frontiers | Frontiers in Neuroscience

TYPE Brief Research Report
PUBLISHED 03 May 2024
DOI 10.3389/fnins.2024.1373515

OPEN ACCESS

EDITED BY
Jürgen Dammers,
Helmholtz Association of German Research Centres (HZ), Germany

REVIEWED BY
Philipp Lohmann,
Research Center Juelich, Germany
Yongxia Zhou,
University of Southern California,
United States

*CORRESPONDENCE
Geoffrey Brookshire
✉ geoff.brookshire@sparkneuro.com

†PRESENT ADDRESS
Nicholas M. Blauch,
Harvard University, Cambridge, MA,
United States

‡These authors have contributed equally to this work and share first authorship

RECEIVED 19 January 2024
ACCEPTED 04 April 2024
PUBLISHED 03 May 2024

CITATION
Brookshire G, Kasper J, Blauch NM, Wu YC, Glatt R, Merrill DA, Gerrol S, Yoder KJ, Quirk C and Lucero C (2024) Data leakage in deep learning studies of translational EEG. *Front. Neurosci.* 18:1373515. doi: 10.3389/fnins.2024.1373515

Data leakage in deep learning studies of translational EEG

Geoffrey Brookshire[1*‡], Jake Kasper[1‡], Nicholas M. Blauch[1,2†], Yunan Charles Wu[1], Ryan Glatt[3], David A. Merrill[3,4,5], Spencer Gerrol[1], Keith J. Yoder[1], Colin Quirk[1] and Ché Lucero[1]

[1]SPARK Neuro Inc., New York, NY, United States, [2]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, United States, [3]Pacific Brain Health Center, Pacific Neuroscience Institute and Foundation, Santa Monica, CA, United States, [4]Saint John's Cancer Institute at Providence Saint John's Health Center, Santa Monica, CA, United States, [5]Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine at University of California, Los Angeles, Los Angeles, CA, United States

A growing number of studies apply deep neural networks (DNNs) to recordings of human electroencephalography (EEG) to identify a range of disorders. In many studies, EEG recordings are split into segments, and each segment is randomly assigned to the training or test set. As a consequence, data from individual subjects appears in both the training and the test set. Could high test-set accuracy reflect data leakage from subject-specific patterns in the data, rather than patterns that identify a disease? We address this question by testing the performance of DNN classifiers using segment-based holdout (in which segments from one subject can appear in both the training and test set), and comparing this to their performance using subject-based holdout (where all segments from one subject appear exclusively in either the training set or the test set). In two datasets (one classifying Alzheimer's disease, and the other classifying epileptic seizures), we find that performance on previously-unseen subjects is strongly overestimated when models are trained using segment-based holdout. Finally, we survey the literature and find that the majority of translational DNN-EEG studies use segment-based holdout. Most published DNN-EEG studies may dramatically overestimate their classification performance on new subjects.

KEYWORDS

electroencephalography, deep neural networks, data leakage, cross-validation, Alzheimer's disease, epilepsy

1 Introduction

Translational neuroscience studies increasingly turn to deep neural network (DNN) models to find structure in neural data. The power of DNN models comes from their ability to discover patterns in the data that researchers would not have been able to specify. DNN classifiers have the potential to revolutionize medical care by increasing the speed, accuracy, and availability of diagnosis (Mall et al., 2023). DNNs have been trained on a variety of imaging techniques to identify a wide range of clinical conditions. Many of these studies use DNNs to diagnose diseases based on anatomical neuroimaging. For example, DNN models can identify Alzheimer's disease (AD) using structural magnetic resonance imaging (MRI) (Wen et al., 2020), and a variety of cancers and brain injuries using CT scans (Hosny et al., 2018; Kaka et al., 2021). In addition to anatomical data, a large number of studies have used DNNs to identify diseases from functional neuroimaging data. For example, DNNs with functional MRI show promise for identifying AD, Autism spectrum disorders, attention-deficit/hyperactivity disorder (ADHD), and schizophrenia (Wen et al., 2018). Furthermore, DNNs have been used with electroencephalography (EEG) to study a variety of different neural and cognitive disorders (de Bardeci et al., 2021).

Deep learning helps to reveal previously-unknown patterns in neuroimaging data, but it also presents researchers with subtle pitfalls. One set of challenges concerns how the data are split into separate training and test sets. The training set is used to fit the model's parameters, and the test set is used to estimate the model's performance on new data (a third subset of the data is often held aside as a validation set, used to tune the model's hyperparameters and to determine when to stop training the model). In some cases, researchers train their model on one subset of the available data, and then evaluate the model's performance on a separate test set. In other cases, researchers use cross-validation (CV) to train and test models on multiple subsets of the data. Under both of these approaches, researchers must be careful to avoid "data leakage" when splitting the data into training and test sets. Data leakage, which arises when information about the test set is present in the training set, results in a positively-biased estimate of the model's performance (Kaufman et al., 2012). For example, in a data-mining competition focused on identifying patients with breast cancer, one team of researchers found that the patient ID number carried predictive information about cancer risk (Rosset et al., 2010). These ID numbers may have appeared after compiling data from different medical institutions. Because the ID number was assigned based on patients' diagnosis, it constitutes a source of data leakage (Rosset et al., 2010). In general, data leakage occurs when an experimenter handles the data in a way that artificially introduces correlations between the training and test sets.

DNN models typically require a large amount of training data to perform well, but neural datasets are usually expensive and difficult to obtain. To increase the number of observations available to train the model, these studies often split a single neural recording into multiple samples, and use each sample as a separate observation during training or testing. For example, a 3D structural MR volume could be split into multiple 2D slices, and an fMRI time-series could be split into multiple segments of time (Wen et al., 2020). When multiple observations from a single subject are included in both the training and test sets, it constitutes data leakage: Instead of learning a generalizable pattern, these models could learn characteristics of the individual subjects in the training set, and then simply recognize those familiar subjects in the test set. As a result, these models perform well in the study's test set, leading the researchers to believe they have a robust classifier. In new subjects, however, the model may fail to generalize.

Prior research has shown that leakage of subject-specific information—sometimes referred to as "identity confounding" (Chaibub Neto et al., 2019)—occurs in a number of different research areas. For example, this type of data-leakage occurs in published MRI studies (Wen et al., 2020). Furthermore, leakage of subject-specific information is widespread in translational studies using optical coherence tomography (OCT), and leads to strongly inflated estimates of test accuracy (Tampu et al., 2022). Identity confounding has also been demonstrated in studies that make clinical predictions on the basis of smartphone data, wearable sensor data, and audio voice recordings (Saeb et al., 2017; Tougui et al., 2021).

Studies using DNNs with EEG are particularly susceptible to data leakage. In these studies, each subject's full EEG time-series (lasting several minutes) is commonly divided up into brief segments (lasting several seconds) (de Bardeci et al., 2021). Each segment is then used as a separate observation during training or testing. This segmentation procedure is meant to ensure that DNN models have enough training data to learn robust representations of the patterns that characterize a disease, and to prepare the data for commonly-used model architectures. However, EEG segmentation leads to data leakage if the same subjects appear in both the training and test sets. Segments of EEG from one subject are more similar to each other than to segments from different subjects (Demuru and Fraschini, 2020). Instead of learning an abstract representation that would generalize to new subjects, a DNN model could therefore achieve high classification accuracy by associating a label with each subject's idiosyncratic pattern of brain activity. As a consequence, randomly splitting EEG segments into training and test sets results in data leakage, and a biased estimate of test performance: accuracy is high on the researchers' test set, but the classifier will generalize poorly to new subjects. In a clinical setting, this leads to an apparently-promising diagnostic tool that fails when applied to new patients. To avoid this kind of data leakage, all segments from a given subject must be assigned to only a single partition of the data (i.e., train *or* validation *or* test).

