IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

A systematic review of electroencephalography open datasets and their usage with deep learning models

Alberto Nogales[1] and Álvaro J. García-Tejedor[1]

[1]CEIEC, Research Institute, Universidad Francisco de Vitoria, Ctra. M-515 Pozuelo-Majadahonda km 1800, 28223 Pozuelo de Alarcón, Spain

Corresponding author: Alberto Nogales (e-mail: alberto.nogales@ceiec.es).
"No funding received."

ABSTRACT

Data are the main headache for machine learning, both because of their varied nature and their limited availability. The medical field brings together both situations: tables, images, text, or signals that are difficult to acquire due to the number of patients, the complexity and time of acquisition, or ethical constraints. The existence of open datasets is the best option for researchers in this field. Electroencephalograms are a good example of this situation. This paper identifies the primary open datasets of electroencephalography tests and how they are used in deep learning models. The aim is to provide structured information that can be consulted by researchers in the field (both physicians and computer scientists) in order to know which datasets are available, which characteristics they have, or which deep learning models could be applied to them. The process followed the PRISMA methodology for systematic reviews applying different inclusion and exclusion criteria to obtain a set of high-quality papers on which the data sets used were analyzed. The databases included in the searches were Scopus, PubMed, Web of Science (WOS), Science Direct, IEEE Explorer, and SpringerLink. In total, 37 papers were selected which included 30 datasets that have been considered. Then, the DL models used in the papers and the different characteristics of the datasets have been statistically analyzed by obtaining different measures and graphs. The most relevant conclusions are the widespread use of convolutional neural networks (the less innovative among the different models) as the main tool for EEG data analysis. Against this position, we found the use of hybrid models and the family of RNNs as techniques to use in cases of brain stimuli, classification of levels of fatigue, and diagnosis of diseases. Related to the datasets' features, we demonstrate the difficulty in compiling this data due to the number of tests and that should be studied the minimum of channels or sampling frequency recommended to obtain good accuracies in the model.

INDEX TERMS Systematic review, Deep learning, Open datasets, Electroencephalograms.

I. INTRODUCTION

Most people are connected every day through their mobile phones or computers. This entails the creation of vast amounts of data through organizations or private companies every day. According to (Völske et al. 2021), in 2020, 44 zettabytes were produced, and by 2025 is estimated to be between 163 and 175 zettabytes. The trend remains the same in the medical field due to new applications and the wide range of data from demographic information to images resulting in medical tests like radiographs or 3D scanners passing through those that collect the biomedical signal.

(Chang and Moura 2010) define biosignal processing as extracting relevant information from biomedical signals. These are also described as physiological activities from organisms that can comprise neural, cardiac rhythms, or others. Among all the medical tests related to signals, electroencephalograms (EEGs) are considered the most beneficial for compiling brain signals.

EEGs are a type of data called time series, defined in (Velicer and Molenaar 2013) as sets of repeated observations of a single unit or individual at regular intervals over many instances. The case of EEGs corresponds to a test used to diagnose neurological diseases based on a set of electrodes placed around the scalp. EEGs compile a lot of data being very complex to analyze. They need professionals with high skills acquired through years of training. The problem with EEGs is that they are studied by eye, and due to their complexity, the professionals miss a lot of information.

There is a trend toward integrating and leveraging these enormous amounts of data to make medicine more personalized, efficient and focused on the patient. Nevertheless, classical methods like statistics are not powerful enough to manage that large number of variables and data. At this point, using more modern techniques, such as Artificial Intelligence (AI), is a great benefit. They allow us to find new patterns and predict how different variables behave or identify new ones not considered in complex medical problems, (Normandeau 2013)

AI is a computer science field aiming to analyze and decipher human mechanisms related to intelligent behaviours that are, later, reproduced in machines, (Russell and Norvig 2016). Among all the AI techniques, Machine Learning (ML) has stood out from the rest in recent years. ML is defined by (Samuel 1959) as a discipline that studies and develops algorithms that create systems that learn by finding patterns in datasets. ML comprises a wide range of techniques, with Artificial Neural Networks (ANN) obtaining the best results recently. (Hecht-Nielsen 1988) defines ANN as a computational model formed by several simple units that are strongly connected and can process information by responding to external stimuli. The benefits of ANN remained not very useful until deeper architectures, called Deep Learning (DL) models, arose. DL consists of ANN models with several layers that can learn data representation using more abstraction levels, (LeCun, Bengio, and Hinton 2015). Figure 1 depicts the hierarchy of the fields in AI described previously.

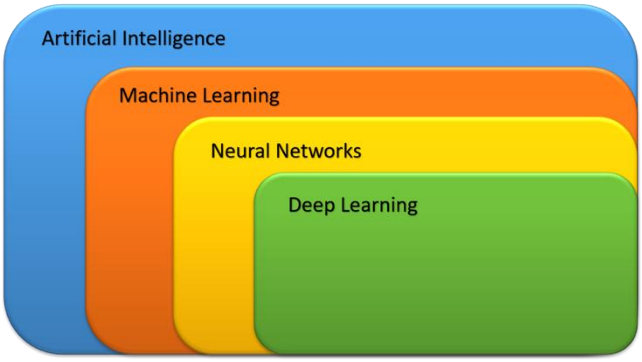

FIGURE 1. The hierarchical situation of the deep learning models among different artificial intelligence disciplines.

(Roy et al. 2019), a review of deep learning models with EEGs, states that there are a large number of works that cannot be reproduced as data is unavailable. It also points out that more than half of the studies use publicly available datasets. Considering also the difficulty of obtaining EEGs and the computational cost of developing deep neural models, it seems clear that there is a need to have a reference resource that can be consulted by researchers in this field (both physicians and computer scientists) to know which datasets are available or which models perform better.

This paper also presents an innovative character because, to the best of our knowledge, it is the only one that has studied what open EEG datasets can be found in the scientific community.

The main contribution of this paper is to present a systematic review of open EEG datasets used in works using DL techniques. The paper follows a methodology to obtain scientific papers utilizing this kind of dataset. Papers have been searched in the best-known scientific sources using a set of keywords to focus the searches. However, as EEGS datasets are scarce, not many open datasets are available, so there are not many papers that meet the selection criteria. After discarding some of them that either did not use an open EEG dataset with a DL model, did not provide model performance metrics, or did not include a description of the dataset and a link to download it, we remained with 37 works.

In the process, we provide a set of statistical metrics alongside some graphics that let to understand the information. This is useful in the following cases. The content will let researchers know which are the most used deep learning techniques and which accuracies they get depending on the dataset. It also could help scientists choose which models perform well with their specific data or use case. Another provided information is which datasets are available and how they perform, this is useful when researchers want to develop a new model/method and test it or know which models are not applied a lot. Another interesting usage is also to know which type of use cases does not have an open dataset that could be used by the scientific community, so people could create a new one. Finally, compiling the information of which are the most common values for the main features of the datasets (number of channels, sample rate, etc.) could help us to build a golden standard of the dataset.

This work has the following sections: Section 2 describes the DL techniques used in the papers. Section 3 details the method followed in compiling the documents. Section 4 compiles all the studied features. Section 5 contains an in-depth discussion of the datasets and their use. Finally, Section 6 presents some conclusions about the research.

II. STATE OF THE ART

By considering the creation of AlexNet as the main milestone in deep learning (Krizhevsky, Sutskever, and Hinton 2012), the number of papers in medical bibliographic databases has been growing exponentially yearly. Figure 2 shows the number of publications from 2012 to 2023 containing the word deep learning. We can see that, in some cases, the number of publications doubled from one year to another. Then, considering the period from 2018 to 2023, we can see that most of the scientific production in this field is during those years. Even during this year, more than 4000 papers have been published in about 2 months and a half which means that at the end of the year, the number of papers will be greater than in 2022.

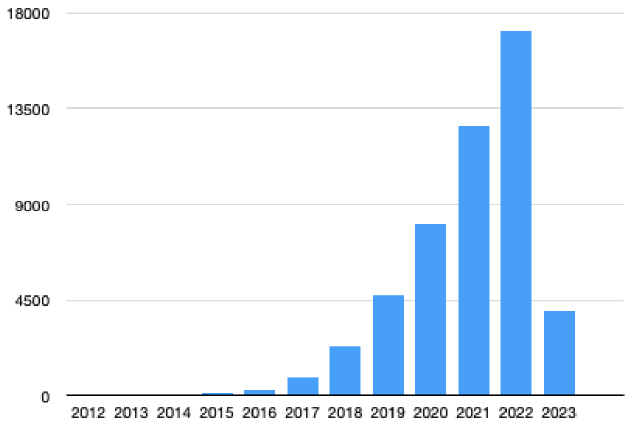

FIGURE 2. Distribution by the publication year of the deep learning papers indexed in PubMed from 2012 to 2023 (n=4524).

DL techniques are based on multiple models and architectures, although their effectiveness is directly related to the nature and quality of the data used in the training stage. This section describes the architectures and models that can use EEGs.

DL models can be classified depending on how they learn from the data. This case has three main classes: supervised, unsupervised, and semi-supervised.

Supervised models need labelled data to perform the training. In this case, the model knows the relation between input data and the expected output and uses the following classification.

Multilayer perceptron (MLP) is the simplest case of a DL model. The architecture comprises an input layer, several hidden layers, and an output layer. (Lin et al. 2007) use an MLP to classify EEG signals depending on the music some subjects are listening to.

Convolutional neural networks (CNNs) are the most used models with several applications in computer vision. Its primary ability is to detect patterns in a delocalized way. This characteristic lets to learn a particular pattern in an image that can later be seen in another part of another image. Recently, a specific type of CNN that manages graphs called Graph Convolutional Neural Networks (GCNN) has arisen. (Kipf and Welling 2016) presents this model as a method that encodes a graph's structure and its nodes' features using a special type of CNNs. CNNs are used by (M. Zhou et al. 2018) to classify epileptic seizures. GCNN recognizes emotions by analyzing EEGs, (Song et al. 2018).

Recurrent neural networks (RNNs) are defined by (Elman 1990) as a model that uses an input vector of arbitrary length and applies a transition function recursively to its internal hidden state vector h_t. It uses data structures that are time series, for example, EEGs. Within RNNs, a particular type is called Long Short-Term Memory networks (LSTM) or Gated Recurrent Unit (GRU). LSTMs were proposed to work with noisy or incomprehensible input data without information loss (Hochreiter and Schmidhuber 1997). In the case of RNNs, (Ruffini et al. 2016) applies them to the prognosis in patients with neurodegeneration. Then, LSTMs have applications in emotion recognition (Alhagry et al., 2017). Finally, GRU has been applied in emotion classification, (Chen, Jiang, and Zhang 2019)

The other leading group of models belongs to the category of unsupervised models. In this case, data is unlabeled, and there is no a priori knowledge about the final results (Sathya, Abraham, and others 2013).

Deep Autoencoders (DAE) use unsupervised learning. Defined in (Ballard 1987), its particular characteristic is that both the input and output layers have the same or similar size and two processing structures. The first one is the decoder which starts from the input data and reduces its size to a small piece that contains its main characteristics. The second part is the decoder which aims to upsample the previous small piece of data by upsampling it until reaches the input data size. In (Qiu et al. 2018), autoencoders classify ictal EEG. We consider Restricted Boltzmann Machines (RBM) as a particular type of Autoencoder introduced by (Smolensky 1985) that can learn a probability distribution. In DL, RBMs were used to implement Deep Boltzmann Machines (DBMs), (Salakhutdinov and Hinton 2009). The field of EEGs has applications like (Lu et al. 2016) that applies it to motor imagery.

The previous learning types generate a new one by mixing them and are called semi-supervised. Generative adversarial networks (GAN) are under this class. GANs need neural models, the generator, and the discriminator. Both work in a training type called adversarial process, (Goodfellow et al. 2014). This architecture aims to learn and imitate a data distribution. The generative model is responsible for creating synthetic instances of the input data. Then, the discriminator evaluates these data and decides if it is similar enough to the input data or not. This task gives a probability of being authentic (input data) or synthetic (created by the generator). By repeating this process, the generator learns how to create data more like the input one. In this case, GANs are applied to perform data augmentation strategies with EEGs (Luo and Lu 2018).

