the dataset includes the collection protocol, an explanation of the file structure and internal variables, and the code used for the analysis. The datasets that provide the collection protocol only in the associated publication have been noted and a 0.5 score deduction was applied.

III. RESULTS

A. OVERVIEW

From the 141 records (datasets) sought for retrieval one dataset was not published yet, so it was not retrieved. Of the 140 records assessed for eligibility 37 were journal articles that were not describing a dataset, 23 records did not have EEG data, 13 records were not exploring a specific diagnosis / condition, 9 did not have a DOI, 8 were general BCI datasets (not healthcare related), 7 datasets had no ethics disclosed either in the dataset metadata or the associated publication, 5 had no link or reference to where to find the data, 3 were

TABLE 2. List of EEG datasets included in this review.

Ref.	Year	Condition / focus	Population	Access	Licence
[44]	2020	ADHD	34 ADHD, 25 HC (all aged 9-16yr)	Open	CC BY 4.0
[45]	2020	Alzheimer's Disease	34 SCD, 79 MCI, 48 AD, 33 HC	Restricted	CC BY 4.0
[46]	2021	Autism Spectrum Disorder	28 ASD, 28 HC	Open	CC BY 4.0
[47]	2017	Brain Injury	14 patients PCF lesions, 20 HC	Credentialed	n/d
[48]	2021	Brain Injury	38 sub-acute mTBI, 23 chronic TBI, 24 HC	Open	CC0
[49]	2022	Chronic Post-Burn Itch	15 patients, 14 HC	Restricted	CC BY 4.0
[50]	2021	Depression	75 HC, 46 depressive symptom group	Open	CC0
[51]	2022	Depression	24 patients (depression), 29 HC	Open	n/d
[52]	2022	Depression	76 HC, 46 depressive symptom group	Open	CC0
[53]	2018	Hearing Impairment	17 normal hearing younger adults (18-30), 14 normal hearing older adults (>=60), 17 hearing impaired older adults (>= 60)	Open	n/d
[54]	2021	Migraine	17 migraineurs, 18 healthy controls	Open	CC BY 4.0
[55]	2017	Multiple	children and young adults	Credentialed	CC BY NC SA 4.0
[56]	2021	Parkinson's Disease	14 PD w/o FOG, 14 PD w/ FOG, 13 HC	Credentialed	CC BY 4.0
[57]	2021	Parkinson's Disease	15 PD, 16 HC	Open / Request	CC0
[58]	2021	Parkinson's Disease	28 PD, 28 HC	Open	CC0
[59]	2021	Parkinson's Disease	28 PD, 28 HC	Open	CC0
[60]	2021	Parkinson's Disease	25 PD, 25 HC	Open	CC0
[61]	2021	Phobia	9 claustrophobic, 13 HC	Open	CC BY 4.0
[62]	2022	Phobia	70 patients, 53 controls	Open	CC BY 4.0
[63]	2017	Schizophrenia	14 paranoid schizophrenia, 14 HC	Open	CC0
[64]	2022	Schizophrenia	19 patients w/ psychosis episodes, 24 HC	Open	CC0
[65]	2022	Schizophrenia	31 first-episode schizophrenia spectrum psychosis, 22 HC	Open	CC0
[66]	2012	Sleep	16 HC, 2 bruxism, 9 insomnia, 5 narcolepsy, 40 nocturnal frontal lobe epilepsy, 10 periodic leg movement, 22 REM behavior disorder, 4 sleep-disordered breathing	Open	ODC By 1.0
[67]	2016	Sleep	100 participants (one recording), 8 participants (2 recordings), 10 healthy controls	Open	n/d
[68]	2018	Sleep	22 participants	Open	BSD 3-Clause
[69]	2021	Sleep	3,984 paediatric sleep studies on 3,673 unique patients	Credentialed	PhysioNet Credentialed Health Data License 1.5.0
[70]	2021	Sleep	22 infants	Credentialed	n/d
[71]	2022	Sleep	151 recordings	Open	CC BY 4.0
[72]	2022	Sleep	33 healthy	Open	CC0
[73]	2021	Stroke	91 able-bodied, 65 post-stroke participants	Open	CC0

* HC: healthy controls, ADHD: Attention Deficit Hyperactivity Disorder, SCD: Subjective Cognitive Decline, MCI: Mild Cognitive Impairment, AD: Alzheimer's Disease, ASD: Autism Spectrum Disorder, PFC: Pre-Frontal Cortex, TBI: Traumatic Brain Injury, mTBI: mild TBI, PD: Parkinson's Disease, FOG: Freeze of Gait
* n/d: not disclosed, n/a: not applicable

repeated (same dataset but different DOIs), 2 were seizure related, 2 had no explanation about what each file represented, and 1 was not in English. In total, 110 records were excluded and 30 records included.

In general, the datasets included fall into three broad categories of condition of focus: sleep studies and related disorders, psychiatric conditions, and neurological conditions. Sleep and related issues have the biggest number of datasets found for a condition, with seven datasets. The number of datasets found for each condition is shown in Figure 2a. Two datasets were found related to phobias, each focusing on different phobias (claustrophobia and arachnophobia) but were placed under the same condition. One dataset contains multiple conditions for the paediatric population (developmental disorders), thus it was not included in any single condition.

The number of datasets published in each country is shown in Figure 2b. Some datasets were collected by researchers collaborating in multiple locations, so the same dataset was counted for each country in the collaboration. For example, [73] was collected in multiple sites in Italy, Germany, and Switzerland, and [47] was collected in the USA and Norway.

The year of publication for each dataset was also recorded and shown in Figure 2c. The number of published datasets has increased in the last years, which could be due to recent efforts on open science and data sharing.

The main data modalities other than EEG included in the datasets are shown in Figure 2d. The most common data modalities included are physiology related: electromyography (EMG), electrocardiography (ECG), and electrooculography (EOG). Due to the considerable number of sleep datasets in this review, a lot of the included types of data are pertaining to PSG recordings, such as EMG to detect muscle activity, ECG to capture heart rate metrics, EOG to record eye movements, respiratory rate, airflow, audio recording for snoring, body position, oxygen saturation (SaO2), and CO2 measurement. Structural Magnetic Resonance Imaging (MRI) and functional MRI (fMRI) are also included in some datasets for further brain analysis. Other than the modalities listed above, no other physiological data collection was found in the datasets included in this review.

The following sections will list and discuss the datasets included in this review. The datasets are organized according to the health condition of focus on (as per the research question RQ2), grouped by three broad categories: sleep studies, psychiatric conditions (such as depression, schizophrenia, among others), and neurological conditions (such as Parkinson's and Alzheimer's Disease).

B. CONDITIONS

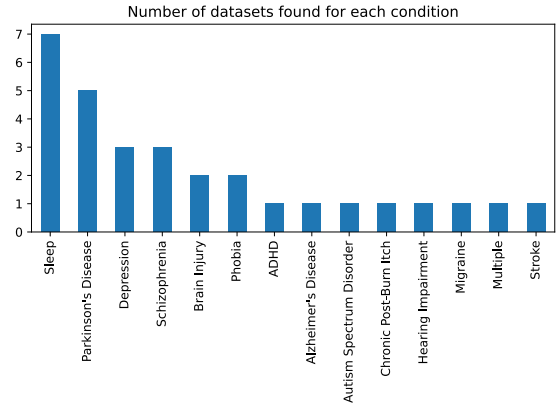
1) SLEEP

The CAP Sleep Database [66] is a dataset of polysomnographic recordings from 108 participants – 16 healthy controls, and other sleep related pathologies, such as insomnia, narcolepsy, sleep-disordered breathing, among others. Each recording contains three or more EEG channels, EOG, EMG, airflow, respiration, SaO2, and ECG. The EEG recordings are annotated with sleep stage and cyclic alternating pattern (CAP) events which are correlated with sleep instability and related pathologies. The associated publication [74] describes how to classify CAP events, and how they present in an EEG recording.

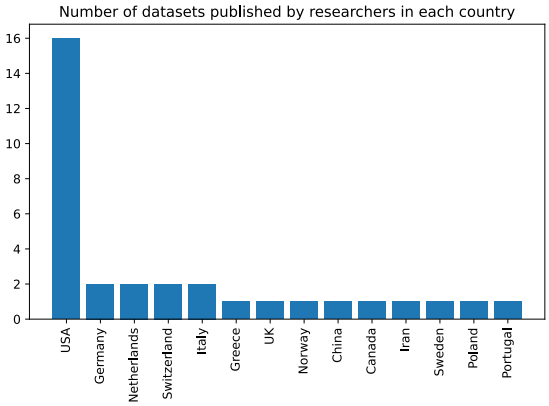
The ISRUC-Sleep Dataset [67] consists of polysomnography recordings of 108 participants with sleep disorders, of which 8 have two different recording sessions, and 10 healthy participants. Each recording contains 6 EEG channels, EOG channels for each eye, EMG (measured on chin and legs), ECG, airflow, respiration, SaO2, and body position. For each subject, the dataset includes details such as demographic data (age and gender), medication taken, and sleep scoring. The associated publication describes in detail the dataset and a method of automatic sleep stage classification and its performance evaluation compared with manual expert annotation. The automatic sleep stage classification method is comprised of preprocessing, feature extraction based on maximal overlap discrete wavelet transform (MODWT), and a supervised learning step which uses a Support Vector Machine classifier.

The dataset described in [68] is composed of recordings of 64-channel EEG from 22 participants during short periods of sleep, with manually annotated sleep stages and spindles (specific pattern of EEG that occurs during the second part of REM sleep). Recordings were carried out on two separate days, after the participants undertook a high- or low-load visual working memory task. Added to the EEG raw data, the authors also provide code in python for signal processing and artifact correction using Independent Component Analysis (ICA). This is especially useful for reusing the dataset, as it enables researchers to replicate the authors' data processing and verify their findings. It also enables other researchers to build solutions based on the data collected.