How does leakage of subject-specific information bias the results of translational DNN-EEG studies? Here we address this question by examining the effects of data leakage in two case studies, and then reviewing the published literature to gauge the prevalence of this leakage. In the case studies, we reproduce two convolutional neural network (CNN) architectures used by published studies—both of which used a train-test split that introduced data leakage. In order to focus on the ways in which leakage results from the train-test split, and to facilitate comparison with prior literature, we reuse these published model architectures without any modification. First, we use a CNN to classify subjects as either healthy or as having dementia due to Alzheimer's disease. Second, we use a CNN to classify whether segments of time contain an epileptic seizure. In both datasets, we find that real-world performance is dramatically overestimated when data from individual subjects is included in both the training and test sets. In the literature review, we find that the majority of translational DNN-EEG studies suffer from data leakage due to data from individual subjects appearing in both the training and test sets.

2 Method

2.1 Deep neural network analysis overview

To investigate how segment-based holdout leads to data leakage, we reproduced the model architectures from two published studies (Oh et al., 2020; Rashed-Al-Mahfuz et al., 2021). The goal of these analyses was not to develop an optimal architecture, but rather to evaluate the impact of different cross-validation choices on the estimated model performance. We therefore re-used the published architectures and data processing pipelines without modification, and without any model selection or hyperparameter tuning. The code necessary to reproduce both of these DNN models is provided in the Supplementary material.

2.2 Experiment 1: Alzheimer's disease diagnosis

2.2.1 EEG data

We analyzed EEG data that was collected for a previously published study (Ganapathi et al., 2022). These EEG recordings were provided to us by the Pacific Neuroscience Institute. All procedures were approved by the St. John's Cancer Institute Institutional Review Board (Protocol JWCI-19-1101) in accordance with the Helsinki Declaration of 1975. Patients were evaluated by a dementia specialist as part of their visit to a specialty memory clinic (Pacific Brain Health Center in Santa Monica, CA) for memory complaints. This evaluations included behavioral testing as well as EEG recordings. After these evaluations, subjects were selected by retrospectively reviewing charts for patients aged 55 and older seen between July 2018 and February 2021.

Patients received a consensus diagnosis from a panel of board-certified dementia specialists. Diagnoses were performed using standard clinical methods on the basis of neurological examinations, cognitive testing (MMSE Folstein et al., 1975 or MoCA Nasreddine et al., 2005), clinical history (e.g., hypertension, diabetes, head injury, depression), and laboratory results (e.g., vitamin B-12 levels, thyroid stimulating hormone levels, and rapid plasma regain testing). These tests were used to rule out reversible causes of memory loss and to diagnose subjective cognitive impairment (SCI), mild cognitive impairment (MCI), and dementia. EEG data was not included in the diagnostic process. Cognitive impairment was diagnosed on the basis of MMSE [or MoCA scores converted to MMSE (Bergeron et al., 2017)], with MCI diagnosed according to established criteria (Langa and Levine, 2014). MCI was distinguished from dementia on the basis of preserved independence in functional abilities, and a lack of significant impairment in social or occupational functioning. SCI was diagnosed in patients with subjective complaints but without evidence of MCI. Diagnostic categorization was based on the clinical syndromes (Langa and Levine, 2014), and did not consider disease etiology or subtypes within each stage.

EEG data were recorded at 250 Hz using the eVox System (Evoke Neuroscience), with a cap that included 19 electrodes following the International 10-20 system (FP1, FP2, F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, O1, and O2). The full EEG session included a 5-min block of eyes-open rest, a 5-minute block of eyes-closed rest, and a 15-min go/no-go task. In this study, we analyzed only the eyes-open resting-state data. Recordings were low-pass filtered below 125 Hz, and split into non-overlapping segments of 2 s (500 samples) for model training. Channels were stacked to produce matrices of shape (500, 19) as model inputs.

We selected all 49 subjects in the dataset who were diagnosed with dementia due to Alzheimer's disease (18 male, 31 female; age 73.9 ± 6.8 years). As a comparison, we selected an equal number of subjects with subjective cognitive impairment (SCI; $n = 49$, 18 male, 31 female; age 63.9 ± 11.4 years).

2.2.2 Architecture

Because our goal was to evaluate the effects of different cross-validation strategies on generalizability, we re-used a previously-published model architecture without modification. We reproduced the model architecture from Oh et al. (2020); this model is a 1D convolutional neural network trained to classify segments of time-series EEG data as SCI or AD.

This model learns temporal filters that are applied equivalently across each EEG channel. Progressing through the network, subsequent layers build more complex features that take into account a larger temporal receptive field, and some invariance is achieved through pooling over time. The model consisted of four convolutional layers, each followed by rectification, max pooling, and batch normalization; convolutional layers were followed by two dense fully-connected layers of 20 and 10 hidden units, respectively, each rectified, and finally a dense connectivity to the output layer with 2 units representing AD yes/no probability logits. All deep learning models were trained with Keras and Tensorflow. The exact Keras code used to specify the architecture can be found in the Supplementary material.

2.2.3 Training

Models were trained for 70 epochs without any early stopping or hyperparameter tuning. A batch size of 32, initial learning rate of 0.0001, and the Adam optimizer were used to optimize models. Training accuracy was computed and stored online during each epoch, and averaged across batches to report the training accuracy for each epoch. To visualize how quickly the models reached their final performance, test set accuracy was also computed after each epoch, averaged across batches. Since we reused the model architecture from prior published work, no model selection was performed; performing ongoing validation on the test is therefore not a source of data leakage. For segment-based holdout, data were split using 10-fold cross-validation (see "Cross-validation" for details).

2.3 Experiment 2: seizure detection

2.3.1 EEG data

We analyzed data from the Siena Scalp EEG Database (Detti, 2020; Detti et al., 2020) hosted on PhysioNet (Goldberger et al., 2000). These recordings were collected in accordance with the Declaration of Helsinki, and approved by the Ethical Committee of the University of Siena. Participants provided written informed consent before beginning data collection. This dataset includes recordings from 14 epilepsy patients (age 20–71 years, nine male) digitized at 512 Hz with electrodes arranged following the International 10-20 system. Seizures in the data were labeled by an expert clinician. This dataset contains 47 seizures in ∼128 h of recorded EEG. To ensure that the data were balanced between seizure and non-seizure epochs, we selected non-seizure data from the beginning of each subject's recordings to match the duration of their seizure-labeled data. This led to 47 min 21 s of data in each condition (1 h 34 min 42 s in total).

In contrast to the previous section where raw time series were used, EEG data were prepared for the classifier analysis in the frequency domain, following the approach used by Rashed-Al-Mahfouz and colleagues (Rashed-Al-Mahfuz et al., 2021).

Spectrograms were computed with a window length of 256 samples (0.5 s) overlapping by 128 samples (0.25 s), using a Hann taper. Spectrograms were then divided into segments of 1.5 s. As in the original study, we used the RGB representation of the spectrogram (viridis color-map), and exported as 224 × 224 × 3 images for training and testing with the CNN models.

2.3.2 Architecture

The aim of this study was to evaluate the impact of different cross-validation choices, not to identify a highly-performing model architecture. We therefore reused the model architecture presented by Rashed-Al-Mahfuz et al. (2021) without modification. No model selection or hyperparameter tuning was performed. To handle 3D spectrogram data (vs. 2D time-series used in the previous section), a 2D convolutional neural network was used. This model learns 2D spectrotemporal features that are applied equivalently across the spectrogram. The model contains four convolutional layers, each followed by rectification, pooling, and batch normalization, followed by two hidden fully-connected layers of 256 and 512 units each, dropout, and a final classification layer of 2 units corresponding to seizure yes/no. The exact Keras code used to specify the architecture can be found in the Supplementary material.