It is noteworthy that in recent years a trend in the creation of hybrid models has been detected. These types of models are seen as an important area of development within the DL

in the near future, (Kurz et al. 2022). These architectures join two or more models generating a CNN-LSTM or an Autoencoder-LSTM.
To summarize all the information above, Figure 3 shows a taxonomy with all the deep learning models.

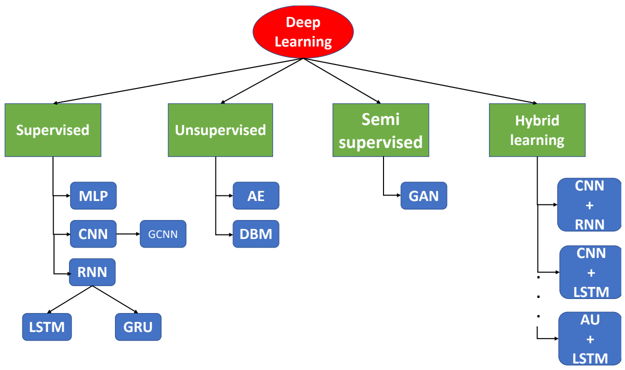

FIGURE 3. Taxonomy of the different deep learning models considering the different approaches to train the models.

Metrics are an important aspect when evaluating a DL model. Four are the most important in this type of analysis: accuracy, specificity, sensitivity, and F-1 score. Accuracy is defined as the ratio between successful predictions and the total number of predictions. This metric is used as a way to measure the performance of a model in the first moment. Specificity measures the ratio between the number of true negatives (healthy people diagnosed correctly) and the total of those predicted as true negatives and false positives (healthy people diagnosed as sick). Sensitivity is similar to specificity but considers true positives instead true negatives and false negatives instead of false positives. F1-score considers true positives, false positives, and false negatives as described in the following Equation.

$$\frac{2*TP}{2*TP+FP+FN} \tag{1}$$

III. METHODOLOGY

The method used to determine which research works are framed in a particular field or respond to the needs of certain research questions is called Systematic Literature Review (SLR) or just systematic review. There are different guidelines for conducting an SLR. For example (Keele and others 2007) includes: necessity of the review, research questions, development of the protocol, identifying the research works, establishing some inclusion and exclusion criteria, analyzing some features of the papers and creating the review as a paper. As the phases of the review process can differ, we have used (Barua, Ahmed, and Begum 2023), (Thongchotchat et al. 2023), (Marican et al. 2022) and (Bujang et al. 2022) as references to design our process which is described in Figure 4. First, we formulate a set of research questions. Then, we start the process of finding and selecting the research works, from where we collect the datasets. Following, we analyze them, and the papers where they are used. Finally, we describe all this information in the presented paper.

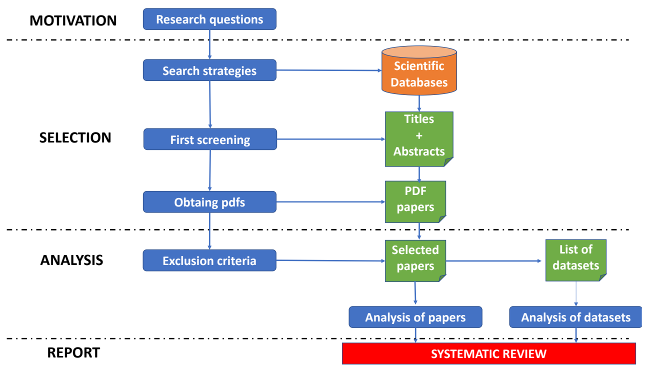

FIGURE 4. The review process starts with the research questions conducted by the researchers and ends with the written present work.

A. FORMULATED RESEARCH QUESTIONS

As the main aim of this systematic review is compiling open datasets of EEGs that have been used with DL models, some information could be analyzed like the characteristics of the datasets or the deep learning models. In this way, the following research questions are proposed as a way to understand the purpose of the review and its utility.

Question 1: Which EEG datasets are freely available to researchers so they can perform their studies in deep learning?
Motivation 1: As EEGs are difficult to compile due to the time needed to do the test or the number of patients and controls, this data is scarce. This information source can be consulted by them to find data for their research.

Question 2: Which values have the main characteristics of the datasets? The number of channels, frequency, etc.
Motivation 2: This is key for researchers when establishing a protocol to compile their own data. This decision must be taken by both profiles: physicians and computer scientists. This assures that the data accomplishes with a minimum quality so the deep learning models could be appropriate and cover a wide range of use cases.

Question 3: Which deep learning models perform better with electroencephalograms and their use cases?
Motivation 3: Given the metrics compiled in this work and the deep learning models that have been obtained, researchers can know which deep learning models best fit the different datasets depending on the characteristics and the use cases: diagnosis, motor imagery, etc. It also lets researchers know if the datasets are good enough to apply DL techniques.

B. SEARCH STRATEGY FOR IDENTIFYING THE STUDIES

To obtain the papers, we have set the following keywords to be used in every scientific source: (("open dataset") OR ("free dataset") OR ("freely available dataset") OR ("open data") OR

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

("free data") OR ("freely available data")) AND (("EEG") OR ("electroencephalogram")) AND (("deep learning") OR ("neural network") OR ("neural networks")). The search and collection of papers include everything published until March 15, 2023. The following sources were used to make the searches: Scopus[1], PubMed[2], Web of Science (WOS)[3], Science Direct[4], IEEE Explorer[5], and SpringerLink[6]. After discarding repeated items, conference papers surveys, or arxiv papers, the final selection of works has been made to apply more restrictive criteria.

C. CRITERIA FOR SELECTING PAPERS

A group of computer scientists has set out the following criteria to obtain the final set of papers.

The first selection of works consisted of a single screening where titles and abstracts are read to check if they meet the minimum inclusion criterion of "a paper that uses an EEG open dataset to train a deep learning model". Searches in scientific resources were made. Then titles and abstracts were read, and those that did meet the criteria of including an open EEG dataset used with DL models to solve a particular use case were included for the following step.

As there is no way to automatize a more exhaustive process of selecting the papers, several quality requirements have been set out. This is a set of exclusion criteria that discard papers accomplishing the following:

1. The paper does not describe the DL model which is trained with a dataset of EEGs.
2. Metrics about the performance of the models are not included in the evaluation.
3. The paper does not include a detailed description of the dataset and a link to download it. Datasets available upon request are not considered. The EEGs are obtained from humans.

D. PRISMA FLOW DIAGRAM

This systematic review compiles a set of papers by using the following methodology. Figure 5 contains a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Moher et al. 2009), which summarizes how we achieved the selection of papers used to compile the datasets reported in the paper.

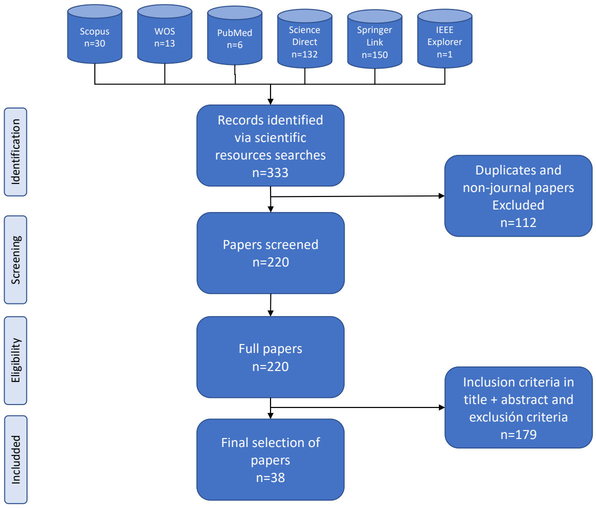

FIGURE 5. PRISMA diagram of the bibliographic review conducted starting with the findings in each scientific resource and finishing with the number of reviewed papers.

From the first search, a total of 331 works has been obtained which are distributed as follows: Scopus (30), PubMed (5), WOS (13), Science Direct (132), IEEE Explorer (1), and SpringerLink (150). After eliminating duplicates and papers not published in journals (conferences, arxiv, etc.), it remained a total of 219 papers were. The next step was to analyze their titles and abstracts to check if the paper applies deep learning models in an open dataset of EEGs. If yes, we must check if they accomplish the three exclusion criteria. The previous decisions eliminated different papers, including one whose dataset was unavailable for download, MERTI-Apps (Maeng et al., 2020). After this process, the final set had 38 papers from which we are analyzing some features related to the used deep learning models and obtaining the report of the open EEG datasets.

E. THREATS TO VALIDITY

A systematic review can be put at risk due to a potential biasness and the imprecise application of the extraction method. To evaluate this, four dimensions are considered: internal validity, external validity, construct validity, and conclusion validity.

Internal validity: depending on the search process a validity threat can impact the representativeness of the selected scientific works. To avoid that, we have used (Barua, Ahmed, and Begum 2023), (Thongchotchat et al. 2023), (Marican et al. 2022), and (Bujang et al. 2022) as guidelines to adjust our process. The research questions have been a guide to constructing the searches and thinking about the inclusion and exclusion criteria that best fit them. The selection of keywords and scientific resources could be a

[1] https://www.scopus.com/
[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://www.webofscience.com/
[4] https://www.sciencedirect.com/
[5] https://ieeexplore.ieee.org/
[6] https://link.springer.com/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3293421

limitation, but we have tried to avoid it by using very clear terminology and using official resources like WOS, PubMed, etc.
External validity: We have limited the papers to works published in scientific journals discarding preprints, books, conferences, etc. This allows for obtaining strong conclusions that could be useful for scientists in related fields.
Construct validity: Research questions are one of the main pieces to guide the review and think about its utility. For this purpose, these research questions have been discussed among the authors and other researchers in the field.
Conclusion validity: To address the potential subjectivity of our study, the authors read the title and abstract for reviewing the first screening studies. The aim of this approach was to minimize bias in the extraction of data. In the event of disagreements between them, a consensus was reached through discussions. This approach ensured that the data collected was both reliable and objective.

F. DATA EXTRACTION AND CLASSIFICATION OF THE STUDIES

The paper comprises two types of studies on this set of papers. The first is on the deep learning models used with the datasets, the metrics to measure performance, the use case solved in the paper, and the paper's year of publication. The second compiles some characteristics related to the dataset: number of channels, number of individuals, distribution system, sampling frequency, or the format of the file.

IV. RESULTS

Some statistics have been obtained based on the previous characteristics used in the papers and datasets. This information was compiled by developing some Python scripts and using Matplolib, which provides a graphic set of charts, (Barrett et al. 2005). Apart from that, we have developed a VOSviewer[7] with the keywords in the selected paper to confirm their relation to the keywords used in the original searches. Figure 6 shows a VOSviewer with this information.

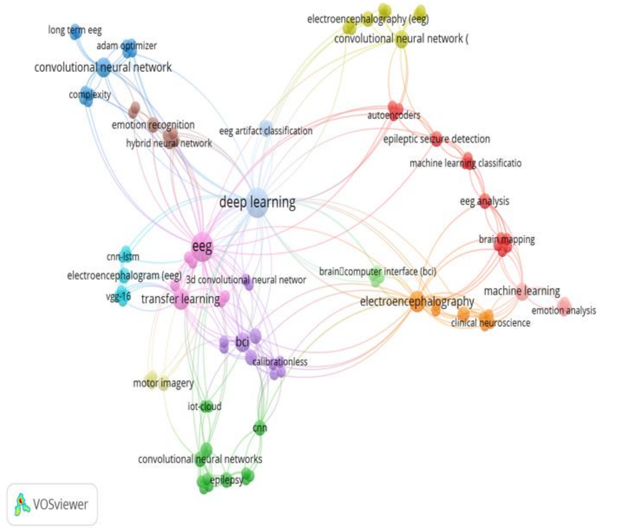

FIGURE 6. **VOSviewer network visualization describing the relationship between keywords in the encelography and deep learning fields.**

In the Figure above, we can see that the biggest nodes correspond to "deep learning" and "eeg" which makes sense as are key terms in our searches. Then, there are other keywords similar to the latter like "electroencephalogram" or "electroencephalography". Related to DL, there are keywords that describe the different models: "convolutional neural network", "autoencoders" or "cnn-lstm". There are also some keywords that define the use cases: "emotion recognition", "motor imagery" or "epileptic seizure detection" which, as we will see later, are important in the study of datasets.
For the analysis of papers and datasets, the following graphs have been provided. Looking at the papers: bar diagrams with the publication year, percentages of the deep learning models used, DL metrics, the use case that has been solved, and the relation between the deep learning models and the use cases. By taking into account the datasets' characteristics: number of individuals and tests, length of the tests, number of channels, distribution of the electrodes systems that have been used, frequency in hertz during the test, and the file format provided to work with the data.