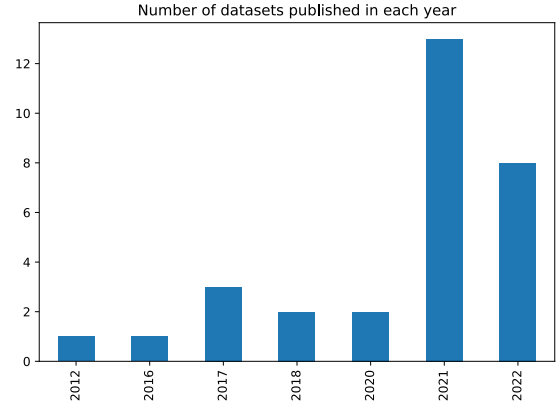
The NCH Sleep DataBank [69] is a large dataset containing 3,984 paediatric sleep studies on 3,673 unique patients associated with EHR, which includes medications, measurements, diagnoses, etc. The data modalities included are EEG for sleep stage identification, EMG channels on chin and leg, EOG, ECG, airflow and respiratory effort, blood oxygen saturation, and carbon dioxide measurement of exhaled air. The data is annotated in real-time by technicians and reviewed later by another expert. This dataset has restricted access by means of credential assignment. Code in python for data analysis is also provided in a separate github repository [75]. The publication [76] specifies the transformations carried out on the raw data (filename, EDF header, random date shift, etc.) to anonymize patients. The publication is a full description of the dataset and what is included in it, going into detail such as file naming conventions, variables included in the files, and how annotation was made. The description of files included is helpful to potential users as they can identify what is included and if it is relevant for their purposes. Also, due to the huge amount of annotated raw EEG data, it is very well suited for the development of deep learning



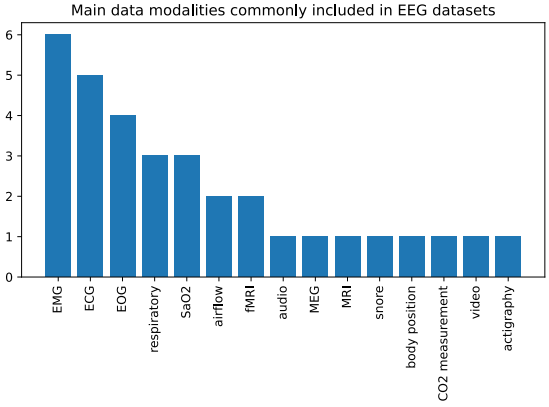
(a) Number of datasets in this scoping review for each health condition.



(b) Number of datasets published by country.



(c) Number of datasets published in each year. No datasets were found published in 2013, 2014, 2015, and 2019.



(d) Data modalities included in the datasets.

FIGURE 2. **Distribution of datasets found for each health condition, country, year of publication, and secondary data modalities included.**

models – which the authors highlight in the publication. They also suggest potential applications of the data and provide starter code for analysis.

The dataset in [70] is another dataset dealing with paediatric sleep which contains sleep recordings of infants (between one week and seven months of age), with a high-density EEG cap and video to measure movements. The access to the files in this dataset is restricted to credentialed authorized users. Analysis of the data is carried out in the associated publication [77]. The publication reports that the MATLAB code used for the analysis is shared in a GitHub repository, but as of April of 2023, the repository does not exist. Data collection protocol is described in the publication.

The Haaglanden Medisch Centrum sleep staging database [71] is comprised of 151 whole night polysomnography recordings, containing EEG, EOG, chin EMG and ECG, including the manual annotations of technicians. The dataset webpage describes the data collection protocol and the data description of what each file contains and their formats, as well as usage notes on how to open and interact with the files provided. The associated publication [78] presents a deep learning model for automatic sleep staging using this dataset and other open sleep datasets.

The dataset in [72] includes EEG and fMRI signals collected from 33 healthy participants during resting state and sleep sessions, before and after a visual-motor adaptation task. The data collection protocol is briefly described in the dataset description, but expanded and detailed on the associated publication [79].

2) PSYCHIATRIC CONDITIONS

a: DEPRESSION

The dataset in [50] and [52] collected EEG recordings of 122 university students aged 18-25 years during rest [50], with eyes open and eyes closed, and during cognitive tasks [52]. The participants were screened for major depressive disorder (MDD) using the Beck Depression Inventory, and classified in one of 4 groups: MDD, past MDD diagnosis (but not current), not meeting the diagnosis criteria for MDD, and not tested for MDD. The article associated with this dataset is [80] which explores how depression and anxiety relate to reward systems in the brain. The dataset is provided in the BIDS format, which is a standard in the field [81], [82]. The data collection protocol for this dataset is not described fully, and the dataset description notes that some of the data might be mislabelled, some EEG channels have been

TABLE 3. Data modalities included in each dataset.

Ref. Data modalities	Demographics	Clinical History	Psychometric	Physiological	Neuroimaging
[44] EEG					
[45] EEG	✓	✓			
[46] EEG					
[47] EEG					
[48] EEG	✓	✓	✓		
[49] EEG	n/d				
[50] EEG	✓		✓		
[51] EEG, audio	✓		✓		
[52] EEG	✓		✓		
[53] EEG, MEG					
[54] EEG	✓	✓			
[55] EEG, MRI	✓	✓	✓		
[56] EEG, EMG					
[57] EEG	✓	✓	✓		
[58] EEG	✓	✓			
[59] EEG	✓	✓	✓		
[60] EEG	✓	✓	✓		
[61] EEG	✓				
[62] EEG					
[63] EEG					
[64] EEG					
[65] EEG	✓	✓	✓		
[66] EEG, EOG, EMG, respiratory, SaO2, ECG	✓	✓			
[67] EEG, EOG, EMG, ECG, snore, airflow, respiratory, SaO2, body position	✓	✓	✓	✓	
[68] EEG	✓		✓		
[69] EEG, EMG, EOG, ECG, airflow, respiratory, SaO2, CO2 measurement	✓	✓			
[70] EEG, video	✓				
[71] EEG, EOG, EMG, ECG					
[72] EEG, fMRI					✓
[73] EEG, ECG, EMG, actigraphy, fMRI	✓	✓		✓	

* EMG: electromyography, ECG: electrocardiography, EOG: electrooculography, PSG: polysomnography, SaO2: oxygen saturation, MRI: Magnetic Resonance Imaging (MRI), fMRI: functional MRI.

interpolated, and no raw data is available leading to some question as to its useability in future work. The cognitive task part of the dataset contains the associated code to produce the stimuli presented to participants, and the code developed in MATLAB to analyse the data and replicate the study.

Another dataset developed in the study of depression is [51] which provides EEG and audio data from 24 patients clinically diagnosed with depression, and 29 matched healthy controls. The EEG recordings were created with a standard medical 128-channel device, and a novel 3-channel wearable device, during rest and activities such as responding to questions, reading, and picture description. The dataset is provided in the BIDS format, and contains a README file, a file describing the methodology used for the data collection, and the information sheet given to patients for consent in participating in the study. The publication associated with this dataset is [83] which describes the dataset in detail.

b: DEVELOPMENTAL DISORDERS: ASD, ADHD, LEARNING DISORDERS, ETC.

The dataset in [44] contains the raw data from a study [84] on working memory and response inhibition. This study was carried out on participants aged 9-16 years old, 34 of those diagnosed with attention deficit hyperactivity disorder (ADHD) and 25 typically developing participants. The protocol for the data collection is in the publication, and no other demographic data is given. In the study, the participants had to perform working memory tasks and response inhibition tasks. The EEG data is from 21 channels sampling at 500 Hz, provided in `.cnt` format.

Another dataset [46] has EEG data acquired from adult individuals aged 18-68 years, 28 of those with a diagnosis of autism spectrum disorder (ASD), and 28 neurotypical controls. The study investigated how brain aging is different for adults with ASD compared to healthy controls. The data was collected during rest for 150 seconds of eyes closed for each subject. The EEG data is provided for each subject with a pair of files `.fdt` for the raw EEG, and `.set` with details on the recording parameters. However, no information is provided for which participants have ASD and which are the controls.

The dataset "Healthy Brain Network (HBN) Biobank" [55] includes data on over 4,000 participants aged 5 to 21 years old. The data modalities are EEG, MRI, behavioural and cognitive phenotyping, actigraphy, eye tracking, genetics, audio, and video. The publication details the complete methodology of data collection: subject screening, inclusion and exclusion criteria, assessment tests, and recording protocols for each data modality. For each different diagnosis included in the dataset, the respective assessment tests that

are relevant are mentioned. The assessments included are anxiety, ADHD, ASD, cognitive and executive functioning, depression and mood, obsessive-compulsive disorder, physical tests, sleep, family structure and trauma, substance abuse or addictive behaviour, verbal learning, and others. Added to that, the publication also contains the lessons learned over the implementation of a large-scale data collection, which is of great benefit for other researchers also collecting large amounts of data. The access to the full dataset is restricted by request / project submission for validation.

c: SCHIZOPHRENIA

The dataset in [63] consists of EEG recordings of 14 patients with paranoid schizophrenia and 14 healthy controls. In the dataset repository, there is only information on the sampling frequency and electrodes used. The groups are identified via the filename, starting with `h` or `s`. The publication that analyses this dataset [85] describes the protocol used for the EEG recordings and the analysis implemented comparing the controls to patients using connectivity measurements extracted from the EEG data.

The dataset in [64] is an EEG dataset of 19 patients and 24 healthy controls recorded during a task of ultimatum game. The protocol for the data collection is described in the associated paper [86]. EEG was recorded with 128-channels at 2048 Hz. No other information about patients or about data pre-processing is available.

Another dataset for schizophrenia is divided into two DOIs: [65] and [87]. In this dataset, EEG and MEG (magnetoencephalography) data were recorded for two independent samples of healthy controls and First Episodic Psychosis (FEP) individuals. For each subject, the dataset includes 5 minutes of resting EEG collected using a 60-channel cap setup. The associated paper [88] has the test protocol used and analyses run on the collected data.

d: PHOBIAS

For the study of reactions of people suffering from phobias, two datasets were found.