2.3.3 Training

Models were trained for 70 epochs with no early stopping. We used the RMSProp optimimzer with a batch size of 32 and a learning rate of 0.00001. Training accuracy was computed and stored online during each epoch, and averaged across batches to report the training accuracy for each epoch. To visualize how quickly the models reached their final performance, test set accuracy was also computed after each epoch, averaged across batches. Since we reused the model architecture from prior published work, no model selection was performed; performing ongoing validation on the test is therefore not a source of data leakage.

2.4 Cross-validation

This study is primarily concerned with the consequences of different approaches to splitting the data between training and test sets. We assess two types of train-test split: (1) holding out individual segments of EEG data without regard for subject ID ("segment-based holdout"), and (2) holding out entire subjects, ensuring that all segments for a given subject appear in only the training or the test set ("subject-based holdout"; Figure 1).

2.4.1 Segment-based holdout

Segment-based cross-validation considers all EEG segments to be equivalent, and divides them into training and validation partitions without considering subject ID. This segment-holdout approach will lead to data leakage if there is statistical non-independence due to multiple EEG segments coming from each subject. Given n segments and m time-points per segments, we construct a matrix X of EEG segments of size (n, m), and a vector $\boldsymbol{y}$ of diagnostic label of length n. The cross-validation is a simple partition of the index vector $\boldsymbol{\alpha} = \{1, 2, ..., n\}$ into disjoint subsets $\boldsymbol{\alpha}_{\text{train}}$ and $\boldsymbol{\alpha}_{\text{test}}$. Where X_i gives the ith segments of X, we then have $X_{\text{train}} = \{X_i\} \forall i \in \boldsymbol{\alpha}_{\text{train}}$, $X_{test} = \{X_i\} \forall i \in \boldsymbol{\alpha}_{\text{test}}$, and $\boldsymbol{y}_{\text{train}} = \{\boldsymbol{y}_i\} \forall i \in \boldsymbol{\alpha}_{\text{train}}$, $\boldsymbol{y}_{\text{test}} = \{\boldsymbol{y}_i\} \forall i \in \boldsymbol{\alpha}_{\text{test}}$.

2.4.2 Subject-based holdout

Subject-based cross-validation takes into account which subject each EEG segment comes from. This approach enforces that each subject appears in only one partition of the cross-validation, ensuring there is no leakage of subject-level information across training and test sets. To create this split, we consider an additional subject vector $\boldsymbol{s}$, which is used to constrain the partition of X and $\boldsymbol{y}$. Concretely, rather than partitioning the index vector $\boldsymbol{\alpha}$, we partition the unique subject vector $\boldsymbol{s}_u$, which gives the unique entries of $\boldsymbol{s}$, and collect all corresponding segments from each subject contained in train and validation partitions into $\boldsymbol{\alpha}_{\text{train}}$ and $\boldsymbol{\alpha}_{\text{test}}$. This enforces the constraint that $\boldsymbol{s}_i \neq \boldsymbol{s}_j \forall i \in \boldsymbol{\alpha}_{\text{train}}, j \in \boldsymbol{\alpha}_{\text{test}}$. To perform k-fold cross-validation, we first divide $\boldsymbol{s}_u$ into k non-overlapping chunks, and each chunk to serve as the validation data in each fold of cross-validation, where the remaining $k - 1$ chunks are reserved for training.

2.5 Literature review

We searched the literature for studies that used deep learning with segments of EEG to classify a variety of diseases. We searched Google Scholar for papers investigating Alzheimer's disease, Parkinson's disease, attention-deficit/hyperactivity disorder (ADHD), depression, schizophrenia, and seizures. We then searched the references of these papers to identify additional publications for inclusion. Following this search, we included every study that used a DNN to identify psychiatric or neurological conditions using EEG. This non-exhaustive search included 63 papers, all of which were published since 2018 and used deep learning to study one of the conditions named above.

Next, we examined how the training and test sets were determined in these studies. If a paper specified that the EEG recordings were split into segments, but did not specify that they used subjects as an organizing factor of the train-test split, we labeled that study as using "segment-based" holdout. Some papers specifically stated that segments from individual subjects were included in both the training and test sets (for example, studies that trained separate models for each subject); these studies were also labeled as segment-based holdout. If a paper specified that all the segments from a single subject were assigned to only the training or the test set, we labeled that study as using "subject-based" holdout. If a study used both segment-based and subject-based holdout in different analyses, we labeled the study as "both". We labeled studies as "unclear" if we could not determine whether the models were trained on segments of EEG recordings, and it was not explicitly stated that subjects were used as a factor in the holdout procedure.

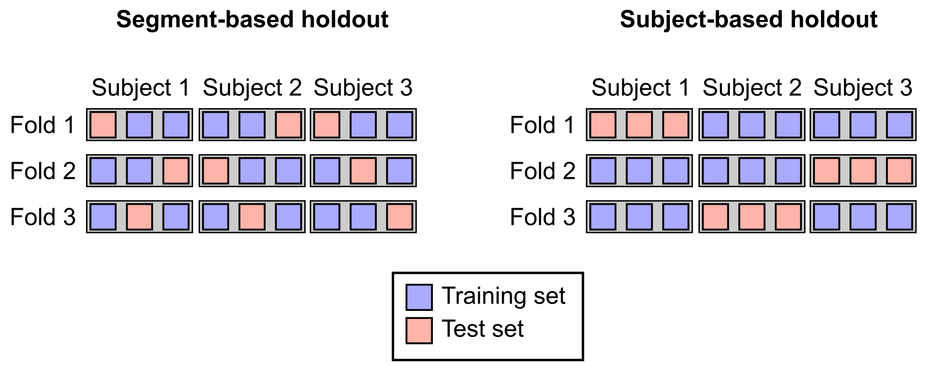

FIGURE 1
Illustration of segment-based and subject-based holdout. This example shows cross-validation with three participants, each of whom have three segments of data, and 3-fold cross-validation (CV). Each row shows a separate CV fold. Each square illustrates a single EEG segment, with blue squares indicating observations in the training set and red squares indicating observations in the test set. Gray rectangles are drawn around observations from the same subject.

3 Results

3.1 Data leakage leads to biased test-set accuracy

We analyze two datasets to test how the estimated accuracy of a DNN classifier depends on the train-test split. First, we examine the effects of data leakage in a patient-level classifier by training a model to diagnose Alzheimer's disease. Second, we examine the effects of data leakage in a segment-level classifier by training a model to identify periods of time that include an epileptic seizure. In each of these analyses, we reuse a published DNN architecture to analyze an existing dataset.

3.1.1 Identifying patients with Alzheimer's disease

To determine whether segment-based holdout leads to a biased estimate of accuracy, we first trained a CNN to diagnose Alzheimer's disease using segments of EEG. When the EEG segments were split into training and test sets without considering subject ID, the model showed nearly perfect test-set accuracy of 99.8% (99.1–100.0%) (Figure 2A). Performance quickly approached ceiling within the first 15 training epochs (Figure 3A). This high accuracy is consistent with prior studies that use segment-based holdout and report high accuracy for CNNs at identifying neurological disorders (Acharya et al., 2018b; Lee et al., 2019; Oh et al., 2020). Could this pattern of high accuracy reflect data leakage, instead of a robust and generalizable classifier?