A. SUMMARY OF PAPERS

Table 1, attached at the end of the paper shows the selected set of papers with the following information: the paper's reference, the deep learning model used, the metrics applied during the experimentation, the tasks performed by the individuals while compiling the data, and the year of publication of the work. Apart from that, we have added a matrix of evaluation by using the Composite Quality Indicator (CQI) index of the i-th paper is determined by combining the normalized indicator values of the other indexes within the range of 0 to 1.

[7] https://www.vosviewer.com/

TABLE I
SUMMARY OF SELECTED PAPERS

B. STATISTICS AND ANALYSIS OF THE INCLUDED STUDIES

This section provides graphs and statistics from analyzing the selected papers related to the use of open datasets. Figure 7 shows a bar chart distributing the 38 papers by year of publication.

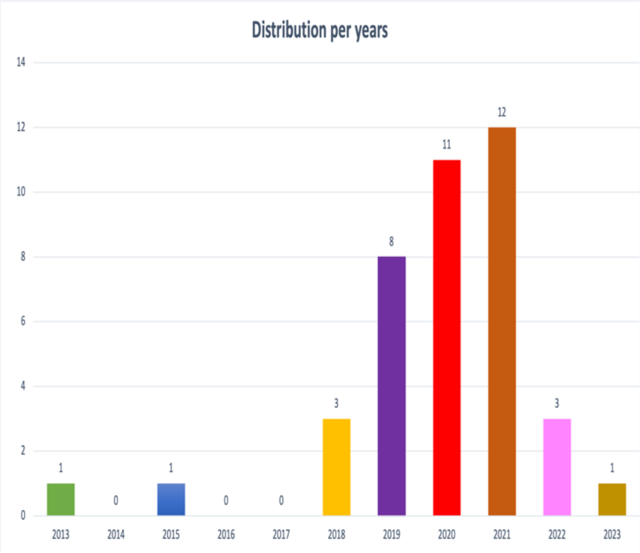

FIGURE 7. Distribution by number of selected papers published each year (n=38).

As you can see a total amount of 38 papers have been selected, each of them covering different use cases, applying different deep-learning models with their own experiments. Following, we briefly described each paper:

1. (Cui et al. 2021) presents easily interpretable CNN designed to identify shared EEG features among various subjects for the detection of driver drowsiness. The model incorporates the Global Average Pooling (GAP) layer within its structure, which enables to leverage of the Class Activation Map (CAM) method. This utilization of CAM allows to localize of the regions in the input signal that contributes the most to the classification process.
2. (Abdelhameed and Bayoumi 2021) describes a novel approach that exploits the automatic feature learning capabilities of a two-dimensional deep convolution autoencoder (2D-DCAE), which is coupled with a neural network-based classifier. This unified system is trained to optimize classification accuracy between the ictal and interictal brain state signals.
3. (Shalash 2021) introduces a method to estimate driver fatigue state using a single EEG channel signal. The process involved pre-processing the EEG signals, transforming them into color images using spectrogram analysis, and then classifying them as fatigue or normal fatigue using CNNs.
4. (Juárez-Guerra et al. 2020) introduces a model named Multidimensional RadialWavelons Feed-Forward Wavelet Neural Network (MRW-FFWNN), for classifying epileptic seizures based on EEG signals. The network differentiates between three brain states associated with epilepsy: ictal, interictal, and healthy.
5. (G. Xu et al. 2019) proposes a framework that combines a pre-trained VGG-16 model with a target CNN model for MI EEG signal classification.
6. (Wu et al. 2019) introduces a parallel multiscale filter bank convolutional neural network (MSFBCNN) for MI classification. The proposed network consists of a layered end-to-end structure, where a feature-extraction network is employed to extract both temporal and spatial features. This approach enables effective feature representation for accurate MI classification.
7. (Korkalainen et al. 2019) develops a deep learning method for automatically classifying sleep stages. Additionally, the study aims to study the impact of obstructive sleep apnea (OSA) severity on the classification.
8. (Eldele et al. 2021) presents AttnSleep a novel attention-based deep learning architecture proposed for sleep stage classification using single-channel EEG signals. It consists of a feature extraction module utilizing a multi-resolution convolutional neural network (MRCNN) with adaptive feature recalibration (AFR), and a temporal context encoder (TCE) with multi-head attention.
9. (Baser et al. 2022) creates a real-time algorithm for the non-invasive detection of spike-and-wave discharges (SWDs) in EEG recordings of individuals with absence epilepsy. The proposed approach involves utilizing CNNs and Thomson's multitaper power spectral density estimation analysis to represent the power of EEG signals as a function of frequency and time.
10. (Yan et al. 2021) utilizes 2D-CNN to extract features from polysomnographic signals. The model incorporates a "squeeze and excitation" block to recalibrate channel-wise features and includes an LSTM module to capture long-range contextual relationships. The learned features are then passed to the decision layer to generate sleep stage predictions.
11. (San-Segundo et al. 2019) describes a model that consists of two convolutional layers for feature

extraction and three fully connected layers for classification. Multiple EEG signal transforms, including Fourier, wavelet, and empirical mode decomposition were used as inputs of the model.

12. (Alhussein et al. 2018) uses deep learning to develop a cognitive Internet of Things (IoT) framework using CNN and Autoencoders. In this approach, smart EEG sensors are utilized specifically for recording and transmitting EEG signals from epileptic patients.
13. (L. Xu et al. 2020) validates deep learning models for MI and addresses the issue of cross-dataset variability. To mitigate this problem, an online pre-alignment strategy is proposed to align the EEG distributions. Experiments are performed with EEGNet and ShallowNet.
14. (Schirrmeister et al. 2017) investigates various deep CNNs for decoding imagined or executed movements from raw EEG signals. The results demonstrate that incorporating as batch normalization and exponential linear units, along with a cropped training strategy, significantly improved the decoding performance of the deep learning models.
15. (Maeng, Kang, and Kim 2020) develops an annotation labeling program for emotions by arousal and valence. During the learning phase, a Genetic Algorithm (GA) was employed to select optimal parameters for an LSTM model and determine the active feature group from EEGs in the time, frequency, and time-frequency domains.
16. (J. Zhang, Xu, and Yin 2023) implements a tool for depression screening by using EEG signals from 128 channels. The experiments consist of a 2D-CNN-LSTM classifier, support vector machine, K-nearest neighbor, and decision tree were employed.
17. (Y. Li, Zhang, and Ming 2023) examines the integration of EEG and fNIRS data through a bi-modal fusion approach. A Y-shaped neural network was developed to combine the bimodal information at different stages of the network.
18. (Dar et al. 2022) combines 1D-CRNN with an Extreme Learning Machine (ELM). The proposed architecture is robust for emotion detection in Parkinson's disease (PD) patients and can handle cross-dataset learning with different emotions and experimental settings.
19. (Islam et al. 2021) introduces a CNN for emotion recognition. The EEG data were transformed into images representing Pearson's Correlation Coefficient (PCC). These images were then fed into the CNN model to classify emotions. Two protocols were conducted, protocol-1 for two-level emotion identification and protocol-2 for three-level recognition of valence and arousal.
20. (J. Liu et al. 2020) introduces a model for emotion classification using EEGs. The proposed approach combines CNN, Sparse Autoencoder (SAE), and Deep Neural Network to enhance feature extraction and classification. The CNN extracts features, which are then encoded and decoded by the SAE to reduce redundancy. The resulting features are fed into the DNN for the classification task.
21. (Zhu et al. 2021) aims to re-implement a compact CNN called EEGNet and evaluated its feasibility for decoding SSVEP in ear-EEG signals. To further improve classification accuracy different kernel numbers were used.
22. (Kwon, Shin, and Kim 2018) uses a CNN for emotion recognition. Classification performance was improved by combining EEG and galvanic skin response (GSR) signals. The GSR signals were preprocessed using the zero-crossing rate.
23. (Topic and Russo 2021) employs Deep learning as a feature extraction method, where feature maps are utilized to extract relevant features. These extracted features are then fused together to enable the classification of various types of emotions.
24. (F. Li et al. 2021) implements a 34-layer deep residual CNN model for sleep staging classification. The network takes raw single-channel electroencephalogram (Fpz-Cz) signal as input and yields hypnogram annotations for each 30s segment as output.
25. (Bairagi et al. 2021) creates a patient-specific computer-aided model that can detect epileptic seizures. This model incorporates an FIR filter, DWT (Discrete Wavelet Transform), ANN, and a newly proposed sequential window algorithm (SWA).
26. (Salehzadeh, Calitz, and Greyling 2020) presents a deep learning-based framework for accurately classifying EEG artifacts based on a person's physiological activity. The framework utilizes a processing pipeline that combines CNN and LSTM.
27. (Y. Liu et al. 2020) introduces methods for EEG-based mental fatigue recognition that utilize inter-subject transfer learning, eliminating the need for calibration. To identify the most informative features, it employs different machine learning techniques such as deep learning.
28. (D. Wang and Shang 2013) propose a system based on Deep Belief Networks (DBNs) for automatic feature extraction from raw physiological data. The extracted features are then

used to build classifiers for predicting arousal, valence, and liking levels.

29. (von Atzingen et al. 2022) uses CNN to analyze the brain signals of 11 healthy subjects during the tasting of drinks with different sweeteners. The study aimed to gain insights into the neural responses associated with them.
30. (Podmore et al. 2019) studies how a CNN can classify frequency and phase-encoded SSVEP signals.
31. (Sarmiento et al. 2021) develops a novel algorithm called CNNeeg1-1 using DL techniques to recognize EEG signals associated with imagined vowel tasks.
32. (Ma et al. 2021) proposes a model for the automatic diagnosis of epilepsy. It replaces the traditional CNN with a one-dimensional CNN. Additionally, the model combines an independent recurrent neural network (indRNN) with CNN.
33. (Guillot and Thorey 2021) presents RobustSleepNet as a deep learning model designed for automatic sleep stage classification. The model is trained and evaluated using a diverse set of 8 sleep staging datasets. This approach aims to ensure the model's robustness to demographic variations and improve its generalizability across different populations.
34. (Huang et al. 2020) introduce an approach based on timestamp-based segmentation (TSS) and multichannel analysis for sleep stage classification. TSS utilizes a CNN to predict sleep stage labels at each timestamp. Then incorporates multiple channels of PSG recordings, including EEG, EOG, EMG, and leg electromyogram.
35. (Miao et al. 2020) proposes a novel deep learning methodology for spatial-frequency feature learning and classification of motor imagery EEGs. The approach involves a CNN model specifically designed to capture the spatial-frequency characteristics.
36. (Pedoeem et al. 2020) presents a tool to support neurologists to expedite the analysis of EEG data, enabling doctors to assess more patients and allocate additional time to treatment. The model adopts a transformer-based neural network architecture for seizure detection.
37. (Bassi, Rampazzo, and Attux 2021) employs CNNs to classify EEG signals in a single-channel BCI based on SSVEP. Notably, the approach eliminates the need for calibration with the user, enhancing the usability of the system.
38. (Banville et al. 2021) explores self-supervised learning (SSL) techniques for EEG analysis, specifically temporal context prediction and contrastive predictive coding. Applying SSL to EEG-based sleep staging and pathology detection, the model us able to discover meaningful patterns in unlabeled data.