The first dataset [61] consists of 9 participants with self-identified claustrophobia and 13 healthy controls. The EEG was recorded under 3 different conditions: in a well-lit spacious room, in a chamber with moderate light, and a moderate lit smaller room. Information about each participant, such as age, gender, and group are present in the dataset. The associated publication [89] describes in detail the dataset organization and files, and the data collection protocol.

The second dataset in [62] consists of EEG recordings of 40 patients with arachnophobia and 53 controls. No demographic information is given for the participants. The data collection protocol is detailed in the associated publication [90]. However, the dataset includes the code used for the analysis done for the publication, which allows other researchers to review the methodology and replicate results.

3) NEUROLOGICAL CONDITIONS

a: PARKINSON'S DISEASE

The dataset in [56] studies freezing of gait in Parkinson's Disease. The test cohort consists of 14 participants with Parkinson's Disease experiencing gait freezing, 14 patients with Parkinson's Disease but without freeze of gait, and 13 healthy controls. No data other than to which group the participants belong to is presented in the dataset. The associated paper [91] describes the data collection protocol, i.e., how the participants were screened, and the motor task carried out during the EEG recordings. The participants were recorded performing ankle dorsiflexion while sitting in a chair and with gaze fixed at a specified point.

The dataset in [57] is a curated dataset derived from [92]. It contains EEG recording of 15 patients with Parkinson's Disease and 16 healthy controls. The dataset provides other information regarding participants, such as age, gender, handedness, and clinical history. The data collection protocol is explained in [92], which consisted of rest and a stop-signal task. The dataset curators ask that researchers using the dataset approach them via email, as they have published the dataset with open access, but would like to have people ask for permission and guidance to use.

The three datasets in [58], [59], and [60] were all collected at the same institute by the same research group. The dataset consist of EEG recordings for 28 participants with Parkinson's Disease and 28 healthy controls (the last dataset [60] contains recordings of just 25 participants for each group). The EEG was collected during a cognitive task [58], reinforcement learning [58], and rest and auditory stimuli [60]. Some brief description of the data collection protocol is present in the dataset descriptions, and the author cites an article that explains the task, as well as supplies the Matlab code necessary for replicating it. Also present in the dataset are other code files for further data analysis. Results from analysis were published in [93] for the first dataset, [94] for the second dataset, and [95] for the third.

b: ALZHEIMER'S DISEASE

For Alzheimer's Disease, the dataset [45] includes 230 participants, with 5 subgroups: 33 healthy control elders, 34 with subjective cognitive decline, 79 with mild cognitive impairment, 48 participants with Alzheimer's Disease, and 36 healthy young controls. The full dataset has restricted access, but 4 samples are included in the open version. The dataset description is very informative on the data collection protocol, EEG device and setup, and preprocessing done to the data. The analysis of this dataset was published in [96].

c: BRAIN INJURY

The dataset in [47] contains EEG recordings of 14 patients with unilateral prefrontal cortex lesions and 20 matched healthy controls. Raw and pre-processed data are included, as well as analysis code to replicate results. The data collection protocol and results are published in [97]. The

access to the dataset is credentialed — e.g., the user needs to create an account to download the dataset.

The dataset in [48] consists of EEG recordings of patients with traumatic brain injury (TBI). The population is composed of 23 chronic TBI patients, 38 sub-acute mild TBI, and 24 healthy controls. The dataset also contains demographic data, and neurophysiological assessments. Some of the data collection protocol is explained in the dataset description, but a full detailed explanation is present in the associated article [98]. The task performed in this dataset generated auditory evoked potentials using a 3 stimulus oddball paradigm (where presentations of the different stimuli have different probabilities of occurring), and EEG was also collected during rest.

d: OTHERS - COGNITIVE IMPAIRMENT, HEARING IMPAIRMENT, STROKE, MIGRAINE, ETC.

The dataset in [73] is a multi-center (Italy, Germany, and Switzerland) dataset of various modalities of data collected to study upper-limb movements. The dataset contains 65 post-stroke participants and 91 healthy participants. Data modalities included are EEG, ECG, EMG, actigraphy and kinematic data, and fMRI. The publication [99] has all of the details of data collection procedures and what is contained in the dataset. The dataset also contains a Matlab script to plot data.

The dataset in [54] contains high-density (128-channels) EEG recordings of 17 migraineurs and 18 healthy controls. The recordings were obtained for rest and evoked potentials, both visual and auditory. Included in the dataset files are README files with description of the files, Matlab code used to generate the visual and auditory stimuli, and a document describing the data collection protocol in detail. The analysis of the data is found in the publication [100].

EEG and MEG recordings to study hearing impairment are present in the dataset [53]. It contains recordings of 17 normal hearing younger adults (18-30 years old), 14 normal hearing older adults (over 60 years old), and 17 hearing impaired older adults (over 60 years old). The dataset description contains an overview of the data collection protocol and file naming convention. More details and the analysis of the data are found in the publication [101].

The dataset in [49] investigates brain functionality involved in chronic post-burn itch. It contains EEG data from 15 patients and 14 healthy controls recorded during rest and stimulation of skin. The dataset also contains documents describing the collection protocol, SPSS scripts for statistical analysis, and the literature review completed by the authors. The access to the dataset is restricted, i.e., the user needs to request access to the dataset owner to download the files.

C. EEG CONFIGURATION

The Table 4 shows the configuration of EEG system used for the data collection of each dataset in this review, as well as the recording type, and annotation method.

D. ASSESSMENT: HOW REUSABLE ARE THE DATASETS?

The Table 5 shows the result of using the ''F-UJI Automated FAIR Data Assessment Tool'' [42] to programmatically evaluate the datasets. For each category (Findability, Accessibility, Interoperability, and Reusability), the value shown represents what percentage of the FAIRsFAIR Data Object Assessment Metrics [43] the dataset fulfils.

Since the automated tool is not able to test if the dataset uses the standard format for EEG data (BIDS), or determine the level of documentation of the dataset, we collected the information shown on table 6. It indicates if a dataset has a comprehensive explanation, a README file, some indication of the file structure used, if the filenames are descriptive enough to enable another researcher to identify what the file is about, where the data collection protocol is located (in a file inside the dataset files, in the associated publication, or not disclosed), and if there is any code provided by the authors to create the environment and stimuli used and/or the data analysis done.

IV. DISCUSSION

Electroencephalography (EEG) is a reliable clinical diagnostic method which can be used for different health conditions, such as epilepsy, Alzheimer's Disease, sleep disorders, and anaesthesia monitoring. On the research side, EEG is widely used for psychiatric, psychological, and neurophysiological studies of varied cognitive and psychological conditions. It is also used in Brain-Computer Interfaces, where the EEG signals can be used to control computers, speech devices, and other accessibility tools. Recent research in BCI also uses EEG to identify emotional states in participants. With these varied use-cases and combined with its relatively low-cost (when compared to other neurological diagnostic tools), EEG is a crucial tool for research in neurology and related areas.

Recent advances in Machine Learning and Deep Learning have opened up various new opportunities in research. However, these methodologies require large volumes of data. Open datasets greatly advance research on these methods.

A number of considerations (i.e., ethics, data availability, etc.) needs to be taken in to account when analysing such datasets as those selected in this review and these are discussed below.

A. ETHICS

Seven datasets were excluded from this review as they had not indicated whether or not they had ethics approval for the studies undertaken. In general, with some exceptions, the included datasets disclosed ethics approval only on the associated publication. Also, some of the datasets did not include a copyright disclaimer, whereas some of the repositories (e.g., PhysioNet) included copyright by default. These disclosures are important from a legal perspective, as it informs potential users of who owns the data, and what can be done with it. The most common copyright of the datasets included in this review is ''CC0'' in which the researcher gives up all the copyright, placing the dataset in the public domain, where it

TABLE 4. EEG data collection characteristics: device, number of channels, and sampling frequency used, type of recording and annotations, file format, and where the recording protocol is available.

Ref.	Type of recording	Device	Channels	Frequency (Hz)	Annotation	Format	Protocol
[44]	task: n-back / nogo	SynAmps amplifier	21	500	stim	.cnt	publication
[45]	rest	EGI 300 Geodesic EEG system	256	250	stim	.mat	publication
[46]	rest	Biosemi Active two EEG system	64	2048	none	.set + .fdt	publication
[47]	task: working memory	64 + 8 channel BioSemi ActiveTwo amplifier with Ag-AgCl pin-type active electrodes	64+8	1024	stim	.bdf	publication + file (dataset)
[48]	task: AEP	Brain Vision	64	500	stim	.set + .fdt	publication
[49]	rest, stim	ActiveTwo BioSemi system	32	1024	stim	.bdf	publication
[50]	rest	SynAmps amplifier	64	500	none	.set + .fdt	publication
[51]	rest, task	HydroCel Geodesic Sensor Net, 24bit A/D conv.	128, 3	250	stim	.mff , .txt	publication + file (dataset)
[52]	task: learning	SynAmps amplifier	64	500	stim	.set + .fdt	publication
[53]	rest	Biosemi ActiABR-200	5	16384	stim	.bdf	brief
[54]	EP (visual and auditory), rest	BioSemi ActiveTwo system	128	512	stim	.bdf	publication + file (dataset)
[55]	rest, task	EEG Geodesic Hydrocel	128	500	stim, manual	bids	publication + file (dataset)
[56]	task: motor	g.Nautilus, Guger Technologies	17	250	stim	n/d (.zip)	publication
[57]	rest, task	Active Two BioSemi system	32	512	stim	.bdf	publication
[58]	task	Brain Vision	64	500	stim, events	.set + .fdt	publication
[59]	task	Brain Vision	64	500	stim, events	.set + .fdt	publication
[60]	task: AEP	Brain Vision	64	500	stim	.set + .fdt	publication
[61]	rest, stim	Electro-Cap Mitsar	31	500	stim	.txt	publication
[62]	rest, stim	Active Two BioSemi system	64	512	stim	bids	publication
[63]	rest	n/d	19	250	n/d	.edf	publication
[64]	task: game	ActiveTwo MARK II Biosemi	128	2048	stim	.eeg	publication
[65]	rest	Elekta Neuromag Vectorview	60	n/d	n/d	.eeg	publication
[66]	sleep	n/d	3+	n/d	manual	.edf	yes
[67]	sleep	SomnoStar Pro	19	200	manual	.rec	publication
[68]	task: working memory	actiCHamp active electrode system	64	1k	manual	.eeg	publication
[69]	sleep	n/d	26+	256 (+)	manual	.edf	somewhat
[70]	sleep	GES 400	128	1000	manual	.txt	publication
[71]	sleep	SOMNOscreen	4	256	manual	.edf	yes
[72]	rest, task, sleep	BrainAmp	32	5000	manual	.eeg, .tsv	publication + file (dataset)
[73]	task: SoftPro protocol	Geodesic high-density EEG System	128	500	events	n/d	publication

* n/d: not disclosed, n/a: not applicable, stim: stimuli via STIM channel on EEG recorder, EP: evoked potential, AEP: auditory evoked potential.

can be used with no conditions. The second most common is "CC BY 4.0" which is similar to CC0 but requires the user to attribute credit to the original publisher.