When we used subject-based holdout, ensuring that individual subjects' data did not appear in both the training and test sets, test accuracy dropped to 53.0% (43.1–64.8%), with 95% confidence intervals that included chance performance of 50%. Performance remained low throughout the training epochs (Figure 3B). Compared with subject-based holdout, segment-based holdout significantly overestimates the model performance on previously-unseen subjects (Wilcoxon $T = 0.0$, $p = 0.002$).

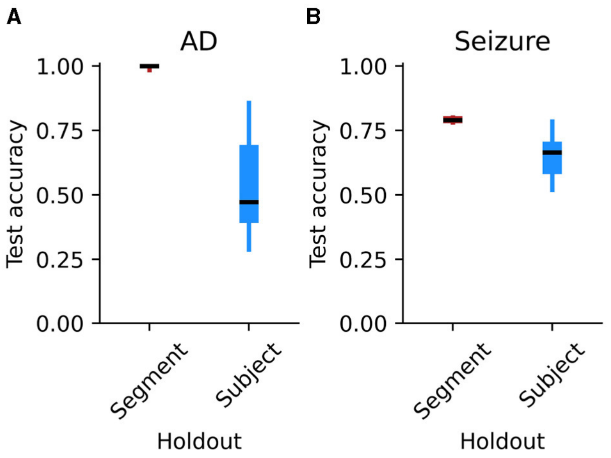

FIGURE 2
Test-set accuracy of CNN models predicting held-out data, plotted separately for segment-based holdout and subject-based holdout. **(A)** Accuracy for models trained to classify Alzheimer's disease in individual subjects. Boxes show the inter-quartile range, dark lines show the median, and whiskers extend to the minimum and maximum points. **(B)** Accuracy for models trained to identify seizures in segments of EEG data. Details as in **(A)**.

3.1.2 Identifying segments containing epileptic seizures

In some cases, artificial neural network models have been used to identify time-limited events within ongoing brain activity, such as epileptic seizures. Does segment-based holdout also lead to data leakage when labeling periods of time within subjects? To answer this question, we trained a CNN to classify segments of EEG data as containing an epileptic seizure or not.

When the EEG segments were split into training and test sets without considering subject ID, the model reached a high test-set accuracy of 79.1% (78.8–79.4%) (Figure 2B). Accuracy leveled out within 10 training epochs (Figure 3C). When individual subjects' data segments were restricted to appear in only the training or test set, however, accuracy fell to 65.1% (61.3–69.1%). Accuracy

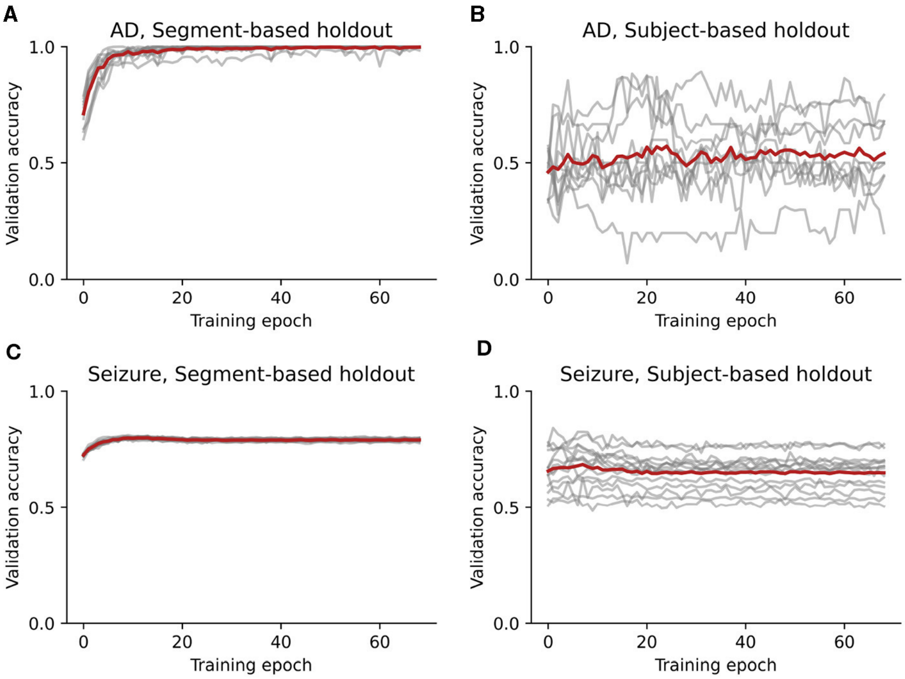

FIGURE 3
Test-set accuracy of CNN models plotted as a function of the training epoch. Gray lines show accuracy in individual cross-validation folds, and red lines show the average across folds. **(A)** Accuracy for models trained to classify Alzheimer's disease using segment-based holdout. **(B)** Accuracy for models trained to classify Alzheimer's disease using subject-based holdout. **(C)** Accuracy for models trained to identify seizures using segment-based holdout. **(D)** Accuracy for models trained to identify seizures using subject-based holdout.

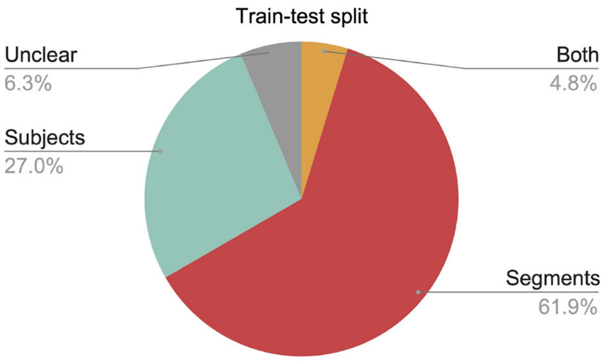

FIGURE 4
Number of studies using each type of test-split. "Segments": Segments of EEG data were assigned to the training and test sets without regard to subject; this approach leads to data leakage. "Subjects": Each subject's data appeared in only the training set or the test set. "Both": Both the Subjects and Segments approaches were used in different analyses. "Unclear": We could not determine which approach was used for train-test splits.

remained low throughout training epochs (Figure 3D). Even when the model is tasked with labeling periods of activity within subjects, segment-based holdout significantly overestimates performance on previously-unseen subjects (Wilcoxon $T = 0.0, p = 0.0001$).

3.2 Data leakage in published EEG studies

Do published translational EEG studies suffer from subject-specific data leakage, or do they avoid it by computing their test-set accuracy on held-out subjects? We examined the train-test split strategies in published studies that attempted to identify a clinical disorder using DNNs with EEG recordings. Out of the 63 relevant papers we found, only 17 (27.0%) unambiguously avoided this type of data leakage (Figure 4; Table 1). Leakage of subject-specific information is pervasive in the translational EEG literature.

4 Discussion

In EEG studies using deep learning, data leakage can occur when segments of data from the same subjects are included in both the training and test sets. Here we demonstrate that leakage of subject-specific information can dramatically overestimate the real-world clinical performance of a DNN classifier. Our Alzheimer's CNN classifier appeared to have an accuracy of above 99% when using segment-based holdout, but its true performance on previously-unseen subjects was indistinguishable from chance. We found this bias in test-set performance both in a between-subjects task (identifying patients with Alzheimer's disease in Experiment 1) and in a within-subjects task (identifying segments that contain a seizure in Experiment 2). Next, we show that this type of data leakage appears in the majority of published translational DNN-EEG studies we examined. Together, these results illustrate how an improperly-designed training-test split can bias the results of

TABLE 1 Prior translational studies using deep learning with EEG.