Another relevant piece of information that can be obtained from this preliminary analysis is related to the first research question; the type of DL model used. This knowledge is helpful for researchers to determine which are the most powerful models for processing EEGs. The usage percentages are collected in the following pie chart, Figure 8. As more than one model can be used in a paper, the number of instances is bigger than 38.

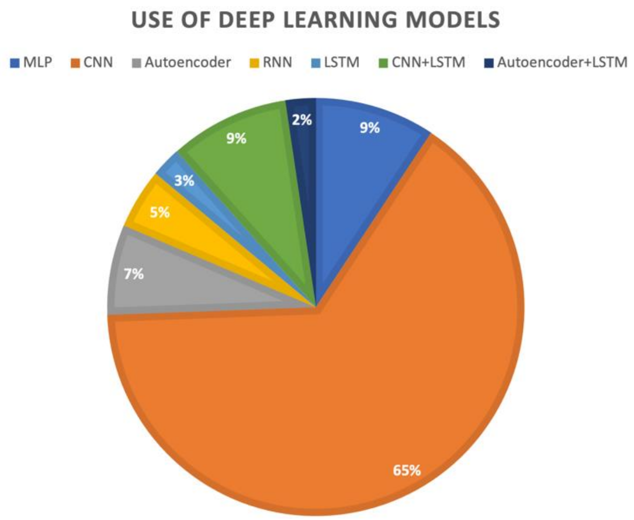

FIGURE 8. Pie chart with the percentages of the deep learning models applied in the selected papers (n=40).

Another quality criterion for selecting a paper is the use of metrics to evaluate the performance of the models. Following, we compile a pair of aspects related to them. Again, it should be highlighted that a paper can use more than one metric. Figure 9 represents a diagram of bars that counts the times each DL metric appears in the set of selected papers which is interesting to know which of them researchers should apply in their works.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3293421

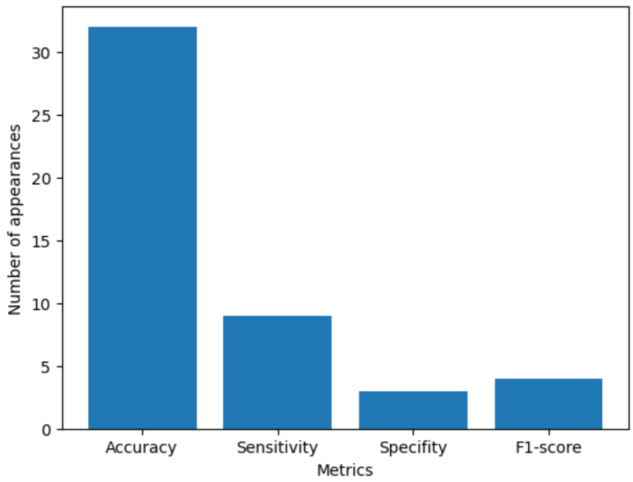

FIGURE 9. Distribution of the machine learning metrics applied in the selected papers (n=48).

Another graphic related to metrics is the following boxplot, where we represent the distribution of the values obtained in the different papers after training the different DL models, Figure 10.

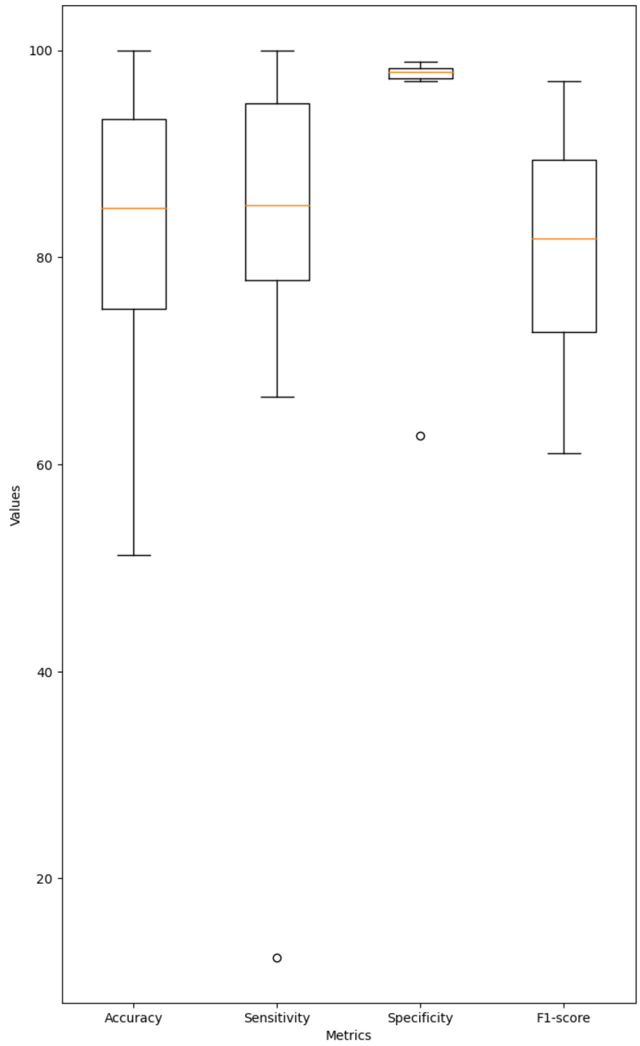

FIGURE 10. Boxplots describing the distribution of the precise values of each metric in the experiments performed in selected papers (n=99).

EEGs can solve several use cases. This information helps us know which application fields are less exploited, so there is scope for further research. Figure 11 uses a pie chart to describe this information which we have classified the datasets into 8 general categories:

- Motor imagery (MI) classification. This application aims to recognize a subject's intention, (Lu et al. 2016).
- Seizure management. EEGs of patients with epilepsy, a brain disorder that consists of abnormal cerebral activities.
- Estimation of sleep stages. Datasets collect the five possible stages a human can experiment with while sleeping.
- Recognize emotions. This task consists of classifying human emotional states as the domains of arousal and valence.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3293421

- Classify levels of fatigue. Mental fatigue happens when a subject has paid attention to a task for a long time. These datasets can measure different levels of fatigue, in some cases while driving.
- Disease diagnosis. In the medical field, we typically find datasets of epilepsy, but others can diagnose diseases such as Attention Deficit and Hyperactivity Disorder (ADHD).
- Brain stimuli. It measures how the brain responds to different perception tasks. For example, the response to images or the consumption of sweetened drinks.
- Human activity recognition. This is a way to detect artifacts while performing tasks such as reading, watching, and speaking.
- Depression screening. the process of assessing individuals for symptoms and indicators of depression in order to identify those who may be at risk or currently experiencing depression.

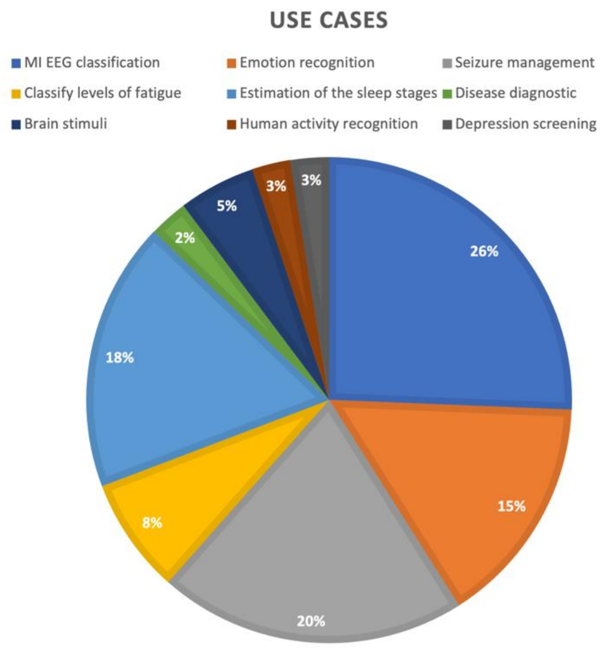

FIGURE 11. Pie chart with the percentages of the different use cases conducted in the reviewed datasets (n=38).

Figure 12 combines the results of both previous analyses in a bubble diagram where the X-axis represents the deep learning model and the Y-axis the possible use cases. This information is interesting when a scientist needs to decide what DL models could be used depending on the use case they are working on. The bubble size and colour depend on the number of instances.

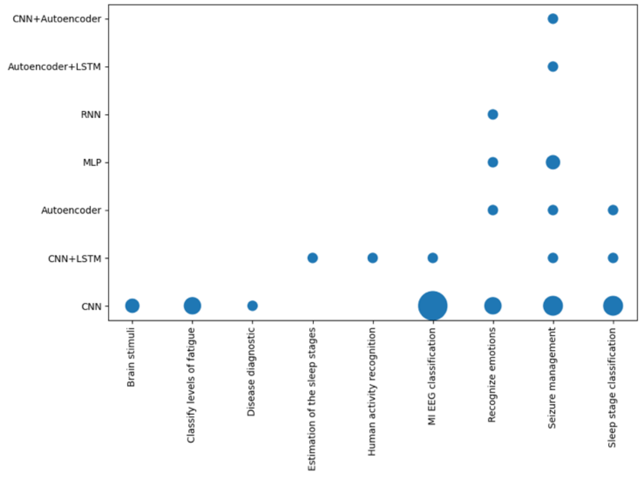

FIGURE 12. Relationship between the deep learning models used in the selected papers and use cases carried out in the reviewed datasets.

C. SUMMARY OF DATASETS

As we have said before, we have applied the PRISMA method to obtain a set of papers from which we are analyzing the datasets used in them. The following is a brief description of them.

1. BCI competition IV 2a[8]: the imagination of movement of the left hand, right hand, both feet and tongue, (Ang et al. 2012)
2. BCI competition IV 2b[9]: motor imagery of left hand and right hand, (Brunner et al. 2008)
3. DEAP and video signals[10]: emotion recognition of low arousal and low valence (LALV), high arousal and low valence (HALV), low arousal and high valence (LAHV,) and high arousal and high valence (HAHV), (Koelstra et al. 2011)
4. Multichannel EEG sustained attention driving task[11]: fatigue and non-fatigued during driving, (Zhao et al. 2019)
5. Temple University EEG Corpus[12]: a compilation of different neural diseases, (Obeid and Picone 2016)
6. CHB-MIT Scalp EEG Database[13]: seizure and non-seizure states in epileptic patients, (Shoeb 2009)
7. MAHNOB-HCI[14]: a scale of valence and arousal, (Lichtenauer et al. 2011)

[8] https://www.bbci.de/competition/iv/#dataset2a
[9] https://www.bbci.de/competition/iv/#dataset2b
[10] https://www.eecs.qmul.ac.uk/mmv/datasets/deap/
[11] https://figshare.com/articles/dataset/Multi-channel_EEG_recordings_during_a_sustained-attention_driving_task/6427334/5
[12] https://isip.piconepress.com/projects/tuh_eeg/
[13] https://physionet.org/content/chbmit/1.0.0/
[14] https://mahnob-db.eu/hci-tagging/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3293421