B. HEALTH CONDITIONS PREVALENT IN THE DATASETS

Of the different conditions covered by the datasets, the most common is sleep-related problems. This was not unexpected, as it is standard clinical practice to record EEG associated with other sensors (e.g., PSG) for the diagnosis of sleep problems. Added to that, the use of EEG in research and clinical settings has originated from the first use of EEG in humans by Berger [3], [4] monitoring sleep.

Although some diagnoses are very prevalent in the general population, such as migraine, depression or anxiety, there are very few datasets found in this review that study that condition: 1 for migraine, 3 for depression, and none for anxiety. In the case of migraine, for example, the prevalence in the general population is estimated to be 1 billion people worldwide, i.e., 15% of the general population is estimated to have a migraine attack in a year [102]. However, there is only one relevant dataset available (from [100]) which, while very valuable in terms of amount of data, still does not cover the condition as a whole since only EEG recordings from migraine patients in the interictal phase (between the attacks, i.e., at least 3 days before or 3 days after a migraine attack) are available. No open EEG dataset that has recordings of patients in different migraine phases is openly available for use at the time of publication.

C. DATA AVAILABILITY

Various research studies published in recent years have a data availability statement, where the authors indicate that

TABLE 5. F-UJI FAIR evaluation of the datasets. *F* is the Findability category, *A* is the Accessibility category, *I* is Interoperability, and *R* is Reusability. The last column shows all of the criteria combined. The percentage values indicate the extent to which the dataset complies to the FAIRsFAIR metrics [43]. A score of 100% indicates that the dataset was identified by the automated tool to be compliant with all of the metrics for that category.

Ref.	F	A	I	R	FAIR
[44]	100%	67%	75%	40%	67%
[45]	100%	100%	75%	40%	71%
[46]	100%	67%	75%	40%	67%
[47]	86%	33%	25%	20%	42%
[48]	71%	33%	25%	30%	42%
[49]	93%	33%	75%	50%	65%
[50]	71%	33%	25%	30%	42%
[51]	86%	50%	75%	20%	52%
[52]	71%	33%	25%	30%	42%
[53]	79%	0%	25%	20%	35%
[54]	100%	67%	75%	50%	71%
[55]	71%	67%	100%	40%	63%
[56]	79%	0%	25%	40%	44%
[57]	71%	33%	25%	30%	42%
[58]	71%	33%	25%	30%	42%
[59]	71%	33%	25%	30%	42%
[60]	71%	33%	25%	30%	42%
[61]	86%	33%	75%	40%	58%
[62]	100%	67%	75%	50%	71%
[63]	100%	83%	75%	60%	77%
[64]	71%	33%	25%	30%	42%
[65]	71%	33%	25%	30%	42%
[66]	86%	67%	75%	30%	58%
[67]	36%	33%	75%	10%	31%
[68]	36%	33%	75%	10%	31%
[69]	86%	67%	100%	30%	63%
[70]	86%	50%	25%	30%	48%
[71]	86%	67%	75%	40%	63%
[72]	71%	33%	25%	30%	42%
[73]	100%	67%	75%	80%	83%

"data is available upon request". However, as Tedersoo et al. [103] have found, only approximately 40% of data requests, in general, are fulfilled. As the authors of [29] have discovered, availability of research data decreases with publication age.

The availability of data is not only important from a replication perspective, but it creates opportunities for further analysis with new methodologies and tools created in the future (i.e., developing deep learning approaches to diagnosis and treatment of diseases). The benefit of using AI models in medical applications is the ability of the model to deal with noisy data, uncertainties, and detect patterns in data that are not found using more standard statistical methods. The increase in available data also allows researchers to compare and validate results, and further strengthens the derived conclusions.

However, data sharing increases the burden on the researchers, as the dataset needs to be prepared to be reused. As Perrier et al. [21] identified, researchers may also be weary of sharing data due to possible misuse or misinterpretation of data. Indeed, one of the datasets in this review has a note by the authors asking potential users to correspond with them to make sure the data is being interpreted correctly. Also identified by [21] is the lack of incentives for researchers to share data as it does not have the same weight as published articles in regards to academic appointment and promotion.

One of the datasets in this review indicates that the associated code is available and provides a link, but as of the time of writing this review it was no longer available in the link provided (dataset [70], the associated paper [77] links to https://github.com/gsokoloff/Infant-Sleep-Study-I which does not exist). This could be solved by treating source code for analysis and models as datasets by hosting them in a repository and providing a DOI and metadata. For example, the F-UJI automated tool provides a DOI associated with a Zenodo record containing a version of the automated tool (DOI 10.5281/zenodo.4063720) [40].

D. DATA COLLECTION PROTOCOL

The Table 6 shows which datasets include the recording protocol in their description or files. Of the 30 datasets included in this review, 16 of them did not have a data collection protocol within their files or description. Most of them have the protocol explained on the associated publication, and not all of those had a direct citation or link to the publication — e.g., it was necessary to search the publication databases with author names and general keywords. In some cases, the publication was hard to find, or not found. The protocol details the assumptions, EEG recording task description and method, and other particularities of the methodology used.

In the case of sleep related datasets, there is a standardization of the recording protocol, e.g., the polysomnography study protocol is generally the same across datasets, therefore, the comparison of recordings from different datasets is possible. The same might not be the case for other tasks. Even for tasks observing the same phenomena (for example, visually evoked potentials) the recording protocols might be different, which makes it harder to compare different datasets and their results.

Also, without the protocol, data analysis is either very hard or impossible, as the assumptions made when collecting the data are not clear, and therefore the data analyst can "go fishing" for significant results (also called p-hacking) [104].

A number of datasets in this review, however, have included the code used to generate the stimuli used in the recordings. This enables researchers to use the same recording protocol, down to the same stimuli, which makes the comparison between different recordings more reliable.

E. DATASETS FILE TYPE AND STANDARDS

Although there is not an enforced standard on the type of file for EEG recordings, recent advancements have been made with the publishing of the BIDS standard by Gorgolewski and colleagues [81], and the further extension for EEG by Pernet and colleagues [82]. The BIDS aims to standardize the file format, organization, metadata, and distribution of neuroimaging data. This makes it easier for researchers to access and reuse data. Added to that, the BIDS standard allows analysis, software packages, and processes to be

TABLE 6. Dataset reusability: does it have a description? Is a README file included? Is there a document or text that indicates the file organization / structure of the dataset? Are the filenames descriptive? Where is the protocol located? Is the code used for recording and data analysis available?

Ref.	Description	README	File organization	Descriptive filename	Protocol	Recording code	Analysis code
[44]	×	×	×	✓	Publication	×	×
[45]	×	×	×	×	Dataset	n/a	×
[46]	×	×	×	×	Brief	n/a	×
[47]	✓	✓	✓	✓	Publication	×	✓
[48]	×	✓	×	✓	Publication	✓	✓
[49]	✓	✓	✓	✓	Dataset	×	✓
[50]	×	✓	×	✓	Publication	✓	✓
[51]	✓	✓	✓	✓	Publication	×	×
[52]	×	✓	×	✓	Publication	✓	✓
[53]	✓	×	×	✓	No	✓	×
[54]	✓	✓	×	✓	Dataset	✓	×
[55]	✓	✓	✓	✓	Dataset	✓	×
[56]	×	×	×	✓	Publication	×	×
[57]	Brief	✓	×	✓	Publication	×	×
[58]	×	✓	×	✓	Publication	✓	✓
[59]	×	✓	×	✓	Publication	✓	✓
[60]	×	✓	×	✓	Publication	✓	✓
[61]	×	×	×	✓	Dataset	n/a	×
[62]	✓	✓	×	✓	Publication	×	✓
[63]	×	×	×	✓	Somewhat	n/a	×
[64]	×	✓	×	✓	Publication	×	×
[65]	✓	✓	×	✓	No	n/a	×
[66]	✓	×	×	✓	Dataset	n/a	✓
[67]	Brief	×	✓ (website)	✓	Dataset	n/a	×
[68]	✓	✓	✓	✓	Dataset	n/a	✓
[69]	✓	n/d	n/d	✓	Dataset	n/a	✓, elsewhere
[70]	×	×	×	✓	Publication	n/a	×
[71]	✓	✓ (website)	✓ (website)	✓	Dataset	n/a	×
[72]	✓	✓	×	✓	Dataset	×	×
[73]	Publication	✓	✓ (README)	✓	Dataset	×	×

* n/d: not disclosed, n/a: not applicable.

interoperable with any dataset that follows the standard. This negates the need to develop analysis and software for each dataset, reducing possible errors as well.

For example, datasets that distribute the files in .mat format require specific software (Matlab) to open the files, while the BIDS files are an open standard that can be opened by any software framework or tool that can deal with EEG data. This allows a greater flexibility and accessibility to the data. Nine of the 30 datasets are published using BIDS due to the requirement enforced by the repository PhysioNet where they are hosted.