Article	Target	Test split
Ahmadi et al. (2021)	ADHD	Segments
Bakhtyari and Mirzaei (2022)	ADHD	Segments
Chang et al. (2022)	ADHD	Subjects
Chen et al. (2019a)	ADHD	Segments
Chen et al. (2019b)	ADHD	Segments
Dubreuil-Vall et al. (2020)	ADHD	Subjects
Mafi and Radfar (2022)	ADHD	Segments
Moghaddari et al. (2020)	ADHD	Segments
TaghiBeyglou et al. (2022)	ADHD	Subjects
Tosun (2021)	ADHD	Segments
Vahid et al. (2019)	ADHD	Subjects
Zhou et al. (2022)	ADHD	Unclear
Kim et al. (2018)	Alcoholism	Segments
Bi and Wang (2019)	Alzheimer's	Segments
Gkenios et al. (2022)	Alzheimer's	Both
Huggins et al. (2021)	Alzheimer's	Segments
Ieracitano et al. (2019)	Alzheimer's	Both
Kim and Kim (2018)	Alzheimer's	Subjects
Morabito et al. (2016)	Alzheimer's	Subjects
You et al. (2020)	Alzheimer's	Segments
Zhao and He (2015)	Alzheimer's	Segments
Acharya et al. (2018b)	Depression	Segments
Ay et al. (2019)	Depression	Segments
Kwon et al. (2019)	Depression	Subjects
Li et al. (2019)	Depression	Subjects
Li X. et al. (2020)	Depression	Subjects
Mumtaz and Qayyum (2019)	Depression	Segments
Uyulan et al. (2021)	Depression	Unclear
Xie et al. (2020)	Depression	Unclear
Zhang et al. (2020)	Depression	Segments
Khare et al. (2021)	Parkinson's	Segments
Lee et al. (2019)	Parkinson's	Segments
Loh et al. (2021)	Parkinson's	Segments
Oh et al. (2020)	Parkinson's	Segments
Shaban (2021)	Parkinson's	Segments
Shaban and Amara (2022)	Parkinson's	Subjects
Shi et al. (2019)	Parkinson's	Subjects
Ahmedt-Aristizabal et al. (2020)	Schizophrenia	Subjects
Chu et al. (2017)	Schizophrenia	Segments
Oh et al. (2019)	Schizophrenia	Both
Shalbaf et al. (2020)	Schizophrenia	Segments

(Continued)

TABLE 1 (Continued)

Article	Target	Test split
Acharya et al. (2018a)	Seizure	Segments
Avcu et al. (2019)	Seizure	Subjects
Choi et al. (2019)	Seizure	Subjects
Daoud and Bayoumi (2019)	Seizure	Segments
Emami et al. (2019)	Seizure	Subjects
Fürbass et al. (2020)	Seizure	Subjects
Gao et al. (2020)	Seizure	Segments
Hussein et al. (2019)	Seizure	Segments
Iešmantas and Alzbutas (2020)	Seizure	Subjects
Jana et al. (2020)	Seizure	Segments
Khan et al. (2017)	Seizure	Segments
Li Y. et al. (2020)	Seizure	Segments
Liang et al. (2020)	Seizure	Segments
Raghu et al. (2020)	Seizure	Unclear
Rashed-Al-Mahfuz et al. (2021)	Seizure	Segments
Truong et al. (2018)	Seizure	Segments
Ullah et al. (2018)	Seizure	Segments
Wei et al. (2018)	Seizure	Segments
Wei et al. (2019)	Seizure	Segments
Zhao et al. (2020)	Seizure	Segments
Zhou et al. (2018)	Seizure	Segments
Boualleque et al. (2020)	Seizure and autism	Segments

Each line in the table describes one published translational study using a DNN with EEG data. The "Target" column holds the clinical condition being classified. The "Test split" column shows the approach used to determine how the data were divided into training and test sets. "Segments": Segments of EEG data were assigned to the training and test sets without regard to subject; this approach leads to data leakage. "Subjects": Each subject's data appeared in only the training set or the test set. "Both": Both the Subjects and Segments approaches were used in different analyses. "Unclear": We could not determine which approach was used for train-test splits.

DNN studies, and show that biased results are widespread in the published literature.

To be useful in a clinical setting, a diagnostic classifier must be able to identify a disease in new patients. Models trained using segment-based holdout, however, strongly overestimate their ability to perform this task. Instead, these models may learn patterns associated with individual subjects, and then associate those idiosyncratic patterns with a diagnosis. As a consequence, performance of these models drops precipitously when they are tested in new subjects, and performance is unlikely to generalize to a new dataset. When training a translational DNN classifier, the model must be tested with subjects who were not included in the training set.

Our results show that segment-based cross-validation inflates estimates of out-of-sample model performance when training on segments from resting-state EEG. However, the same principles of data leakage will apply to task-based EEG; providing a classifier with person-specific information enables it to artificially inflate performance.

Although this study focused on Alzheimer's Disease and epileptic seizures, our findings are not particular those diseases. Classification studies will overestimate model generalization whenever data from individual participants is present in both the training and test sets. Prior review articles have summarized the details and idiosyncrasies of DNN models in the context of AD (Cassani et al., 2018; Wen et al., 2020) and seizures (Rasheed et al., 2020; Shoeibi et al., 2021).

4.1 Data leakage in between- and within-subjects designs

We find that segment-based cross-validation overestimates performance for both between-subjects (Alzheimer's disease, Experiment 1) and within-subjects comparisons (seizures, Experiment 2). However, the magnitude of this overestimate was smaller in a within-subjects comparison (Figure 2). What leads to this difference in the size of the effect between the two tasks? In a between-subjects task, the classifier can simply associate a label with each individual participant. In a within-subjects task, however, this shortcut is not available to the model. Instead, it must learn a representation of the labels—albeit a representation that may be contaminated by multiple segments coming from the same event, or one that may be specific to a given participant.

4.2 Data leakage when identifying events within subjects

Instead of identifying a disease in each subject, some studies attempt to identify a diseased process in each segment of time (see Table 1). DNN models of epilepsy, for example, often aim to classify the segments of data that contain a seizure. We demonstrated in Experiment 2 that those studies are not immune to data leakage in training-test splits: the accuracy in novel subjects is strongly overestimated when the test set includes subjects who were also in the training set. This result could arise if the model uses different patterns to identify seizures in each subject.

Subject-specific studies indicate that a bespoke classifier could be trained to identify seizures in each new patient (Jana et al., 2020; Liang et al., 2020; Li Y. et al., 2020). However, this would require every patient to have a large dataset of recordings that have already been labeled, which limits the clinical utility of this approach. A more realistic approach is to train DNN models to identify events in unseen patients.

4.3 Data leakage in other methods

In studies which have only one observation per subject, cross-validation is trivial – single observations are simply assigned to the training or test set. However, in EEG and many other medical imagining methods, the data from each subject is routinely split into multiple segments. In this paper, we showed how data leakage can arise when a long recording is split into multiple shorter segments. However, the same principles apply to any other method that introduces statistical non-independence between the training and test sets. For example, some EEG-based DNNs treat every channel independently, and use information from each channel as a separate observation (Loh et al., 2021). Those studies are likely to suffer from substantial data leakage, since physiological sources of electrical activity appear redundantly across multiple EEG scalp electrodes (Michel and He, 2019).

These principles also apply to other medical imaging methods and classifiers. Similar patterns of "identity confounding" data leakage have been documented in studies using functional (Wen et al., 2018) and anatomical (Wen et al., 2020) MRI, optical coherence tomography (OCT) (Tampu et al., 2022), accelerometer and gyroscope recordings from smartphones (Saeb et al., 2017), audio voice recordings (Chaibub Neto et al., 2019; Tougui et al., 2021), and performance on motor tasks (Chaibub Neto et al., 2019). Furthermore, data leakage due to identity confounding is not limited to deep neural networks, and has been uncovered using random forests (Saeb et al., 2017; Chaibub Neto et al., 2019; Tougui et al., 2021) and support vector machines (Tougui et al., 2021).