8. Sleep EDF[15]: sleep stages after temazepam intake and after placebo intake, (Goldberger et al. 2000)
9. Motor Imagery dataset from Weibo et al. 2014[16]: simple MI (left hand, right hand, and feet) and compound MI (both hands, left hand combined with the right foot, right hand combined with the left foot), (Yi et al. 2014)
10. PhysioNet/CinC Challenge 2018[17]: wakefulness, stage 1, stage 2, stage 3, rapid eye movement (REM), and undefined, (Ghassemi et al. 2018)
11. Open source SSVEP dataset[18]: healthy subjects focused on 40 characters flickering at different frequencies, (Y. Wang et al. 2016)
12. BCI Competition III IVa[19]: MI of the left hand, right hand, and right foot, (Blankertz et al. 2004)
13. EEG data for driver fatigue detection[20]: drivers suffering fatigue or not, (Min, Wang, and Hu 2017).
14. University of Bonn[21]: seizure and non-seizure states, (Andrzejak et al. 2001).
15. Motor Imagery dataset from Zhou et al. 2016[22]: MI of the left hand, right hand, and feet, (B. Zhou et al. 2016)
16. Sleep Heart Health Study[23]: sleep scores, (Quan et al. 1997)
17. EEG datasets for motor imagery brain-computer interface[24]: data for non-task-related and task-related states, (Cho et al. 2017)
18. DOD-O[25]: scored apnea patients, (Guillot et al. 2020).
19. DOD-H[26]: scored sleep stages, (Guillot et al. 2020).
20. CAP sleep database[27]: activity during NREM sleep, (Terzano et al. 2002)
21. Bern-Barcelona EEG database[28]: patients have pharmacoresistant focal-onset epilepsy, (Andrzejak, Schindler, and Rummel 2012)
22. MrOS Sleep[29]: sleep study, (G.-Q. Zhang et al. 2018)
23. Database-Imaged-Vowels-1[30]: pronounce the five main vowels "a", "e", "i", "o", and "u" and six Spanish words, (Coretto, Gareis, and Rufiner 2017)
24. EEG+NIRS Single-Trial Classification[31]: it conducts two BCI experiments: left versus right-hand motor imagery; mental arithmetic versus resting state, (Shin et al. 2016).
25. MODMA[32]: this is a dataset for mental-disorder analysis which includes clinically depressed patients and controls, (Cai et al. 2020)
26. BehaveNET[33]: human task recognition of reading, speaking and watching TV.
27. EEG Sweeteners AI[34]: this study evaluated brain signals from 11 healthy subjects when they tasted passion fruit juice equivalently sweetened with sucrose, sucralose, and aspartame, (von Atzingen et al. 2022)
28. MESA[35]: sleep study to understand how variations in sleep and sleep disorders vary across gender and ethnic groups and relate to measures of subclinical atherosclerosis, (G.-Q. Zhang et al. 2018)
29. CBICIC2019[36]: it comprises two subsets of MI with left and right-hand tasks.
30. Deep BCI[37]: classification of steady-state visual evoked potentials (SSVEPs) based BCI from earEEG, (Kwak and Lee 2019).
31. AMIGOS[38]: valence, arousal, dominance, familiarity, and liking, and selected basic emotions.
32. SEED[39]: report of emotional reactions.

All the information in the datasets has been collected in the following table attached at the end of the paper. The most used datasets are Sleep EDF and DEAP.

TABLE II
SUMMARY OF SELECTED DATASETS

[15] https://www.physionet.org/content/sleep-edfx/1.0.0/
[16] http://moabb.neurotechx.com/docs/generated/moabb.datasets.Weibo2014.html
[17] https://archive.physionet.org/physiobank/database/challenge/2018/
[18] http://bci.med.tsinghua.edu.cn/download.html
[19] https://www.bbci.de/competition/iii/desc_IVa.html
[20] https://figshare.com/articles/dataset/The_original_EEG_data_for_driver_fatigue_detection
[21] https://ebrary.net/59044/education/details_public_databases
[22] http://moabb.neurotechx.com/docs/generated/moabb.datasets.Zhou2016.html
[23] https://sleepdata.org/datasets/shhs
[24] https://academic.oup.com/gigascience/article/6/7/gix034/3796323
[25] https://dreem-dod-o.s3.eu-west-3.amazonaws.com/index.html
[26] https://dreem-dod-h.s3.eu-west-3.amazonaws.com/index.html
[27] https://archive.physionet.org/physiobank/database/capslpdb
[28] https://www.upf.edu/web/mdm-dtic/-/1st-test-dataset#.YfgOG1jMIUo
[29] https://sleepdata.org/datasets/mros
[30] http://www.ifp.illinois.edu/speech/speech_web_lg/data/mri/index.html
[31] http://doc.ml.tu-berlin.de/hBCI.
[32] http://modma.lzu.edu.cn/data/index/
[33] https://zenodo.org/record/2552600#.ZBdONuzMJ
[34] https://github.com/Atzingen/EEG_Sweetners_AI
[35] https://sleepdata.org/datasets/mesa
[36] https://www.datafoundation.cn/competitions/342/datasets
[37] http://deepbci.korea.ac.kr/
[38] http://www.eecs.qmul.ac.uk/mmv/datasets/amigos/
[39] https://bcmi.sjtu.edu.cn/home/seed/

IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

D. STATISTICS AND ANALYSIS OF THE OPEN EEGS'DATASETS

The first important feature in a dataset is the number of individuals which is directly related to the model behavior. (Roy et al. 2019) show that models increase their performance when the number of subjects exceeds 15. The number of tests is logically related to this feature. The values of both characteristics are compiled in Figure 13 which shows a double diagram bar with their distribution per dataset.

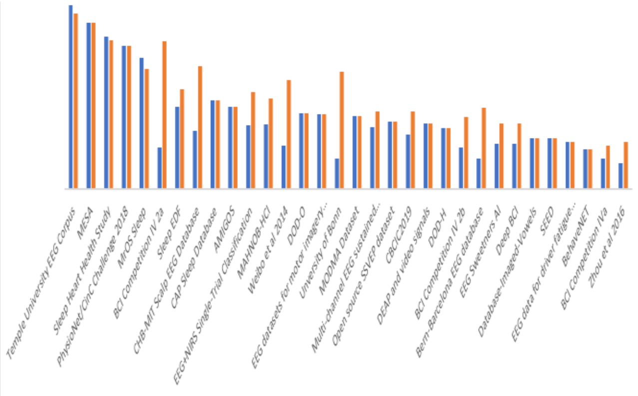

FIGURE 13. **Distribution of the number of individuals and test collected in each reviewed dataset (n=32).**

The number of channels is also a critical decision depending on the use case. (Jasper 1958) tells that a minimum of 21 channels should be used to examine an adult brain. This information is collected in Table II.
Another particular feature of EEGs is that of the electrodes system which indicates how electrodes are placed around the scalp. Figure 14 shows a pie chart with the percentage of datasets according to the system.

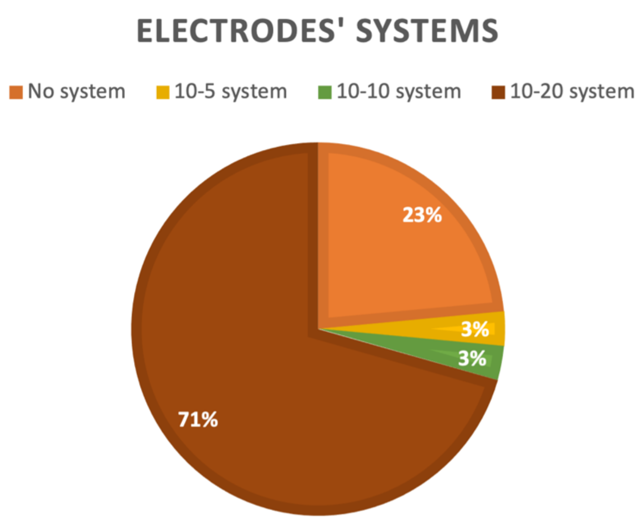

FIGURE 14. **Pie chart with the percentages of the electrodes' systems used in the reviewed datasets (n=32).**

Another interesting measure that will determine the performance of the model is sample frequency. The following bar diagram (Figure 15) represents the distribution of studies according to the frequency used to represent the data. This measure is directly related to the machine used to collect the data. In this case, different frequencies can be used in the same dataset.

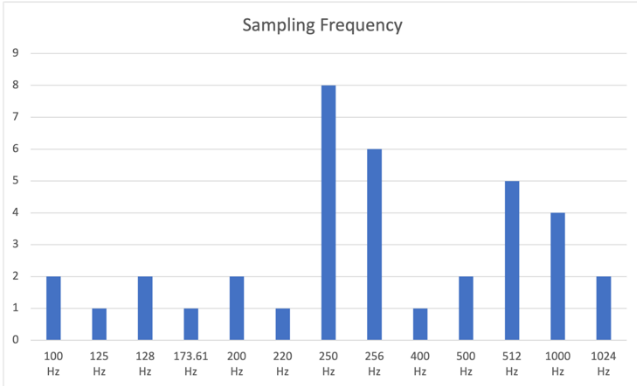

FIGURE 15. **Distribution of the sampling frequency used in the reviewed datasets (n=37).**

Finally, we have a pie chart that compiles the file format used, Figure 16. This depends on the different software used when doing the test.

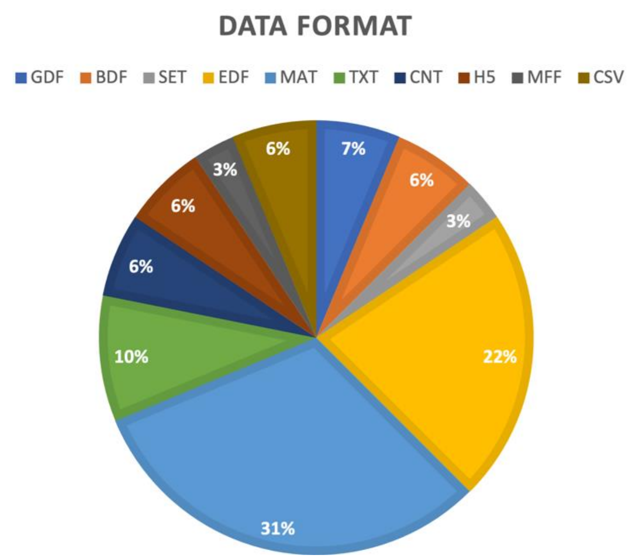

FIGURE 16. **Pie chart with the percentages of the file formats used in the reviewed datasets (n=32).**

V. DISCUSSION

This work provides a compilation of open EEG datasets analyzed using DL models in a set of papers selected by applying the PRISMA method in a systematic review. The results of the previous section are discussed below from a double perspective: on the one hand, the papers and the DL models used, and on the other, the datasets.

The first part of the statistical analysis starts with the year of publication of the papers. Figure 7 verifies, in part, the trend of papers in deep learning, mentioned in Figure 2. The number of articles published between 2018 and 2021 shows a significant increase. However, the total number is still small, and we can conclude that there is room for creating new open datasets available to the community. It is foreseeable that more papers on EEG and deep learning will be published in the coming years.

As can be seen in Figure 8, the most commonly used DL model, by a wide margin, is the CNN, which appears in 65% of the cases either as a 1-dimensional CNN (EEGs are processed channel by channel) or 2-dimensional CNNs, (EEGs are processed as a whole). Then there is a set of papers that use a hybrid model of CNN with LSTM, 10%. This is followed by hybrid models of Autoencoder and MLP (7%), RNNs (5%) and finally LSTMs or Autoencoder plus LSTM (3%). These numbers give us several ideas. First, using CNNs is successful but less innovative. This makes sense as EEGs can be managed as an image with convolutional filters. Second, using hybrid models seems an opportunity to make new contributions to the field. Finally, GAN and GCNN models are not used which is shocking. The first one has many applications in the creation of synthetic data (very useful considering the shortage of EEGs) or artifact removal (a typical task after collecting this data). The other can be used to model EEGs as graphs and study brain connectivity.

By analyzing the results in Figure 9, we can see that accuracy is the most used metric. This is meaningful as accuracy is the baseline metric to know if a deep learning model performs correctly. Otherwise, the fact of only working with accuracy leads to incomplete experiments as this metric only measures the number of hits. Accuracy has problems in models that use imbalanced datasets and does not give more interpretation of the performance of the model as it does not consider false positives and false negatives like sensitivity, specificity, and F1-score, (Mortaz 2020). Another conclusion obtained from the metrics is that none of the metrics measures the loss of the models which means that all the datasets are considered classification problems.

More information about the metrics is compiled in Figure 10. As we can see, all the metrics obtain values around 90% except specificity which performs near 100%, but with a small set of values. We can also see that the F1-score and sensitivity are more stable with the exception of an outlier in the latter with a poor performance near 10%. We confirm that accuracy is the most used but with a wider range of values which indicates is not as precise as the others.

Regarding Figure 11, the most frequent use case is MI EEG classification, with more than 27% of the cases. This fact is related to BCI competition IV[40], a famous data resource in the field comprising a set of datasets for signal processing and BCI classification. Then, we can highlight three use cases among the rest: seizure management, sleep stage classification, and emotion recognition. The rest of the use cases only occur once, twice, or three times: disease diagnostic, human activity recognition, brain stimuli, and classification in levels of fatigue. We can conclude with this analysis that if we want to publish a dataset that brings value to the field, the last four use cases are not exploited a lot.