Other than the modalities listed in Section III, no other physiological data collection was found in the datasets included in this review. More physiological data would be beneficial in multiple studies; for example, in the case of Parkinson's Disease, an activity tracker (actigraphy) placed in the wrist could help measure the intensity of tremors [105].

F. FAIR GUIDING PRINCIPLES AND DATASET REUSABILITY

The FAIR assessment for the datasets indicates that they are easily findable, as the F score was high in general. However, this may be due to survivor bias, as the inclusion criteria for this review required a DOI. This then enforces that the datasets score at least 70% on *Findable*. The only exceptions to this high value are two datasets [67], [68] for which the DOI correspond to an article describing the dataset and not linked to where the dataset is hosted.

The *Accessible* criteria analyses the metadata related to level of access, and if the data can be accessed using standard networking protocols (such as HTTP, HTTPS, FTP, etc). Since most of the datasets are hosted in repositories which provide the infrastructure for data access, the A score is generally high.

The datasets hosted on data repositories such as Figshare, Mendley Data, and PhysioNet, have a high score on the *Interoperable* principle. The metrics in this principle measure how easy it is for machines to access and read the metadata of the dataset. Most of the data repositories provide by default the metadata since they are built for this purpose. The datasets that score low in this principle are those hosted on different places (other types of repositories, custom servers, etc) or for which the DOI refers to a published article.

Finally, the *Reusable* principle metrics assess if the dataset provides a license, if there is any information about data provenance, and if the dataset follows a metadata standard and provides the files in this standard. Most of the datasets in this review scored 50% or lower, which reflects the lack of information about data provenance and the fact that most of the datasets do not provide files in a standard format.

However, the FAIRsFAIR data object assessment metrics has limitations in the assessment of reusability, which is also highlighted by the metrics authors [43]. Specifically, the metric FsF-R1-01MD related to the Reusability principle requiring rich metadata can only verify if the repository of the dataset contains information described in general metadata

standards. For the case of observational clinical and/or behavioural data, the 'relevant attributes' should contain the data collection protocol, as that directly affects how the data is analysed and interpreted. However, there is no way of automatically or programmatically check this using the F-UJI tool. Added to that, the metric FsF-R1.3-01M, which verifies if the dataset follows the standard recommended by the target research community, does not correctly identify the BIDS format. Therefore, a low score on the *Reusability* category on Table 5 does not necessarily represent a true negative score on reusability for EEG datasets, as the tool cannot identify the community standard BIDS format. Dataset publishers should aim for total compliance with the FAIRsFAIR metrics, but the metrics and the automated tool also need to reflect the requirements and standards of the EEG research community.

The F-UJI tool is capable of checking repository-level information, i.e., the metadata available such as author names, identifiers, keywords, description, and machine-readable information (file names, formats, file structure, dates, and identifiers). However, it is unable to check if the dataset contains documentation such as file description, variable description and codification (if any), data collection protocol, and other information needed for analysis and reuse.

For the assessment of documentation level, we created a scale of 0 to 3 points to represent how well the context of the data provided is explained. We looked specifically for data collection protocols, descriptions of the dataset and its file contents, and software or code used for analysis.

The review by Roy et al. [19], which looks at published deep learning models and results, has identified that 53% of the articles used public data (i.e., data shared) and 42% use private data that is not shared, which means that these models are not reproducible. The review also highlights that 21% of the studies mention that more public data is needed to support the research of deep learning models of EEG data. The other problem raised by the review is that there is a lack of labelled clinical data, and labelling requires time and expertise. As for the source code for the models, Roy et al. found that only 13% of the studies included in their review provided the code, which means that just 12 out of 154 studies are reproducible (with code and data shared) [19].

Added to that, it is hard to combine multiple datasets together for a deep learning model due to the different recording protocols and electrode montages (electrode placement and references). As Yao et al. [106] and Hu et al. [107] the reference method (mono-polar or bipolar) creates systematic changes in the distribution of the signal frequency power.

G. SUMMARY AND FINAL CONSIDERATIONS

EEG is widely used in clinical and research settings. In recent years, EEG data has been used pervasively in Machine Learning algorithms developed to diagnose and further study different health conditions. However, DL models require more data for training and validation. To identify what open datasets have been published with EEG data related to healthcare, and to ascertain how reusable these datasets are, a scoping review was carried out.

A large number of potential datasets were found, but only 30 datasets of the 140 records retrieved and assessed actually fulfilled the inclusion criteria. The main reasons for rejection were: not pointing to a dataset, not related to healthcare or a diagnosis, or had no disclosure of ethics approval for the data collection. Of the 30 included datasets, 7 were related to sleep – which was expected as EEG is widely used in clinical settings during a polysomnographic assessment of sleep. The other diagnosis or conditions covered by the datasets found were either psychiatric or neurological related, such as Parkinson's Disease, Depression, Schizophrenia, Brain Injury, and so on. We found a lack of datasets for some prevalent conditions in the general population. For example, migraine is a neurological condition that afflicts 15% of the population, but we found only one dataset for migraine.

The majority of the datasets found were open access, while the others required either a credential (by registering into the dataset repository) or a request to the dataset publishers. In general, the datasets were licensed as entirely placed in the public domain, or as requiring only attribution to the dataset authors.

The biggest challenge found was that more than half of the datasets in this review did not have any description of its data collection protocol in the dataset. Of these, the majority had the protocol detailed in a different publication – generally a study published analysing the dataset – that in some cases were not linked to in the dataset page. In those cases, to find the protocol we had to search the dataset authors' publications and try to identify which one was based on that dataset. This makes it harder for other researchers to reuse the dataset. However, some datasets had not only the data collection protocol, but also the code used to process and analyse the data by the authors, which provides great documentation and resources for other researchers. By providing the data and the code, researchers make their results more easily reproduceable and verifiable, and also provide resources for the research community in general.

Added to this, the FAIR principles are important guidelines created by the research community to improve data sharing and reuse. They specify requirements for Findability (e.g., DOI registration, metadata, registered in an index), Accessibility (metadata has to be accessible and retrievable using standard protocols), Interoperability (metadata follows a certain standard, easily readable by machines and humans), and Reusability (data provenance, attributes, and other aspects relevant to the domain are clear and detailed). The FAIRsFAIR Project created an automated method of scoring datasets by the FAIR principles [42]. However, the Reusability criteria does not easily translate into scoring criteria that can be verified programmatically. The automated method does not assess if data collection protocols and data explanations are included in the datasets. We created a scale to measure reusability in terms of documentation level of the dataset.

In general, there is a need to: (i) share more data, as there are no datasets for certain health conditions; and (ii) better document the data shared, which makes it easier for other researchers to verify and use the data.

H. RECOMMENDATIONS FOR DATASET PUBLICATIONS

In line with the discussed above, and the reusability metrics created for the dataset assessment in Table 6, we recommend that authors, when publishing research data, follow the guidelines:

- Publish the data in data repositories that assigns a permanent identifier (DOI) to it. Some examples of free data repositories are: Zenodo (https://zenodo.org/) and Figshare (https://figshare.com/) for general data; and OpenNeuro (https://openneuro.org/) and PhysioNet (https://physionet.org/) for EEG and physiological data.
- Include a data availability statement with the link to data repository (or DOI) when publishing analyses of the dataset. Also include a link/reference to the published article in the dataset.
- Explicitly assign a license to the dataset —if the dataset is open to be reused by other researchers, good licenses are *CC0* (''No Rights Reserved'') or *CC-BY 4.0* (attribution required); see https://creativecommons.org/share-your-work/cclicenses/ for more options and information.
- Populate the metadata to make it easier to be discovered, by giving it a meaningful name, adding relevant keywords, and linking to and from the published article.
- Make sure the description clearly identifies the study objectives, the participants, devices and recording protocols used, and any other relevant information. This should be treated as the ''abstract'' for the dataset.
- Include a README file with:
  - The authors or persons responsible for collecting data, including email addresses and affiliations.
  - Dates of data collections.
  - Geographic information of where the data was collected.
  - Ethics approval information.
  - Licensing or other restrictions placed on the data.
  - Links to publications that cite or use the data.
  - Data collection protocol: which devices where used, how many electrodes (and specific montage used), what tasks were performed during the EEG recording. This should have enough information so that another researcher could replicate the data collection protocol.
  - Has any preprocessing been done to the data? If so, make clear what has been done.
  - Describe the file organization / structure of the dataset. Are the recordings separated into folders by participant or by task? Are raw and preprocessed data separated into folders? Where is located the participants data (such as age, gender, which group the participant belongs to, etc.)?
- If possible, include a separate file with the detailed data collection protocol. The code (or a link to its repository) used for stimuli generation, if any, should also be included.
- The filenames should be descriptive. Include the participant ID and the task done in the recording, and any other relevant information (e.g., ''HC1 Rest'' for the recording of Healthy Control ID 1 during rest).
- Make sure the dataset follows a standard format. In the case of EEG recordings, the standard format is BIDS [81], [82].
- Any code used for analysis should also be shared. It can be included with the dataset or hosted in another repository such as GitHub (https://github.com/).

V. CONCLUSION

EEG data holds immense potential for advancing healthcare through machine learning and deep learning. While our scoping review identified numerous datasets on this topic, it is apparent that there is room for improvement in terms of data sharing and documentation. Many datasets lacked essential data collection protocols and explanations, hindering their reusability and the replication of results. As we move forward, the research community must prioritize sharing data for health conditions currently underrepresented and focus on comprehensive documentation by adhering to the FAIR principles with the ultimate goal to enhance dataset transparency.