4.4 Caveats

We find that segment-based cross-validation leads to data leakage, and this type of cross-validation is common in translational EEG studies. This conclusion mirrors results from studies examining a variety of other types of data and classifier models (Saeb et al., 2017; Wen et al., 2018, 2020; Chaibub Neto et al., 2019; Tougui et al., 2021; Tampu et al., 2022). The precise amount of data leakage and the bias that it introduces, however, are likely to differ based on the details of the experiment. For example, if a study trains a classifier to identify individual subjects with a disease, then there may be stronger bias when the study involves fewer participants (Saeb et al., 2017). The model architecture may also influence the amount of data leakage: a model that can more effectively learn subject-specific representations could show stronger bias than a model that cannot learn subject-specific patterns.

5 Conclusion

Data leakage occurs when EEG segments from one subject appear in the both the training and test sets. As a result, the test set accuracy dramatically overestimates the classifier's performance in new subjects. This type of data leakage is common in published studies using DNNs and translational EEG. To accurately estimate a model's performance, researchers must ensure that each subject's data is included in only the training or the test set, but not both.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: EEG data for experiment 1 were provided by the Pacific Neuroscience Institute. These data are described by Ganapathi et al. (2022), and can be accessed through agreement with the authors of that study. EEG data for experiment 2 were downloaded from the publicly-available Siena Scalp EEG Database hosted on PhysioNet (https://physionet.org/content/siena-scalp-eeg/1.0.0/).

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

GB: Conceptualization, Formal analysis, Visualization, Writing – original draft. JK: Data curation, Formal analysis, Software, Visualization, Writing – review & editing. NB: Software, Writing – original draft, Writing – review & editing. YW: Conceptualization, Software, Writing – review & editing. RG: Resources, Writing – review & editing. DM: Resources, Supervision, Writing – review & editing. SG: Funding acquisition, Supervision, Writing – review & editing. KY: Writing – review & editing. CQ: Conceptualization, Data curation, Writing – review & editing. CL: Conceptualization, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by SPARK Neuro, Inc. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Acknowledgments

The contents of this manuscript have previously appeared online as a preprint on medRxiv (https://www.medrxiv.org/content/10.1101/2024.01.16.24301366v1).

Conflict of interest

GB, JK, NB, YW, SG, KY, CQ, and CL were employed at SPARK Neuro Inc., a medical technology company developing diagnostic aids to help clinicians identify and assess neurodegenerative disease.

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2024.1373515/full#supplementary-material

References

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., and Adeli, H. (2018a). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Comput. Biol. Med.* 100, 270–278. doi: 10.1016/j.compbiomed.2017.09.017

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adeli, H., and Subha, D. P. (2018b). Automated EEG-based screening of depression using deep convolutional neural network. *Comput. Methods Programs Biomed.* 161, 103–113. doi: 10.1016/j.cmpb.2018.04.012

Ahmadi, A., Kashefi, M., Shahrokhi, H., and Nazari, M. A. (2021). Computer aided diagnosis system using deep convolutional neural networks for ADHD subtypes. *Biomed. Signal Process. Control* 63:102227. doi: 10.1016/j.bspc.2020.102227

Ahmedt-Aristizabal, D., Fernando, T., Denman, S., Robinson, J. E., Sridharan, S., Johnston, P. J., et al. (2020). Identification of children at risk of schizophrenia via deep learning and EEG responses. *IEEE J. Biomed. Health Inf.* 25, 69–76. doi: 10.1109/JBHI.2020.2984238

Avcu, M. T., Zhang, Z., and Chan, D. W. S. (2019). "Seizure detection using least EEG channels by deep convolutional neural network," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton: IEEE), 1120–1124.

Ay, B., Yildirim, O., Talo, M., Baloglu, U. B., Aydin, G., Puthankattil, S. D., et al. (2019). Automated depression detection using deep representation and sequence learning with EEG signals. *J. Med. Syst.* 43, 1–12. doi: 10.1007/s10916-019-1345-y

Bakhtyari, M., and Mirzaei, S. (2022). ADHD detection using dynamic connectivity patterns of EEG data and convlstm with attention framework. *Biomed. Signal Process. Control* 76:103708. doi: 10.1016/j.bspc.2022.103708

Bergeron, D., Flynn, K., Verret, L., Poulin, S., Bouchard, R. W., Bocti, C., et al. (2017). Multicenter validation of an MMSE-MoCA conversion table. *J. Am. Geriatr. Soc.* 65, 1067–1072. doi: 10.1111/jgs.14779

Bi, X., and Wang, H. (2019). Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning. *Neural Netw.* 114, 119–135. doi: 10.1016/j.neunet.2019.02.005

Bouallegue, G., Djemal, R., Alshebeili, S. A., and Aldhalaan, H. (2020). A dynamic filtering DF-RNN deep-learning-based approach for EEG-based neurological disorders diagnosis. *IEEE Access* 8, 206992–207007. doi: 10.1109/ACCESS.2020.3037995

Cassani, R., Estarellas, M., San-Martin, R., Fraga, F. J., and Falk, T. H. (2018). Systematic review on resting-state EEG for Alzheimer's disease diagnosis and progression assessment. *Dis. Mark.* 2018:5174815. doi: 10.1155/2018/5174815

Chaibub Neto, E., Pratap, A., Perumal, T. M., Tummalacherla, M., Snyder, P., Bot, B. M., et al. (2019). Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ Digit. Med.* 2:99. doi: 10.1038/s41746-019-0178-x

Chang, Y., Stevenson, C., Chen, I.-C., Lin, D.-S., and Ko, L.-W. (2022). Neurological state changes indicative of ADHD in children learned via EEG-based LSTM networks. *J. Neural Eng.* 19:016021. doi: 10.1088/1741-2552/ac4f07

Chen, H., Song, Y., and Li, X. (2019a). A deep learning framework for identifying children with ADHD using an EEG-based brain network. *Neurocomputing* 356, 83–96. doi: 10.1016/j.neucom.2019.04.058

Chen, H., Song, Y., and Li, X. (2019b). Use of deep learning to detect personalized spatial-frequency abnormalities in EEGs of children with ADHD. *J. Neural Eng.* 16:066046. doi: 10.1088/1741-2552/ab3a0a

Choi, G., Park, C., Kim, J., Cho, K., Kim, T.-J., Bae, H., et al. (2019). "A novel multi-scale 3D CNN with deep neural network for epileptic seizure detection," in *2019 IEEE International Conference on Consumer Electronics (ICCE)* (Las Vegas, NV: IEEE), 1–2.

Chu, L., Qiu, R., Liu, H., Ling, Z., Zhang, T., and Wang, J. (2017). Individual recognition in schizophrenia using deep learning methods with random forest and voting classifiers: Insights from resting state EEG streams. *arXiv* [preprint].