The features of DL models and use cases are represented in Figure 12. The biggest bubble representing MI EEG classification with CNN makes sense because the DL model is the most popular in its category and there are several MI datasets. For example, those that are part of the BCI Competition. Regarding this use case, we can see that only models with CNN are used, so there is room to experiment with other models. In the second position, we have papers using CNN in stages of fatigue, sleep classification, and seizure management. The latter has been studied with several DL models, so it seems there are not many opportunities to work with this data. The information on the chart can be used to identify what models can be used with our own dataset. Also, to find combinations that have not been applied before to do new contributions to science. The rest of the combinations have few instances or none, so they can be considered niches to research. For example, using models that are not CNN in cases like levels of fatigue, brain stimuli, classification sleep stages or diagnostic of diseases.

The second part of the statistical analysis comprises the datasets' features. We first find the number of subjects which ranges from 4 to 16,986. This is directly related to the number of tests going from 4 to 10,874, the mismatch with the previous values is because the dataset with the most individuals has not recorded a test for each of them, Figure 13. In fact, this a strange situation as most of the time the number of tests is greater than the number of individuals. As we can see most of the datasets are in the low range which confirms that compiling EEGs is not an easy task.

Test duration ranges from seconds to hours (usually, these are sleep studies or patients with epilepsy). The length of the tests in seconds occurred 6 times ranging from 4 to 30 with an average of 16.93 seconds. In the case of minutes, we found 14 experiments with lengths from 4 to 51 and an average value of about 14 minutes. Finally, there are 12 examples with tests lasting at least 1 hour ranging to 9 with an average of almost 6 hours. This is directly related to the use case as epileptic seizures only need seconds to be analyzed but sleep stages need hours.

The number of channels used in the datasets has different values. In our case, the different options are well distributed

[40] https://www.bbci.de/competition/iv/

with works using only 1 channel and others using 128 channels. However, configurations of 64, 32, 16, and 8 channels, (Montoya-Martinez, Bertrand, and Francart 2019), which are recommended do not outstand. There is no analysis that supports this recommendation for deep learning studies, so it could be a future work to be developed.

Other features that have not been studied under a minimum standard to be met are the electrodes system and the sampling frequency. As can be seen in Figure 15, 10-20 is the most used electrode system by far, which makes sense due to the following aspects. It is an international recommendation, (Yang and Deravi 2017). (Association and others 2013) highlights that it is also one of the most used. Other datasets do not provide this information or do not use one due to the number of channels. Figure 16 describes the use of sample frequency. In the first position, we can find 8 times a sampling frequency of 250 Hz. Then, datasets using 256, 512, and 1000 Hz are also noteworthy. Regarding the minimum Hz to obtain good performances in DL models, (Wen et al. 2021) demonstrate that a higher frequency does not provide better results. Nevertheless, there are no scientific papers that measure the minimum to obtain DL models that perform well, so it could be a future approach.

Finally, Figure 17 gives information about the file formats that have been used. In the first position, we find a format related to EEGLAB[41], a well-known MATLAB tool for brain signal processing. The second position is for European Data Format (EDF) a standard for storing multichannel biological and physiological signals, (Kemp and Olivan 2003). The rest of the formats are widely distributed.

VI. CONCLUSIONS AND FUTURE WORKS

This work provides a compilation of open EEG datasets from papers that apply deep learning models. We have used PRISMA to define a workflow for selecting a set of papers that uses these kinds of datasets. Our initial search returned 331 works which, after screening based on the inclusion/exclusion criteria, were reduced to 37. In these papers, 30 datasets were found. Some clear conclusions related to DL techniques are obtained: convolutional neural networks are widely used due to their link with the nature of the data, MI classification is the most common use case and accuracy is the most used metric, but others are more stable. By combining the first and second conclusions, we know that most of the papers apply CNNs to MI use cases. The conclusions related to the datasets comprise: EGGs are difficult to compile due to the low number of instances in general, the number of channels is not relevant so it should be studied, the most used electrode system is the 10-20 system, most relevant sample frequency should also be analyzed and EDF and MAT file formats stand out from the rest.

The further analysis concludes that the number of published papers per year is remarkable, but it is still worth working in the field. From 2018 to 2021, the amount has increased. But in the last 2 years have decreased a little. So, publishing open datasets is relevant for the scientific community. Related to the DL models, we can see that CNNs are a good solution which is why they have been widely applied. The graphics of the use cases are helpful to find application fields that have not been covered a lot or knowing which kind of datasets can obtain good results. The bubble diagram can be used by researchers to know which DL models should be involved in their datasets depending on the use case. In this way, there are several use cases not very exploited, but the use of CNN is not innovative in any case. The analysis of the dataset's characteristics leads us to conclude that the 10-20 system is the most widely used when collecting the data. No work supports the idea that this is the most efficient one. The sample rate of the datasets is very diverse; therefore, none is a priori better than the other. In the case of the number of channels and sample frequency, values are very distributed and again there are no works supporting which values should be recommended. The main limitation of the study is the amount of selected works because there are not several papers accomplishing the criteria. As EEGs are medical data, people are reluctant to make them freely available, and researchers who compile the EEGs do not want to share them since they prefer to exploit them themselves. Another reason is the difficulty of collecting a good quality bank of EEGs as it is costly in terms of time. Another limitation of the work is that the authors of the papers using the datasets are not the same as those who have published them. This condition supposes a decoupling between the medical and computer science perspectives, not considering that both profiles are necessary. Finally, we have found some papers that do not specify that they are using an open dataset, so this criteria for selecting papers could not have affect in some cases.

Some niches to consider are the following. The use of Natural Language Processing (NLP) techniques such as Transformers and GCNN for not being so exploited. NLP models are one of the most advanced nowadays. If we make a parallelism between texts and EEGs, a sentence can be considered a channel, and a word in the sentence is a particular measure of the channel. This approach could be a starting point for applying these powerful models with EEGs. Another exciting application is studying the network connectivity that can be modeled by representing the EEGs as graphs. In this case, GCNNs are very useful and seem to be a niche.

As future work, the review shows that there is room for finding a gold standard of the characteristics of an EEG

[41] https://sccn.ucsd.edu/eeglab/index.php

dataset to be used in multidisciplinary teams of physicians and computer scientists because sometimes the needs of some do not match those of others. Only one work has been found that studies a single characteristic of the datasets, the number of subjects, (Roy et al., 2019). Thus, we propose to carry out different studies in the future to discover how the electrode system, the number of channels, or the sample rate influence obtaining good results when using DL models.

REFERENCES

Abdelhameed, Ahmed, and Magdy Bayoumi. 2021. “A Deep Learning Approach for Automatic Seizure Detection in Children with Epilepsy.” *Frontiers in Computational Neuroscience* 15: 29.

Alhussein, Musaed, Ghulam Muhammad, M Shamim Hossain, and Syed Umar Amin. 2018. “Cognitive IoT-Cloud Integration for Smart Healthcare: Case Study for Epileptic Seizure Detection and Monitoring.” *Mobile Networks and Applications* 23: 1624–35.

Andrzejak, Ralph G et al. 2001. “Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity: Dependence on Recording Region and Brain State.” *Physical Review E* 64(6): 61907.

Andrzejak, Ralph G, Kaspar Schindler, and Christian Rummel. 2012. “Nonrandomness, Nonlinear Dependence, and Nonstationarity of Electroencephalographic Recordings from Epilepsy Patients.” *Physical Review E* 86(4): 46206.

Ang, Kai Keng et al. 2012. “Filter Bank Common Spatial Pattern Algorithm on BCI Competition IV Datasets 2a and 2b.” *Frontiers in neuroscience* 6: 39.

Association, Management, and others. 2013. *Assistive Technologies: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications*. IGI Global.

von Atzingen, Gustavo Voltani et al. 2022. “The Convolutional Neural Network as a Tool to Classify Electroencephalography Data Resulting from the Consumption of Juice Sweetened with Caloric or Non-Caloric Sweeteners.” *Frontiers in Nutrition* 9.

Bairagi, Ramendra Nath, Md Maniruzzaman, Suriya Pervin, and Alok Sarker. 2021. “Epileptic Seizure Identification in EEG Signals Using DWT, ANN and Sequential Window Algorithm.” *Soft Computing Letters* 3: 100026.

Ballard, Dana H. 1987. “Modular Learning in Neural Networks.” In *AAAI*, , 279–84.

Banville, Hubert et al. 2021. “Uncovering the Structure of Clinical EEG Signals with Self-Supervised Learning.” *Journal of Neural Engineering* 18(4): 46020.

Barrett, Paul et al. 2005. “Matplotlib--A Portable Python Plotting Package.” In *Astronomical Data Analysis Software and Systems XIV*, , 91.

Barua, Arnab, Mobyen Uddin Ahmed, and Shahina Begum. 2023. “A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions.” *IEEE Access*.

Baser, Oguzhan et al. 2022. “Automatic Detection of the Spike-and-Wave Discharges in Absence Epilepsy for Humans and Rats Using Deep Learning.” *Biomedical Signal Processing and Control* 76: 103726.

Bassi, Pedro R A S, Willian Rampazzo, and Romis Attux. 2021. “Transfer Learning and SpecAugment Applied to SSVEP Based BCI Classification.” *Biomedical Signal Processing and Control* 67: 102542.

Blankertz, Benjamin et al. 2004. “The BCI Competition 2003: Progress and Perspectives in Detection and Discrimination of EEG Single Trials.” *IEEE transactions on biomedical engineering* 51(6): 1044–51.

Brunner, C et al. 2008. “Graz Data Set A, Provided by the Institute for Knowledge Discovery (Laboratory of Brain--Computer Interfaces).” *Graz University of Technology*.

Bujang, Siti Dianah Abdul et al. 2022. “Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review.” *IEEE Access*.

Cai, Hanshu et al. 2020. “Modma Dataset: A Multi-Modal Open Dataset for Mental-Disorder Analysis.” *arXiv preprint arXiv:2002.09283*.

Chang, Hsun-Hsien, and José M F Moura. 2010. “Biomedical Signal Processing.” *Biomedical engineering and design handbook* 2: 559–79.

Chen, J X, D M Jiang, and Y N Zhang. 2019. “A Hierarchical Bidirectional GRU Model with Attention for EEG-Based Emotion Classification.” *IEEE Access* 7: 118530–40.

Cho, Hohyun et al. 2017. “EEG Datasets for Motor Imagery Brain--Computer Interface.” *GigaScience* 6(7): gix034.

Coretto, Germán A Pressel, Iván E Gareis, and H Leonardo Rufiner. 2017. “Open Access Database of EEG Signals Recorded during Imagined Speech.” In *12th International Symposium on Medical Information Processing and Analysis*, , 1016002.

Cui, Jian et al. 2021. “A Compact and Interpretable Convolutional Neural Network for Cross-Subject Driver Drowsiness Detection from Single-Channel EEG.” *Methods*.

Dar, Muhammad Najam et al. 2022. “EEG-Based Emotion Charting for Parkinson’s Disease Patients Using Convolutional Recurrent Neural Networks and Cross Dataset Learning.” *Computers in biology and medicine* 144: 105327.

Eldele, Emadeldeen et al. 2021. “An Attention-Based Deep Learning Approach for Sleep Stage Classification With Single-Channel EEG.” *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29: 809–18.

Elman, Jeffrey L. 1990. “Finding Structure in Time.” *Cognitive science* 14(2): 179–211.

Ghassemi, Mohammad M et al. 2018. “You Snooze, You Win: The Physionet/Computing in Cardiology Challenge 2018.” In *2018 Computing in Cardiology Conference (CinC)*, , 1–4.

Goldberger, Ary L et al. 2000. “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.” *circulation* 101(23): e215--e220.

Goodfellow, Ian et al. 2014. “Generative Adversarial Nets.” *Advances in neural information processing systems* 27.

Guillot, Antoine, Fabien Sauvet, Emmanuel H During, and Valentin Thorey. 2020. “Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging.” *IEEE transactions on neural systems and rehabilitation engineering* 28(9): 1955–65.