REFERENCES

[1] S. Sanei and J. A. Chambers, *EEG Signal Processing*. Hoboken, NJ, USA: Wiley, May 2013.
[2] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice* (Issues in Clinical and Cognitive Neuropsychology). Cambridge, MA, USA: MIT Press, 2014.
[3] H. Berger, ''Über das elektroenkephalogramm des menschen,'' *Archiv für Psychiatrie Und Nervenkrankheiten*, vol. 87, no. 1, pp. 527–570, 1929.
[4] R. Ince, S. S. Adanir, and F. Sevmez, ''The inventor of electroencephalography (EEG): Hans Berger (1873–1941),'' *Child's Nervous Syst.*, vol. 37, no. 9, pp. 2723–2724, Sep. 2021, doi: 10.1007/s00381-020-04564-z.
[5] U. R. Acharya, S. Vinitha Sree, G. Swapna, R. J. Martis, and J. S. Suri, ''Automated EEG analysis of epilepsy: A review,'' *Knowl.-Based Syst.*, vol. 45, pp. 147–165, Jun. 2013, doi: 10.1016/j.knosys.2013.02.014.
[6] J. Jeong, ''EEG dynamics in patients with Alzheimer's disease,'' *Clin. Neurophysiol.*, vol. 115, no. 7, pp. 1490–1505, Jul. 2004, doi: 10.1016/j.clinph.2004.01.001.
[7] R. Cassani, M. Estarellas, R. San-Martin, F. J. Fraga, and T. H. Falk, ''Systematic review on resting-state EEG for Alzheimer's disease diagnosis and progression assessment,'' *Disease Markers*, vol. 2018, pp. 1–26, Oct. 2018, doi: 10.1155/2018/5174815.
[8] K. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, ''Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation,'' *Entropy*, vol. 18, no. 9, p. 272, Aug. 2016, doi: 10.3390/e18090272.
[9] M. A. Carskadon and W. C. Dement, ''Chapter 2—Normal human sleep: An overview,'' in *Principles and Practice of Sleep Medicine*, 5th ed., M. H. Kryger, T. Roth, and W. C. Dement, Eds. St. Louis, MO, USA: Elsevier, 2011, pp. 16–26.
[10] A. L. Loomis, E. N. Harvey, and G. Hobart, ''Potential rhythms of the cerebral cortex during sleep,'' *Science*, vol. 81, no. 2111, pp. 597–598, Jun. 1935, doi: 10.1126/science.81.2111.597.
[11] X.-S. Zhang, R. J. Roy, and E. W. Jensen, ''EEG complexity as a measure of depth of anesthesia for patients,'' *IEEE Trans. Biomed. Eng.*, vol. 48, no. 12, pp. 1424–1433, Dec. 2001, doi: 10.1109/10.966601.

[12] J. J. Newson and T. C. Thiagarajan, "EEG frequency bands in psychiatric disorders: A review of resting state studies," *Frontiers Hum. Neurosci.*, vol. 12, p. 521, Jan. 2019.

[13] B. Yakubov, S. Das, R. Zomorrodi, D. M. Blumberger, P. G. Enticott, M. Kirkovski, T. K. Rajji, and P. Desarkar, "Cross-frequency coupling in psychiatric disorders: A systematic review," *Neurosci. Biobehavioral Rev.*, vol. 138, Jul. 2022, Art. no. 104690, doi: 10.1016/j.neubiorev.2022.104690.

[14] S. Das, R. Zomorrodi, M. Mirjalili, M. Kirkovski, D. M. Blumberger, T. K. Rajji, and P. Desarkar, "Machine learning approaches for electroencephalography and magnetoencephalography analyses in autism spectrum disorder: A systematic review," *Prog. Neuro-Psychopharmacology Biol. Psychiatry*, vol. 123, Apr. 2023, Art. no. 110705, doi: 10.1016/j.pnpbp.2022.110705.

[15] A. Kawala-Sterniuk, N. Browarska, A. Al-Bakri, M. Pelc, J. Zygarlicki, M. Sidikova, R. Martinek, and E. J. Gorzelanczyk, "Summary of over fifty years with brain-computer interfaces—A review," *Brain Sci.*, vol. 11, no. 1, p. 43, Jan. 2021, doi: 10.3390/brainsci11010043.

[16] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo, "EEG-based BCI emotion recognition: A survey," *Sensors*, vol. 20, no. 18, p. 5083, Sep. 2020, doi: 10.3390/s20185083.

[17] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Jan. 2019, Art. no. 011001, doi: 10.1088/1741-2552/aaf12e.

[18] D. Painuli, S. Bhardwaj, and U. Köse, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105580, doi: 10.1016/j.compbiomed.2022.105580.

[19] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *J. Neural Eng.*, vol. 16, no. 5, Aug. 2019, Art. no. 051001, doi: 10.1088/1741-2552/ab260c.

[20] *Data Sharing in the Age of Deep Learning | Nature Biotechnology*. Accessed: Nov. 2, 2023. [Online]. Available: https://www.nature.com/articles/s41587-023-01770-3

[21] L. Perrier, E. Blondal, and H. MacDonald, "The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis," *PLoS ONE*, vol. 15, no. 2, Feb. 2020, Art. no. e0229182, doi: 10.1371/journal.pone.0229182.

[22] C. Tenopir, N. M. Rice, S. Allard, L. Baird, J. Borycz, L. Christian, B. Grant, R. Olendorf, and R. J. Sandusky, "Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0229003, doi: 10.1371/journal.pone.0229003.

[23] E. Gibney, "Could machine learning fuel a reproducibility crisis in science?" *Nature*, vol. 608, no. 7922, pp. 250–251, Jul. 2022, doi: 10.1038/d41586-022-02035-w.

[24] M. R. Munafó, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. D. Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis, "A manifesto for reproducible science," *Nature Hum. Behaviour*, vol. 1, no. 1, pp. 1–9, Jan. 2017, doi: 10.1038/s41562-016-0021.

[25] E. Sohn, "The reproducibility issues that haunt health-care AI," *Nature*, vol. 613, no. 7943, pp. 402–403, Jan. 2023, doi: 10.1038/d41586-023-00023-2.

[26] A. L. Beam, A. K. Manrai, and M. Ghassemi, "Challenges to the reproducibility of machine learning models in health care," *JAMA*, vol. 323, no. 4, pp. 305–306, Jan. 2020, doi: 10.1001/jama.2019.20866.

[27] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, and C. S. Greene, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. 14–16, Oct. 2020, doi: 10.1038/s41586-020-2766-y.

[28] C. Summers and M. J. Dinneen, "Nondeterminism and instability in neural network optimization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 9913–9922.

[29] T. H. Vines, A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison, "The availability of research data declines rapidly with article age," *Current Biol.*, vol. 24, no. 1, pp. 94–97, Jan. 2014, doi: 10.1016/j.cub.2013.11.014.

[30] T. Miyakawa, "No raw data, no science: Another possible source of the reproducibility crisis," *Mol. Brain*, vol. 13, no. 1, p. 24, Feb. 2020, doi: 10.1186/s13041-020-0552-2.

[31] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. D. S. Santos, P. E. Bourne, and J. Bouwman, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, Mar. 2016, Art. no. 160018, doi: 10.1038/sdata.2016.18.

[32] P. Handa, M. Mathur, and N. Goel, "Open and free EEG datasets for epilepsy diagnosis," Aug. 2021, *arXiv:2108.01030*.

[33] I. Ahmad, X. Wang, M. Zhu, C. Wang, Y. Pi, J. A. Khan, S. Khan, O. W. Samuel, S. Chen, and G. Li, "EEG-based epileptic seizure detection via machine/deep learning approaches: A systematic review," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–20, Jun. 2022, doi: 10.1155/2022/6486570.

[34] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers Neurosci.*, vol. 10, May 2016, Art. no. 195498.

[35] J. H. Y. Chen and R. M. Mehmood, "A critical review on state-of-the-art EEG-based emotion datasets," in *Proc. 1st Int. Conf. Adv. Inf. Sci. Syst.* New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 1–5, doi: 10.1145/3373477.3373707.

[36] A. Al-Saegh, S. A. Dawwd, and J. M. Abdul-Jabbar, "Deep learning for motor imagery EEG-based classification: A review," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102172, doi: 10.1016/j.bspc.2020.102172.

[37] S. Yasin, S. A. Hussain, S. Aslan, I. Raza, M. Muzammel, and A. Othmani, "EEG based major depressive disorder and bipolar disorder detection using neural networks: A review," *Comput. Methods Programs Biomed.*, vol. 202, Apr. 2021, Art. no. 106007, doi: 10.1016/j.cmpb.2021.106007.

[38] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, Apr. 2019, Art. no. 031001, doi: 10.1088/1741-2552/ab0ab5.

[39] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, D. M. D. Peters, T. Horsley, L. Weeks, and S. Hempel, "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," *Ann. Internal Med.*, vol. 169, no. 7, pp. 467–473, Oct. 2018, doi: 10.7326/m18-0850.

[40] A. Devaraju and R. Huber, "F-UJI—An automated FAIR data assessment tool," Zenodo, Tech. Rep., Oct. 2020, doi: 10.5281/ZENODO.4063720.

[41] (Feb. 2019). *The Project*. [Online]. Available: https://www.fairsfair.eu/the-project

[42] A. Devaraju and R. Huber, "An automated solution for measuring the progress toward FAIR research data," *Patterns*, vol. 2, no. 11, Nov. 2021, Art. no. 100370, doi: 10.1016/j.patter.2021.100370.

[43] A. Devaraju, R. Huber, M. Mokrane, P. Herterich, L. Cepinskas, J. de Vries, H. L'Hours, J. Davidson, and A. White, "FAIRsFAIR data object assessment metrics," FAIRsFAIR, Zenodo, Tech. Rep., Apr. 2022, doi: 10.5281/zenodo.6461229.

[44] C. Breitling, J. Tegelbeckers, H.-H. Flechtner, and K. Krauel, "EEG raw data—Economical assessment of working memory and response inhibition in ADHD," Tech. Rep., 2020, doi: 10.6084/M9.FIGSHARE.12115773.V1.

[45] I. Lazarou, K. Georgiadis, S. Nikolopoulos, V. Oikonomou, and I. Kompatsiaris, "Resting-state high-density EEG using EGI GES 300 with 256 channels of healthy elders, people with subjective and mild cognitive impairment and Alzheimer's disease," Tech. Rep., Dec. 2020, doi: 10.5281/ZENODO.4316608.