Daoud, H., and Bayoumi, M. A. (2019). Efficient epileptic seizure prediction based on deep learning. *IEEE Trans. Biomed. Circuits Syst.* 13, 804–813. doi: 10.1109/TBCAS.2019.2929053

de Bardeci, M., Ip, C. T., and Olbrich, S. (2021). Deep learning applied to electroencephalogram data in mental disorders: a systematic review. *Biol. Psychol.* 162:108117. doi: 10.1016/j.biopsycho.2021.108117

Demuru, M., and Fraschini, M. (2020). EEG fingerprinting: subject-specific signature based on the aperiodic component of power spectrum. *Comput. Biol. Med.* 120:103748. doi: 10.1016/j.compbiomed.2020.103748

Detti, P. (2020). *Siena Scalp EEG Database (version 1.0.0).* PhysioNet. doi: 10.13026/5d4a-j060

Detti, P., Vatti, G., and Zabalo Manrique de Lara, G. (2020). EEG synchronization analysis for seizure prediction: a study on data of noninvasive recordings. *Processes* 8:846. doi: 10.3390/pr8070846

Dubreuil-Vall, L., Ruffini, G., and Camprodon, J. A. (2020). Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG. *Front. Neurosci.* 14:251. doi: 10.3389/fnins.2020.00251

Emami, A., Kunii, N., Matsuo, T., Shinozaki, T., Kawai, K., and Takahashi, H. (2019). Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images. *NeuroImage Clin.* 22:101684. doi: 10.1016/j.nicl.2019.101684

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Fürbass, F., Kural, M. A., Gritsch, G., Hartmann, M., Kluge, T., and Beniczky, S. (2020). An artificial intelligence-based EEG algorithm for detection of epileptiform EEG discharges: validation against the diagnostic gold standard. *Clin. Neurophysiol.* 131, 1174–1179. doi: 10.1016/j.clinph.2020.02.032

Ganapathi, A. S., Glatt, R. M., Bookheimer, T. H., Popa, E. S., Ingemanson, M. L., Richards, C. J., et al. (2022). Differentiation of subjective cognitive decline, mild cognitive impairment, and dementia using qEEG/ERP-based cognitive testing and volumetric MRI in an outpatient specialty memory clinic. *J. Alzheimers Dis.* 90, 1–9. doi: 10.3233/JAD-220616

Gao, Y., Gao, B., Chen, Q., Liu, J., and Zhang, Y. (2020). Deep convolutional neural network-based epileptic electroencephalogram (EEG) signal classification. *Front. Neurol.* 11:375. doi: 10.3389/fneur.2020.00375

Gkenios, G., Latsiou, K., Diamantaras, K., Chouvarda, I., and Tsolaki, M. (2022). "Diagnosis of Alzheimer's disease and mild cognitive impairment using EEG and recurrent neural networks," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine* & *Biology Society (EMBC)* (Glasgow: IEEE), 3179–3182.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, e215—220. doi: 10.1161/01.CIR.101.23.e215

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510. doi: 10.1038/s41568-018-0016-5

Huggins, C. J., Escudero, J., Parra, M. A., Scally, B., Anghinah, R., Vitória Lacerda De Araújo, A., et al. (2021). Deep learning of resting-state electroencephalogram signals for three-class classification of Alzheimer's disease, mild cognitive impairment and healthy ageing. *J. Neural Eng.* 18:046087. doi: 10.1088/1741-2552/ac05d8

Hussein, R., Palangi, H., Ward, R. K., and Wang, Z. J. (2019). Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals. *Clin. Neurophysiol.* 130, 25–37. doi: 10.1016/j.clinph.2018.10.010

Ieracitano, C., Mammone, N., Bramanti, A., Hussain, A., and Morabito, F. C. (2019). A Convolutional Neural Network approach for classification of dementia stages based on 2D-spectral representation of EEG recordings. *Neurocomputing* 323, 96–107. doi: 10.1016/j.neucom.2018.09.071

Iešmantas, T., and Alzbutas, R. (2020). Convolutional neural network for detection and classification of seizures in clinical data. *Med. Biol. Eng. Comp.* 58, 1919–1932. doi: 10.1007/s11517-020-02208-7

Jana, R., Bhattacharyya, S., and Das, S. (2020). "Patient-specific seizure prediction using the convolutional neural networks," in *Intelligence Enabled Research*, eds. S. Bhattacharyya, S. Mitra, and P. Dutta (Springer), 51–60.

Kaka, H., Zhang, E., and Khan, N. (2021). Artificial intelligence and deep learning in neuroradiology: exploring the new frontier. *Can. Assoc. Radiol. J.* 72, 35–44. doi: 10.1177/0846537120954293

Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transact. Knowl. Discov. Data* 6, 1–21. doi: 10.1145/2382577.2382579

Khan, H., Marcuse, L., Fields, M., Swann, K., and Yener, B. (2017). Focal onset seizure prediction using convolutional networks. *IEEE Transact. Biomed. Eng.* 65, 2109–2118. doi: 10.1109/TBME.2017.2785401

Khare, S. K., Bajaj, V., and Acharya, U. R. (2021). PDCNNet: an automatic framework for the detection of Parkinson's disease using EEG signals. *IEEE Sens. J.* 21, 17017–17024. doi: 10.1109/JSEN.2021.3080135

Kim, D., and Kim, K. (2018). "Detection of early stage Alzheimer's disease using EEG relative power with deep neural network," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, HI), 352–355.

Kim, S., Kim, J., and Chun, H.-W. (2018). Wave2vec: vectorizing electroencephalography bio-signal for prediction of brain disease. *Int. J. Environ. Res. Public Health* 15:1750. doi: 10.3390/ijerph15081750

Kwon, H., Kang, S., Park, W., Park, J., and Lee, Y. (2019). "Deep learning based pre-screening method for depression with imagery frontal EEG channels," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)* (Jeju: IEEE), 378–380.

Langa, K. M., and Levine, D. A. (2014). The diagnosis and management of mild cognitive impairment: a clinical review. *J. Am. Med. Assoc.* 312, 2551–2561. doi: 10.1001/jama.2014.13806

Lee, S., Hussein, R., and McKeown, M. J. (2019). "A deep convolutional-recurrent neural network architecture for Parkinson's disease EEG classification," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (Ottawa, ON: IEEE), 1–4.

Li, X., La, R., Wang, Y., Hu, B., and Zhang, X. (2020). A deep learning approach for mild depression recognition based on functional connectivity using electroencephalography. *Front. Neurosci.* 14:192. doi: 10.3389/fnins.2020.00192

Li, X., La, R., Wang, Y., Niu, J., Zeng, S., Sun, S., et al. (2019). EEG-based mild depression recognition using convolutional neural network. *Med. Biol. Eng. Comp.* 57, 1341–1352. doi: 10.1007/s11517-019-01959-2

Li, Y., Liu, Y., Cui, W.-G., Guo, Y.-Z., Huang, H., and Hu, Z.-Y. (2020). Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network. *IEEE Transact. Neural Syst. Rehabil. Eng.* 28, 782–794. doi: 10.1109/TNSRE.2020.2973434

Liang, W., Pei, H., Cai, Q., and Wang, Y. (2020). Scalp EEG epileptogenic zone recognition and localization based on long-term recurrent convolutional network. *Neurocomputing* 396, 569–576. doi: 10.1016/j.neucom.2018.10.108

Loh, H. W., Ooi, C. P., Palmer, E., Barua, P. D., Dogan, S., Tuncer, T., et al. (2021). GaborPDNet: gabor transformation and deep neural network for Parkinson's disease detection using EEG signals. *Electronics* 10:1740. doi: 10.3390/electronics10141740

Mafi, M., and Radfar, S. (2022). High dimensional convolutional neural network for EEG connectivity-based diagnosis of ADHD. *J. Biomed. Phys. Eng.* 12, 645–654. doi: 10.31661/jbpe.v0i0.2108-1380

Mall, P. K., Singh, P. K., Srivastav, S., Narayan, V., Paprzycki, M., Jaworska, T., et al. (2023). A comprehensive review of deep neural networks for medical image processing: recent developments and future opportunities. *Healthc. Anal.* 4:100216. doi: 10.1016/j.health.2023.100216

Michel, C. M., and He, B. (2019). EEG source localization. *Handb. Clin. Neurol.* 160, 85–101. doi: 10.1016/B978-0-444-64032-1.00006-0

Moghaddari, M., Lighvan, M. Z., and Danishvar, S. (2020). Diagnose ADHD disorder in children using convolutional neural network based on continuous mental task EEG. *Comput. Methods Programs Biomed.* 197:105738. doi: 10.1016/j.cmpb.2020.105738

Morabito, F. C., Campolo, M., Ieracitano, C., Ebadi, J. M., Bonanno, L., Bramanti, A., et al. (2016). "Deep convolutional neural networks for classification of mild cognitive impaired and Alzheimer's disease patients from scalp EEG recordings," in *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow (RTSI)* (Bologna), 1–6.