Guillot, Antoine, and Valentin Thorey. 2021. “RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale.” *arXiv preprint arXiv:2101.02452*.

Hecht-Nielsen, Robert. 1988. “Neurocomputing: Picking the Human Brain.” *IEEE spectrum* 25(3): 36–41.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural computation* 9(8): 1735–80.

Huang, Xinyu, Kimiaki Shirahama, Frédéric Li, and Marcin Grzegorzek. 2020. “Sleep Stage Classification for Child Patients Using DeConvolutional Neural Network.” *Artificial Intelligence in Medicine* 110: 101981.

Islam, Md Rabiul et al. 2021. “EEG Channel Correlation Based Model for Emotion Recognition.” *Computers in Biology and Medicine* 136: 104757.

Jasper, Herbert H. 1958. “The Ten-Twenty Electrode System of the International Federation.” *Electroencephalogr. Clin. Neurophysiol.* 10: 370–75.

Juárez-Guerra, Ever et al. 2020. “A New Wavelet-Based Neural Network for Classification of Epileptic-Related States Using EEG.” *Journal of Signal Processing Systems* 92: 187–211.

Keele, Staffs, and others. 2007. “Guidelines for Performing Systematic Literature Reviews in Software Engineering.”

Kemp, Bob, and Jesus Olivan. 2003. “European Data Format ‘plus’(EDF+), an EDF Alike Standard Format for the Exchange of Physiological Data.” *Clinical neurophysiology* 114(9): 1755–61.

Kipf, Thomas N, and Max Welling. 2016. “Semi-Supervised Classification with Graph Convolutional Networks.” *arXiv preprint arXiv:1609.02907*.

Koelstra, Sander et al. 2011. “Deap: A Database for Emotion Analysis; Using Physiological Signals.” *IEEE transactions on affective computing* 3(1): 18–31.

Korkalainen, Henri et al. 2019. “Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea.” *IEEE journal of biomedical and health informatics* 24(7): 2073–81.
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “Imagenet Classification with Deep Convolutional Neural Networks.” *Advances in neural information processing systems* 25: 1097–1105.
Kurz, Stefan et al. 2022. “Hybrid Modeling: Towards the next Level of Scientific Computing in Engineering.” *Journal of Mathematics in Industry* 12(1): 1–12.
Kwak, No-Sang, and Seong-Whan Lee. 2019. “Error Correction Regression Framework for Enhancing the Decoding Accuracies of Ear-EEG Brain--Computer Interfaces.” *IEEE transactions on cybernetics* 50(8): 3654–67.
Kwon, Yea-Hoon, Sae-Byuk Shin, and Shin-Dug Kim. 2018. “Electroencephalography Based Fusion Two-Dimensional (2D)-Convolution Neural Networks (CNN) Model for Emotion Recognition System.” *Sensors* 18(5): 1383.
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *nature* 521(7553): 436–44.
Li, Fan et al. 2021. “End-to-End Sleep Staging Using Convolutional Neural Network in Raw Single-Channel EEG.” *Biomedical Signal Processing and Control* 63: 102203.
Li, Yang, Xin Zhang, and Dong Ming. 2023. “Early-Stage Fusion of EEG and FNIRS Improves Classification of Motor Imagery.” *Multimodal Brain Image Fusion: Methods, Evaluations, and Applications* 16648714: 149.
Lichtenauer, Jeroen, Jie Shen, Michel Valstar, and Maja Pantic. 2011. “Cost-Effective Solution to Synchronised Audio-Visual Data Capture Using Multiple Sensors.” *Image and Vision Computing* 29(10): 666–80.
Lin, Yuan-Pin et al. 2007. “Multilayer Perceptron for EEG Signal Classification during Listening to Emotional Music.” In *TENCON 2007-2007 IEEE Region 10 Conference*, , 1–3.
Liu, Junxiu et al. 2020. “EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder.” *Frontiers in Systems Neuroscience* 14: 43.
Liu, Yisi et al. 2020. “Inter-Subject Transfer Learning for EEG-Based Mental Fatigue Recognition.” *Advanced Engineering Informatics* 46: 101157.
Lu, Na, Tengfei Li, Xiaodong Ren, and Hongyu Miao. 2016. “A Deep Learning Scheme for Motor Imagery Classification Based on Restricted Boltzmann Machines.” *IEEE transactions on neural systems and rehabilitation engineering* 25(6): 566–76.
Luo, Yun, and Bao-Liang Lu. 2018. “EEG Data Augmentation for Emotion Recognition Using a Conditional Wasserstein GAN.” In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, , 2535–38.
Ma, Mengnan et al. 2021. “Research on Epileptic EEG Recognition Based on Improved Residual Networks of 1-D CNN and IndRNN.” *BMC Medical Informatics and Decision Making* 21: 1–13.
Maeng, Jun-Ho, Dong-Hyun Kang, and Deok-Hwan Kim. 2020. “Deep Learning Method for Selecting Effective Models and Feature Groups in Emotion Recognition Using an Asian Multimodal Database.” *Electronics* 9(12): 1988.
Marican, Mohamed Noordin Yusuff, Shukor Abd Razak, Ali Selamat, and Siti Hajar Othman. 2022. “Cyber Security Maturity Assessment Framework for Technology Startups: A Systematic Literature Review.” *IEEE Access*.
Miao, Minmin, Wenjun Hu, Hongwei Yin, and Ke Zhang. 2020. “Spatial-Frequency Feature Learning and Classification of Motor Imagery EEG Based on Deep Convolution Neural Network.” *Computational and mathematical methods in medicine* 2020.
Min, Jianliang, Ping Wang, and Jianfeng Hu. 2017. “Driver Fatigue Detection through Multiple Entropy Fusion Analysis in an EEG-Based System.” *PLoS one* 12(12): e0188756.
Moher, David et al. 2009. “Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement.” *PLoS medicine* 6(7): e1000097.
Montoya-Mart\inez, J, A Bertrand, and T Francart. 2019. “Optimal Number and Placement of Eeg Electrodes for Measurement of Neural Tracking of Speech.” *bioRxiv*.
Mortaz, Ebrahim. 2020. “Imbalance Accuracy Metric for Model Selection in Multi-Class Imbalance Classification Problems.” *Knowledge-Based Systems* 210: 106490.
Normandeau, Kevin. 2013. “Beyond Volume, Variety and Velocity Is the Issue of Big Data Veracity.” *Inside big data*.
Obeid, Iyad, and Joseph Picone. 2016. “The Temple University Hospital EEG Data Corpus.” *Frontiers in neuroscience* 10: 196.
Pedoeem, J, S Abittan, G Bar Yosef, and S Keene. 2020. “TABS: Transformer Based Seizure Detection.” In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, , 1–6.
Podmore, Joshua J, Toby P Breckon, Nik K N Aznan, and Jason D Connolly. 2019. “On the Relative Contribution of Deep Convolutional Neural Networks for SSVEP-Based Bio-Signal Decoding in BCI Speller Applications.” *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27(4): 611–18.
Qiu, Yang, Weidong Zhou, Nana Yu, and Peidong Du. 2018. “Denoising Sparse Autoencoder-Based Ictal EEG Classification.” *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26(9): 1717–26.
Quan, Stuart F et al. 1997. “The Sleep Heart Health Study: Design, Rationale, and Methods.” *Sleep* 20(12): 1077–85.
Roy, Yannick et al. 2019. “Deep Learning-Based Electroencephalography Analysis: A Systematic Review.” *Journal of neural engineering* 16(5): 51001.
Ruffini, Giulio et al. 2016. “EEG-Driven RNN Classification for Prognosis of Neurodegeneration in at-Risk Patients.” In *International Conference on Artificial Neural Networks*, , 306–13.
Russell, Stuart J, and Peter Norvig. 2016. “Artificial Intelligence: A Modern Approach. Malaysia.”
Salakhutdinov, Ruslan, and Geoffrey Hinton. 2009. “Deep Boltzmann Machines.” In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, eds. David van Dyk and Max Welling. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, 448–55. https://proceedings.mlr.press/v5/salakhutdinov09a.html.
Salehzadeh, Amirsaleh, Andre P Calitz, and Jean Greyling. 2020. “Human Activity Recognition Using Deep Electroencephalography Learning.” *Biomedical Signal Processing and Control* 62: 102094.
Samuel, Arthur L. 1959. “Some Studies in Machine Learning Using the Game of Checkers.” *IBM Journal of research and development* 3(3): 210–29.
San-Segundo, Rubén, Manuel Gil-Mart\'\in, Luis Fernando D'Haro-Enr\'\iquez, and José Manuel Pardo. 2019. “Classification of Epileptic EEG Recordings Using Signal Transforms and Convolutional Neural Networks.” *Computers in biology and medicine* 109: 148–58.
Sarmiento, Luis Carlos et al. 2021. “Recognition of EEG Signals from Imagined Vowels Using Deep Learning Methods.” *Sensors* 21(19): 6503.
Sathya, Ramadass, Annamma Abraham, and others. 2013. “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification.” *International Journal of Advanced Research in Artificial Intelligence* 2(2): 34–38.
Schirrmeister, Robin Tibor et al. 2017. “Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization.” *Human brain mapping* 38(11): 5391–5420.
Shalash, W M. 2021. “A Deep Learning Cnn Model for Driver Fatigue Detection Using Single EEG Channel.” *Journal of Theoretical and Applied Information Technology* 99(2): 462–77.
Shin, Jaeyoung et al. 2016. “Open Access Dataset for EEG+ NIRS Single-Trial Classification.” *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25(10): 1735–45.
Shoeb, Ali Hossam. 2009. “Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment.” Massachusetts Institute of Technology.
Smolensky, Paul. 1985. “Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory.” *Parallel distributed processing: explorations in the microstructure of cognition* 1.
Song, Tengfei, Wenming Zheng, Peng Song, and Zhen Cui. 2018. “EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks.” *IEEE Transactions on Affective Computing*

11(3): 532–41.
Terzano, Mario Giovanni et al. 2002. “Atlas, Rules, and Recording Techniques for the Scoring of Cyclic Alternating Pattern (CAP) in Human Sleep.” *Sleep medicine* 3(2): 187–99.
Thongchotchat, Vivat, Yasuo Kudo, Yoshifumi Okada, and Kazuhiko Sato. 2023. “Educational Recommendation System Utilizing Learning Styles: A Systematic Literature Review.” *IEEE Access*.
Topic, Ante, and Mladen Russo. 2021. “Emotion Recognition Based on EEG Feature Maps through Deep Learning Network.” *Engineering Science and Technology, an International Journal* 24(6): 1442–54.
Velicer, Wayne F, and Peter C Molenaar. 2013. “Time Series Analysis for Psychological Research.”
Völske, Michael et al. 2021. “Web Archive Analytics.” *INFORMATIK 2020*.
Wang, Dan, and Yi Shang. 2013. “Modeling Physiological Data with Deep Belief Networks.” *International journal of information and education technology (IJIET)* 3(5): 505.
Wang, Yijun, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. 2016. “A Benchmark Dataset for SSVEP-Based Brain--Computer Interfaces.” *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25(10): 1746–52.
Wen, Tingxi et al. 2021. “A Deep Learning-Based Classification Method for Different Frequency EEG Data.” *Computational and Mathematical Methods in Medicine* 2021.
Wu, Hao et al. 2019. “A Parallel Multiscale Filter Bank Convolutional Neural Networks for Motor Imagery EEG Classification.” *Frontiers in neuroscience* 13: 1275.
Xu, Gaowei et al. 2019. “A Deep Transfer Convolutional Neural Network Framework for EEG Signal Classification.” *IEEE Access* 7: 112767–76.
Xu, Lichao et al. 2020. “Cross-Dataset Variability Problem in EEG Decoding with Deep Learning.” *Frontiers in human neuroscience* 14: 103.
Yan, Rui et al. 2021. “Automatic Sleep Scoring: A Deep Learning Architecture for Multi-Modality Time Series.” *Journal of Neuroscience Methods* 348: 108971.
Yang, Su, and Farzin Deravi. 2017. “On the Usability of Electroencephalographic Signals for Biometric Recognition: A Survey.” *IEEE Transactions on Human-Machine Systems* 47(6): 958–69.
Yi, Weibo et al. 2014. “Evaluation of EEG Oscillatory Patterns and Cognitive Process during Simple and Compound Limb Motor Imagery.” *PloS one* 9(12): e114853.
Zhang, Guo-Qiang et al. 2018. “The National Sleep Research Resource: Towards a Sleep Data Commons.” *Journal of the American Medical Informatics Association* 25(10): 1351–58.
Zhang, Jiao, Baomin Xu, and Hongfeng Yin. 2023. “Depression Screening Using Hybrid Neural Network.” *Multimedia Tools and Applications*: 1–16.
Zhang, Yangsong et al. 2021. “An End-to-End 3D Convolutional Neural Network for Decoding Attentive Mental State.” *Neural Networks* 144: 129–37.
Zhao, Yuxuan et al. 2019. “Multimodal Emotion Recognition Model Using Physiological Signals.” *arXiv preprint arXiv:1911.12918*.
Zhou, Bangyan et al. 2016. “A Fully Automated Trial Selection Method for Optimization of Motor Imagery Based Brain-Computer Interface.” *PloS one* 11(9): e0162657.
Zhou, Mengni et al. 2018. “Epileptic Seizure Detection Based on EEG Signals and CNN.” *Frontiers in neuroinformatics* 12: 95.
Zhu, Yuanlu, Ying Li, Jinling Lu, and Pengcheng Li. 2021. “EEGNet with Ensemble Learning to Improve the Cross-Session Classification of SSVEP Based BCI from Ear-EEG.” *IEEE Access* 9: 15295–303.