[46] E. Milne, "EEG data for 'electrophysiological signatures of brain aging in autism spectrum disorder,'" Tech. Rep., 2021, doi: 10.15131/SHEF.DATA.16840351.

[47] E. L. Johnson, "64-channel human scalp EEG from 14 unilateral PFC patients and 20 healthy controls performing a lateralized visuospatial working memory task," Tech. Rep., 2017, doi: 10.6080/K0ZC811B.

[48] J. F. Cavanagh and D. Quinn, "EEG: Three-stim auditory oddball and rest in acute and chronic TBI," Tech. Rep., 2021, doi: 10.18112/OPENNEURO.DS003522.V1.1.0.

[49] A. Van Laarhoven, S. Millard, K. Bokelmann, R. Schalbroeck, N. Van Der Wee, and N. Van Loey, "No indications for altered EEG oscillatory activity in patients with chronic post-burn itch compared to healthy controls," Tech. Rep., 2022, doi: 10.34894/EOZLXH.

[50] *EEG: Depression Rest*, 2021, doi: 10.18112/OPENNEURO.DS003478.V1.
[51] B. Hu, "Multi-modal open dataset for mental-disorder analysis," Tech. Rep., 2016," 2022, doi: 10.5255/UKDA-SN-854301.
[52] *EEG: Probabilistic Selection and Depression*," 2021, doi: 10.18112/OPENNEURO.DS003474.V1.1.0.
[53] A. Presacco, J. Simon, and S. Anderson, "EEG-MEG," Tech. Rep., 2018, doi: 10.13016/M2599Z52H.
[54] A. C. Zar, S. Haigh, P. Grover, and M. Behrmann, "Ultra high-density EEG recording of interictal migraine and controls: Sensory and rest," Tech. Rep., 2020, doi: 10.1184/R1/12636731.
[55] L. M. Alexander, J. Escalera, L. Ai, C. Andreotti, K. Febre, A. Mangone, N. Vega-Potler, N. Langer, A. Alexander, M. Kovacs, and S. Litke, "An open resource for transdiagnostic research in pediatric mental health and learning disorders," *Sci. Data*, vol. 4, no. 1, Dec. 2017, Art. no. 170181, doi: 10.1038/sdata.2017.181.
[56] F. Karimi, "FOG severity EEG/EMG," Tech. Rep., Oct. 2021, doi: 10.21227/5CH0-8B29.
[57] A. P. Rockhill, N. Jackson, J. George, A. Aron, and N. C. Swann, "UC san Diego resting state EEG data from patients with Parkinson's disease," Tech. Rep., 2022, doi: 10.18112/OPENNEURO.DS002778.V1.0.5.
[58] J. F. Cavanagh, A. Singh, and K. Narayanan, "EEG: Simon Conflict in Parkinson's," Tech. Rep., 2021, doi: 10.18112/OPENNEURO.DS003509.V1.1.0.
[59] J. F. Cavanagh and D. Brown, "EEG: Reinforcement learning in Parkinson's," Tech. Rep., 2021, doi: 10.18112/OPENNEURO.DS003506.V1.1.0.
[60] *EEG: 3-Stim Auditory Oddball and Rest in Parkinson's*," 2021, doi: 10.18112/OPENNEURO.DS003490.V1.1.0.
[61] R. Eyvazpour, "Neural activity in self-identified claustrophobic individuals under in-vivo stimuli: A human electroencephalography dataset," *Data Brief*, vol. 40, Feb. 2021, Art. no. 107733, doi: 10.17632/HRGFWZS3SX.2.
[62] S. Wiens and R. Eklund, "Open data: Electrophysiological correlates of in-vivo and virtual reality therapy in spider phobia," *Psychophysiology*, vol. 59, no. 12, 2022, Art. no. e14117, doi: 10.17045/STHLMUNI.14891607.
[63] E. Olejarczyk and W. Jernajczyk, "EEG in schizophrenia," Tech. Rep., 2017, doi: 10.18150/REPOD.0107441.
[64] A. Padee, P. Missonnier, A. Prevot, G. Favre, I. Gothuey, M. Merlo, and J. Richiardi, "Fribourg ultimatum game in schizophrenia study," Tech. Rep., 2022, doi: 10.18112/OPENNEURO.DS004000.V1.0.0.
[65] D. Salisbury, D. Seebold, and B. Coffman, "EEG: First episode psychosis vs. control resting task 1," Tech. Rep., 2022, doi: 10.18112/OPENNEURO.DS003944.V1.0.1.
[66] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters, "CAP sleep database," Tech. Rep., 2001, doi: 10.13026/C2VC79.
[67] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, Feb. 2016, doi: 10.1016/j.cmpb.2015.10.013.
[68] N. Mei, M. D. Grossberg, K. Ng, K. T. Navarro, and T. M. Ellmore, "A high-density scalp EEG dataset acquired during brief naps after a visual working memory task," *Data Brief*, vol. 18, pp. 1513–1519, Jun. 2018, doi: 10.1016/j.dib.2018.04.073.
[69] H. Lee, B. Li, Y. Huang, Y. Chi, and S. Lin, "NCH sleep DataBank: A large collection of real-world pediatric sleep studies with longitudinal clinical data," *Sci. Data*, vol. 9, no. 1, p. 421, 2022, doi: 10.13026/P2RP-SG37.
[70] M. Blumberg and G. Sokoloff, "Sleep, myoclonic twitching, and EEG activity in human infants," Tech. Rep., 2021, doi: 10.17910/B7.1281.
[71] A.-E. Diego and R. Rijsman, "Haaglanden medisch centrum sleep staging database," Tech. Rep., 2021, doi: 10.13026/T79Q-FR32.
[72] Y. Gu, F. Han, L. E. Sainburg, M. M. Schade, and X. Liu, "Simultaneous EEG and fMRI signals during sleep from humans," Tech. Rep., 2022, doi: 10.18112/OPENNEURO.DS003768.V1.0.9.
[73] G. Averta, F. Barontini, V. Catrambone, S. Haddadin, G. Handjaras, J. P. O. Held, T. Hu, E. Jakubowitz, C. Kanzler, J. Kuhn, O. Lambercy, A. Leo, A. Obermeier, E. Ricciardi, A. Schwarz, G. Valenza, A. Bicchi, and M. Bianchi, "U-limb," Tech. Rep., 2020, doi: 10.7910/DVN/FU3QZ9.
[74] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep," *Sleep Med.*, vol. 2, no. 6, pp. 537–553, Nov. 2001, doi: 10.1016/s1389-9457(01)00149-6.
[75] B. Li, "Sleep study data analysis," Tech. Rep., Mar. 2023. [Online]. Available: https://github.com/liboyue/sleep_study
[76] H. Lee, B. Li, S. DeForte, M. L. Splaingard, Y. Huang, Y. Chi, and S. L. Linwood, "A large collection of real-world pediatric sleep studies," *Sci. Data*, vol. 9, no. 1, p. 421, Jul. 2022, doi: 10.1038/s41597-022-01545-6.
[77] G. Sokoloff, J. C. Dooley, R. M. Glanz, R. Y. Wen, M. M. Hickerson, L. G. Evans, H. M. Laughlin, K. S. Apfelbaum, and M. S. Blumberg, "Twitches emerge postnatally during quiet sleep in human infants and are synchronized with sleep spindles," *Current Biol.*, vol. 31, no. 15, pp. 3426–3432, Aug. 2021, doi: 10.1016/j.cub.2021.05.038.
[78] D. Alvarez-Estevez and R. M. Rijsman, "Inter-database validation of a deep learning approach for automatic sleep scoring," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0256111, doi: 10.1371/journal.pone.0256111.
[79] Y. Gu, F. Han, L. E. Sainburg, M. M. Schade, O. M. Buxton, J. H. Duyn, and X. Liu, "An orderly sequence of autonomic and neural events at transient arousal changes," *NeuroImage*, vol. 264, Dec. 2022, Art. no. 119720, doi: 10.1016/j.neuroimage.2022.119720.
[80] J. F. Cavanagh, A. W. Bismark, M. J. Frank, and J. J. B. Allen, "Multiple dissociations between comorbid depression and anxiety on reward and punishment processing: Evidence from computationally informed EEG," *Comput. Psychiatry*, vol. 3, pp. 1–17, Jan. 2019, doi: 10.1162/cpsy_a_00024.
[81] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, and D. A. Handwerker, "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Sci. Data*, vol. 3, no. 1, Jun. 2016, Art. no. 160044, doi: 10.1038/sdata.2016.44.
[82] C. R. Pernet, S. Appelhoff, K. J. Gorgolewski, G. Flandin, C. Phillips, A. Delorme, and R. Oostenveld, "EEG-BIDS, an extension to the brain imaging data structure for electroencephalography," *Sci. Data*, vol. 6, no. 1, p. 103, Jun. 2019, doi: 10.1038/s41597-019-0104-8.
[83] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, and Q. Zhao, "A multi-modal open dataset for mental-disorder analysis," *Sci. Data*, vol. 9, no. 1, p. 178, Apr. 2022, doi: 10.1038/s41597-022-01211-x.
[84] C. Breitling-Ziegler, J. Tegelbeckers, H.-H. Flechtner, and K. Krauel, "Economical assessment of working memory and response inhibition in ADHD using a combined n-back/nogo paradigm: An ERP study," *Frontiers Hum. Neurosci.*, vol. 14, p. 322, Aug. 2020.
[85] E. Olejarczyk and W. Jernajczyk, "Graph-based analysis of brain connectivity in schizophrenia," *PLoS ONE*, vol. 12, no. 11, Nov. 2017, Art. no. e0188629, doi: 10.1371/journal.pone.0188629.
[86] A. Padée, P. Missonnier, A. Prévot, G. Favre, I. Gothuey, and J. Richiardi, "Anticipatory deficits in psychosis during social cognition task analysed by task-based EEG dynamic functional connectivity," *Schizophrenia Res.*, vol. 254, pp. 11–13, Apr. 2023, doi: 10.1016/j.schres.2023.01.029.
[87] D. Salisbury, D. Seebold, and B. Coffman, "EEG: First episode psychosis vs. control resting task 2," Tech. Rep., 2022, doi: 10.18112/OPENNEURO.DS003947.V1.0.1.
[88] H. Phalen, B. A. Coffman, A. Ghuman, E. Sejdić, and D. F. Salisbury, "Non-negative matrix factorization reveals resting-state cortical alpha network abnormalities in the first-episode schizophrenia spectrum," *Biol. Psychiatry, Cognit. Neurosci. Neuroimaging*, vol. 5, no. 10, pp. 961–970, Oct. 2020, doi: 10.1016/j.bpsc.2019.06.010.
[89] D. Moradi, F. Rahimi, R. Eyvazpour, A. Jahan, S. H. Rasta, and M. Esmaeili, "Neural activity in selfidentified claustrophobic individuals under in-vivo stimuli: A human electroencephalography dataset," *Data Brief.*, vol. 40, pp. 107–733, Feb. 2022, doi: 10.1016/j.dib.2021.107733.
[90] S. Wiens, R. Eklund, M. Szychowska, A. Miloff, D. Cosme, S. Pierzchajlo, and P. Carlbring, "Electrophysiological correlates of in vivo and virtual reality exposure therapy in spider phobia," *Psychophysiology*, vol. 59, no. 12, Dec. 2022, Art. no. e14117, doi: 10.1111/psyp.14117.
[91] F. Karimi, J. Niu, K. Gouweleeuw, Q. Almeida, and N. Jiang, "Movement-related EEG signatures associated with freezing of gait in Parkinson's disease: An integrative analysis," *Brain Commun.*, vol. 3, no. 4, Oct. 2021, Art. no. fcab277, doi: 10.1093/braincomms/fcab277.

[92] J. S. George, J. Strunk, R. Mak-McCully, M. Houser, H. Poizner, and A. R. Aron, "Dopaminergic therapy in Parkinson's disease decreases cortical beta band coherence in the resting state and increases cortical beta band power during executive control," *NeuroImage, Clin.*, vol. 3, pp. 261–270, Jan. 2013, doi: 10.1016/j.nicl.2013.07.013.

[93] A. Singh, S. P. Richardson, N. Narayanan, and J. F. Cavanagh, "Mid-frontal theta activity is diminished during cognitive control in Parkinson's disease," *Neuropsychologia*, vol. 117, pp. 113–122, Aug. 2018, doi: 10.1016/j.neuropsychologia.2018.05.020.

[94] D. R. Brown, S. P. Richardson, and J. F. Cavanagh, "An EEG marker of reward processing is diminished in Parkinson's disease," *Brain Res.*, vol. 1727, Jan. 2020, Art. no. 146541, doi: 10.1016/j.brainres.2019.146541.

[95] J. F. Cavanagh, P. Kumar, A. A. Mueller, S. P. Richardson, and A. Mueen, "Diminished EEG habituation to novel events effectively classifies Parkinson's patients," *Clin. Neurophysiol.*, vol. 129, no. 2, pp. 409–418, Feb. 2018, doi: 10.1016/j.clinph.2017.11.023.

[96] I. Lazarou, K. Georgiadis, S. Nikolopoulos, V. P. Oikonomou, A. Tsolaki, I. Kompatsiaris, M. Tsolaki, and D. Kugiumtzis, "A novel connectome-based electrophysiological study of subjective cognitive decline related to Alzheimer's disease by using resting-state high-density EEG EGI GES 300," *Brain Sci.*, vol. 10, no. 6, p. 392, Jun. 2020, doi: 10.3390/brainsci10060392.

[97] E. L. Johnson, C. D. Dewar, A.-K. Solbakk, T. Endestad, T. R. Meling, and R. T. Knight, "Bidirectional frontoparietal oscillatory systems support working memory," *Current Biol.*, vol. 27, no. 12, pp. 1829–1835.e4, Jun. 2017, doi: 10.1016/j.cub.2017.05.046.

[98] J. F. Cavanagh, J. K. Wilson, R. E. Rieger, D. Gill, J. M. Broadway, J. H. Story Remer, V. Fratzke, A. R. Mayer, and D. K. Quinn, "ERPs predict symptomatic distress and recovery in sub-acute mild traumatic brain injury," *Neuropsychologia*, vol. 132, Sep. 2019, Art. no. 107125, doi: 10.1016/j.neuropsychologia.2019.107125.

[99] G. Averta, F. Barontini, V. Catrambone, S. Haddadin, G. Handjaras, J. P. O. Held, T. Hu, E. Jakubowitz, C. M. Kanzler, J. Kühn, O. Lambercy, A. Leo, A. Obermeier, E. Ricciardi, A. Schwarz, G. Valenza, A. Bicchi, and M. Bianchi, "U-limb: A multi-modal, multi-center database on arm motion control in healthy and post-stroke conditions," *GigaScience*, vol. 10, no. 6, Jun. 2021, Art. no. giab043, doi: 10.1093/gigascience/giab043.

[100] A. Chamanzar, S. M. Haigh, P. Grover, and M. Behrmann, "Abnormalities in cortical pattern of coherence in migraine detected using ultra high-density EEG," *Brain Commun.*, vol. 3, no. 2, Apr. 2021, Art. no. fcab061, doi: 10.1093/braincomms/fcab061.

[101] A. Presacco, J. Z. Simon, and S. Anderson, "Evidence of degraded representation of speech in noise, in the aging midbrain and cortex," *J. Neurophysiol.*, vol. 116, no. 5, pp. 2346–2355, Nov. 2016, doi: 10.1152/jn.00372.2016.

[102] T. J. Steiner, L. J. Stovner, R. Jensen, D. Uluduz, and Z. Katsarava, "Migraine remains second among the world's causes of disability, and first among young women: Findings from GBD2019," *J. Headache Pain*, vol. 21, no. 1, p. 137, Dec. 2020, doi: 10.1186/s10194-020-01208-0.

[103] L. Tedersoo, R. Küngas, E. Oras, K. Köster, H. Eenmaa, Ä. Leijen, M. Pedaste, M. Raju, A. Astapova, H. Lukner, K. Kogermann, and T. Sepp, "Data sharing practices and data availability upon request differ across scientific disciplines," *Sci. Data*, vol. 8, no. 1, p. 192, Jul. 2021, doi: 10.1038/s41597-021-00981-0.

[104] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of P-hacking in science," *PLOS Biol.*, vol. 13, no. 3, Mar. 2015, Art. no. e1002106, doi: 10.1371/journal.pbio.1002106.

[105] M. Sica, S. Tedesco, C. Crowe, L. Kenny, K. Moore, S. Timmons, J. Barton, B. O'Flynn, and D.-S. Komaris, "Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review," *PLoS ONE*, vol. 16, no. 2, Feb. 2021, Art. no. e0246528, doi: 10.1371/journal.pone.0246528.

[106] D. Yao, L. Wang, R. Oostenveld, K. D. Nielsen, L. Arendt-Nielsen, and A. C. N. Chen, "A comparative study of different references for EEG spectral mapping: The issue of the neutral reference and the use of the infinity reference," *Physiological Meas.*, vol. 26, no. 3, pp. 173–184, Feb. 2005, doi: 10.1088/0967-3334/26/3/003.

[107] S. Hu, Y. Lai, P. A. Valdes-Sosa, M. L. Bringas-Vega, and D. Yao, "How do reference montage and electrodes setup affect the measured scalp EEG potentials?" *J. Neural Eng.*, vol. 15, no. 2, Jan. 2018, Art. no. 026013, doi: 10.1088/1741-2552/aaa13f.



CAROLINE PERES DA SILVA (Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from Universidade Estadual de Londrina, Brazil, and the M.Eng.Sc. degree in electrical and electronics engineering from University College Cork (UCC), Cork, Ireland. She is currently pursuing the Ph.D. degree with Tyndall National Institute, Cork, with funding from the Science Foundation Ireland's Centre for Research Training in Artificial Intelligence.

Her research interests include wearable sensor devices for healthcare and well-being, data analysis, signal processing, and artificial intelligence applied to healthcare.



SALVATORE TEDESCO (Member, IEEE) received the B.Sc. degree (Hons.) in information technology engineering and the M.Sc. degree (Hons.) in telecommunications engineering from the University of Salento, Lecce, Italy, and the Ph.D. degree in electrical and electronic engineering from University College Cork (UCC), Cork, Ireland.

Since April 2012, he has been with the Wireless Sensor Networks Group, Tyndall National Institute, UCC, where he is currently a Senior Researcher and a Team Leader, responsible for leading a research team in the wearable and data analytics area. He has authored more than 60 papers in international journals and conference proceedings, and has received over 1.1 million Euro in grant funding, as a PI and a Co-PI. Since joining Tyndall, he has managed and successfully led over 20 industrial and research-oriented projects focused on his main research interests on wearable technologies for healthcare and well-being, human motion analysis in sports and clinical populations, digital health, physiological monitoring, signal processing, edge analytics, and machine learning. Further contributions deal with RFID technology and antenna design, ultra-wideband localization systems for indoor applications, sensor calibration, and industry 4.0.

Dr. Tedesco is the Founder and the past Chair of the Early-Career Researchers' Network, Tyndall National Institute (TEC-Net) (2020–2022). He is currently the Vice-President of the Young European Associated Researchers (YEAR) Network Association.



BRENDAN O'FLYNN (Senior Member, IEEE) received the B.Eng. (Hons.), M.Eng.Sc., and Ph.D. degrees from University College Cork, Cork, Ireland. He is currently a Senior Staff Researcher with Tyndall National Institute, Cork. He is also the Head of the Wireless Sensor Networks (WSN) Group. His research interests include wearable sensing systems, edge-based analytics, sensor system integration, low power embedded systems design and development, system miniaturization, and RF system design and optimization. He has published widely in these areas and has secured significant funding for the development and deployment of "smart sensing" technologies and applied research projects.