A. NOGALES Alberto Nogales is a Computer Scientist with a Master's in Software Engineering and Artificial Intelligence. He earns a Ph.D. at the University of Alcalá (Spain) focused on the field of Semantic Web and Social Network Analysis. He has experience in several European research projects and publications in conferences and journals from JCR. His education has been completed working at Technische Universität Wien (Austria), Tallinna Tehnikaülikool (Estonia), and the University of Málaga (Spain). Since November 2017 works as a post-doctoral researcher and lecturer at Universidad Francisco de Victoria, at the CEIEC research center. Is working in Deep Learning and its use in signal processing, predictive models, and computer vision in medical applications.

A.J. GARCÍA-TEJEDOR Ph.D. in Biochemistry. He has been Lecturer in UCM, UC3M and Universidad Francisco de Vitoria, where he is now Associate Professor in Artificial Intelligence. He has been involved in more than 20 R&D projects mainly in AI. As a researcher, he started doing mathematical modelling of biological systems. Since 2007 he heads CEIEC, a Research Institute of the UFV for technological innovation with a social accent. He has produced several serious games for transmission of educational and cultural content and integration of people with Intellectual Disabilities. His current work is focused on neural models, deep learning techniques and their applications in several fields.

TABLE I
SUMMARY OF SELECTED PAPERS

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3293421

Paper	Preprocessing techniques	Deep learning model	Metrics	Aim	Year	CI	SJR (2021)	SNIP (2022)	CQI
(G. Xu et al. 2019)		CNN	74.2% Accuracy	Motor imagery (MI) electroencephalogram (EEG) signal classification	2019	98	0.927	1.33	0.891345
(Wu et al. 2019)		CNN	75.8 and 84.3 Accuracy	MI classification	2019	74	1.275	1.32	0.992178
(Schirrmeister et al. 2017)		CNN	91.15% Sensitivity	Classification of imagined or executed movements	2018	1157	1.719	1.55	2.169724
(Kwon, Shin, and Kim 2018)		CNN	73.4% Accuracy	Classify emotion based on multimodal data	2018	110	0.803	1.42	0.889157
(D. Wang and Shang 2013)		MLP	60.9%, 51.2%, and 68.4% Accuracy	Predict the levels of arousal, valance, and liking based on the learned features	2013	190	1.257	1.64	1.199843
(Maeng, Kang, and Kim 2020)	Frequency filtering and normalization	RNN	91.3% and 94.8% Accuracy	Recognize emotions	2020	8	0.59	1.01	0.578499
(Podmore et al. 2019)		CNN	86% and 77% Accuracy	Extract stimulus pattern features	2019	34	1.257	1.64	1.065012
(Pedoeem et al. 2020)		AU	12.37% Sensitivity	Predict seizures	2020	5	0.243	0.00	0.091700
(Cui et al. 2021)	FIR filters and artifact removal	CNN	73.22 Accuracy	Detect drivers' drowsy states	2021	12	1.392	0.89	0.827637
(L. Xu et al. 2020)		CNN	71%, 72%, 70% and 72% Accuracy	MI classification	2020	48	0.859	1.25	0.795208
(Miao et al. 2020)		CNN	90% Accuracy	Classification of motor imagery EEG	2020	24	0.522	0.96	0.550083
(Abdelhameed and Bayoumi 2021)	STFT	AU+LSTM	98.79 Accuracy, 98.72 Sensitivity, 98.86 Specificity	Detecting seizures in pediatric patients	2021	27	0.797	1.11	0.704942
(Y. Liu et al. 2020)		CNN	73.01% Accuracy and 68% Accuracy	Mental fatigue recognition	2020	31	1.601	2.20	1.385404
(Shalash 2021)	Bandpass filtering	CNN	94.33%, 92.57 and 93% Accuracy	Detect drivers' fatigue	2021	2	0.195	0.39	0.210637
(Korkalainen et al. 2019)		CNN+LSTM	83.9 and 83.7 Accuracy	Estimation of the sleep stages	2019	60	1.799	2.27	1.506577
(Sarmiento et al. 2021)		CNN	65.62% and 85.66% Accuracy	Recognize EEG signals in imagined vowel tasks	2021	4	0.803	1.42	0.797540
(Bassi, Rampazzo, and Attux 2021)		CNN	82.2% Accuracy and 82.5% F1-Score	BCI Classification	2021	9	1.211	1.86	1.105155
(Guillot and Thorey 2021)		AU	97% F1-Score	Sleep stage classification	2021	25	1.257	1.64	1.057233
(F. Li et al. 2021)		CNN	66.5% Sensitivity, 97.9% Specificity and 67.9% Sensitivity, 97.0% Specificity	Classify sleep staging	2019	20	1.211	1.86	1.114663
(Banville et al. 2021)		CNN	72.3% and 79.4% Accuracy	EEG-based sleep staging and pathology detection	2020	35	1.504	1.66	1.161811
(Yan et al. 2021)		CNN+LSTM	87%, 86% and 86% Accuracy	Automatic Sleep Scoring	2020	9	0.754	0.88	0.592071
(Eldele et al. 2021)		CNN	84.4%, 81.3% and 86.7% Accuracy	Sleep stage classification	2021	89	1.257	1.64	1.112549
(Huang et al. 2020)		CNN	90.89% Accuracy	Sleep Stage Classification	2020	11	1.497	2.24	1.344956
(J. Liu et al. 2020)		AU	89.49%, 92.86% and 96.77% Accuracy	EEG-Based Emotion Classification	2020	74	1.223	1.02	0.866718
(San-Segundo et al. 2019)		CNN	99.5%, 96.5% and 95.7% Accuracy	Classification of epileptic EEG recordings	2019	90	1.309	1.94	1.238873
(Islam et al. 2021)		CNN	78.22% and 74.92% Accuracy	Emotion Recognition	2021	39	1.309	1.94	1.194793
(Y. Zhang et al. 2021)		CNN	70.15% Accuracy, 70.18 F1-Score and 77.07% Accuracy, and 75.48% F1-Score	Detection Attention Deficit and Hyperactivity Disorder (ADHD)	2021	11	2.781	2.81	2.009507
(Baser et al. 2022)		CNN	99.42%, 95.83% Accuracy. 99.55%, 96.29% Specificity. 97.55%, 89.57% Sensitivity 96.39%, 81.86% f1-score	Seizure Management	2022	1	2.781	2.81	2.000864
(J. Zhang, Xu, and Yin 2023)		MLP	93.33% Accuracy and 95.0% Accuracy	Seizure Management	2019	11	2.781	2.81	2.009507
(Alhussein et al. 2018)		CNN, CNN+AU	99.2% Accuracy and 93.5% Sensitivity	Seizure Management	2018	95	0.871	1.37	0.882850

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3293421

TABLE II
SUMMARY OF SELECTED DATASETS

Dataset	Number of subjects	Total tests	Length per-test	Electrodes' system	Nº channels	Sampling frequency	Format	Papers
BCI Competition IV 2a	9	2,591	5 minutes	10-20 system	22 channels	250 Hz	GDF	3
BCI Competition IV 2b	9	45	5 minutes	No system	3 channels	250 Hz	GDF	3
DEAP and video signals	32	32	40 minutes	10-20 system	45 channels	512 Hz	BDF	5
Multi-channel EEG sustained attention driving task	27	62	90 minutes	10-20 system	32 channels	500 Hz	SET	2
Temple University EEG Corpus	16,986	10,874	20 minutes	10-20 system	31 channels	250 Hz (87%) 256 Hz (8.3%) 400 Hz (3.8%) 512 Hz (1%).	EDF	3
CHB-MIT Scalp EEG Database	22	664	1 to 4 hours	10-20 system	23 channels	256 Hz	EDF	3
MAHNOB-HCI	30	120	30 seconds	10-20 system	32 channels	256 Hz	BDF	1
Sleep EDF	78	197	9 hours	No system	2 channels	100 Hz	EDF	6
Weibo et al., 2014	10	320	8 seconds	10-20 system	60 channels	100 Hz	TXT	1
PhysioNet/CinC Challenge 2018	1,985	1,985	7.7 hours average	10-20 system	6 channels	200 Hz	MAT	2
Open source SSVEP dataset	35	35	4 minutes	10-20 system	64 channels	1000 Hz	MAT	2
BCI Competition **III** IVa	5	10	980 seconds	10-20 system	118 channels	1000 Hz	MAT	1
Driver fatigue detection	12	12	5 minutes	10-20 system	40 channels	1000 Hz	CNT	1
University of Bonn	5	500	23.6 seconds	10-20 system	Single-channel	173.61 Hz	TXT	2
Zhou et al., 2016	4	12	750 seconds	10-20 system	14 channels	250 Hz	CNT	1
Sleep Heart Health Study	3,295	2,651	About 8 hours	No system	2 channels	125 Hz	EDF	4
DOD-O	55	55	387 minutes	No system	8 channels	250 Hz	H5	1
DOD-H	25	25	427 minutes	No system	8 channels	250 Hz	H5	1
CAP Sleep Database	108	108	410 minutes	10-20 system	3 channels	From 128 to 512 Hz	EDF	1
Bern-Barcelona EEG database	5	3,740	20 seconds	10-20 system	64 channels	512 or 1024 Hz	TXT	1
MrOS Sleep	1,026	586	341 minutes	10-20 system	5 channels	256 Hz	EDF	1
Database-Imaged-Vowels	15	15	110 seconds	10-20 system	18 channels	1024 Hz	MAT	1
CBCI2019	18 and 6	60 and 40	4 seconds	No system	59 channels	250 and 1000 Hz	MAT	2
Deep BCI	11	33	1200 Seconds	10-20 system	8 channels	500 Hz	MAT	1
EEG Sweetners AI	11	33	16 seconds	10-20 system	2 channels	512 Hz	EDF	1
BehaveNET	8	8	300-380 seconds	No system	4 channels	220 Hz	CSV	1
EEG+NIRS Single-Trial	29	174	147 seconds	10-5 system	30 channels	128 Hz	MAT	2
MODMA Dataset	48	48	5 minutes	10-20 system	128 channels	250 Hz	MFF	1
AMIGOS	77	77	Variable	10-20 system	62 channels	200 Hz	MAT	1

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3293421

SEED	15	15	4575 seconds	10-20 system	17 channels	256 Hz	MAT	2
MESA	6814	6814	6-8 hours	No system	5 channels	256 Hz	CSV	1
EEG datasets for motor imagery brain computer interface	52	52	51 minutes	10-10 system	64 channels	512 Hz	MAT	1