Mumtaz, W., and Qayyum, A. (2019). A deep learning framework for automatic diagnosis of unipolar depression. *Int. J. Med. Inform.* 132:103983. doi: 10.1016/j.ijmedinf.2019.103983

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Oh, S. L., Hagiwara, Y., Raghavendra, U., Yuvaraj, R., Arunkumar, N., Murugappan, M., et al. (2020). A deep learning approach for Parkinson's disease diagnosis from EEG signals. *Neur. Comp. Appl.* 32, 10927–10933. doi: 10.1007/s00521-018-3689-5

Oh, S. L., Vicnesh, J., Ciaccio, E. J., Yuvaraj, R., and Acharya, U. R. (2019). Deep convolutional neural network model for automated diagnosis of schizophrenia using EEG signals. *Appl. Sci.* 9:2870. doi: 10.3390/app9142870

Raghu, S., Sriraam, N., Temel, Y., Rao, S. V., and Kubben, P. L. (2020). EEG based multi-class seizure type classification using convolutional neural network and transfer learning. *Neur. Netw.* 124, 202–212. doi: 10.1016/j.neunet.2020.01.017

Rashed-Al-Mahfuz, M., Moni, M. A., Uddin, S., Alyami, S. A., Summers, M. A., and Eapen, V. (2021). A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (EEG) data. *IEEE J. Transl. Eng. Health Med.* 9, 1–12. doi: 10.1109/JTEHM.2021.3050925

Rasheed, K., Qayyum, A., Qadir, J., Sivathamboo, S., Kwan, P., Kuhlmann, L., et al. (2020). Machine learning for predicting epileptic seizures using EEG signals: a review. *IEEE Rev. Biomed. Eng.* 14, 139–155. doi: 10.1109/RBME.2020.3008792

Rosset, S., Perlich, C., Świrszcz, G., Melville, P., and Liu, Y. (2010). Medical data mining: insights from winning two competitions. *Data Min. Knowl. Discov.* 20, 439–468. doi: 10.1007/s10618-009-0158-x

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., and Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *Gigascience* 6:gix019. doi: 10.1093/gigascience/gix019

Shaban, M. (2021). "Automated screening of Parkinson's disease using deep learning based electroencephalography," in *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 158–161.

Shaban, M., and Amara, A. W. (2022). Resting-state electroencephalography based deep-learning for the detection of Parkinson's disease. *PLoS ONE* 17:e0263159. doi: 10.1371/journal.pone.0263159

Shalbaf, A., Bagherzadeh, S., and Maghsoudi, A. (2020). Transfer learning with deep convolutional neural network for automated detection of schizophrenia from EEG signals. *Phys. Eng. Sci. Med.* 43, 1229–1239. doi: 10.1007/s13246-020-00925-9

Shi, X., Wang, T., Wang, L., Liu, H., and Yan, N. (2019). "Hybrid convolutional recurrent neural networks outperform CNN and RNN in task-state EEG detection for Parkinson's disease," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (Lanzhou: IEEE), 939–944.

Shoeibi, A., Khodatars, M., Ghassemi, N., Jafari, M., Moridian, P., Alizadehsani, R., et al. (2021). Epileptic seizures detection using deep learning techniques: a review. *Int. J. Environ. Res. Public Health* 18:5780. doi: 10.3390/ijerph18115780

TaghiBeyglou, B., Shahbazi, A., Bagheri, F., Akbarian, S., and Jahed, M. (2022). Detection of ADHD cases using CNN and classical classifiers of raw EEG. *Comp. Methods Progr. Biomed.* 2:100080. doi: 10.1016/j.cmpbup.2022.100080

Tampu, I. E., Eklund, A., and Haj-Hosseini, N. (2022). Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Sci. Data* 9:580. doi: 10.1038/s41597-022-01618-6

Tosun, M. (2021). Effects of spectral features of EEG signals recorded with different channels and recording statuses on ADHD classification with deep learning. *Phys. Eng. Sci. Med.* 44, 693–702. doi: 10.1007/s13246-021-01018-x

Tougui, I., Jilbab, A., and El Mhamdi, J. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthc. Inform. Res.* 27, 189–199. doi: 10.4258/hir.2021.27.3.189

Truong, N. D., Nguyen, A. D., Kuhlmann, L., Bonyadi, M. R., Yang, J., Ippolito, S., et al. (2018). Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neur. Netw.* 105, 104–111. doi: 10.1016/j.neunet.2018.04.018

Ullah, I., Hussain, M., Aboalsamh, H., et al. (2018). An automated system for epilepsy detection using EEG brain signals based on deep learning approach. *Expert Syst. Appl.* 107, 61–71. doi: 10.1016/j.eswa.2018.04.021

Uyulan, C., Ergüzel, T. T., Unubol, H., Cebi, M., Sayar, G. H., Nezhad Asad, M., et al. (2021). Major depressive disorder classification based on different convolutional neural network models: deep learning approach. *Clin. EEG Neurosci.* 52, 38–51. doi: 10.1177/1550059420916634

Vahid, A., Bluschke, A., Roessner, V., Stober, S., and Beste, C. (2019). Deep learning based on event-related EEG differentiates children with ADHD from healthy controls. *J. Clin. Med.* 8:1055. doi: 10.3390/jcm8071055

Wei, X., Zhou, L., Chen, Z., Zhang, L., and Zhou, Y. (2018). Automatic seizure detection using three-dimensional CNN based on multi-channel EEG. *BMC Med. Inform. Decis. Mak.* 18, 71–80. doi: 10.1186/s12911-018-0693-8

Wei, X., Zhou, L., Zhang, Z., Chen, Z., and Zhou, Y. (2019). Early prediction of epileptic seizures using a long-term recurrent convolutional network. *J. Neurosci. Methods* 327:108395. doi: 10.1016/j.jneumeth.2019.108395

Wen, D., Wei, Z., Zhou, Y., Li, G., Zhang, X., and Han, W. (2018). Deep learning methods to process fMRI data and their application in the diagnosis of cognitive impairment: a brief overview and our opinion. *Front. Neuroinform.* 12:23. doi: 10.3389/fninf.2018.00023

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., et al. (2020). Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med. Image Anal.* 63:101694. doi: 10.1016/j.media.2020.101694

Xie, Y., Yang, B., Lu, X., Zheng, M., Fan, C., Bi, X., et al. (2020). "Anxiety and depression diagnosis method based on brain networks and convolutional neural networks," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Montreal, QC: IEEE), 1503–1506.

You, Z., Zeng, R., Lan, X., Ren, H., You, Z., Shi, X., et al. (2020). Alzheimer's disease classification with a cascade neural network. *Front. Public Health* 8:584387. doi: 10.3389/fpubh.2020.584387

Zhang, X., Li, J., Hou, K., Hu, B., Shen, J., and Pan, J. (2020). "EEG-based depression detection using convolutional neural network with demographic attention mechanism," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Montreal, QC: IEEE), 128–133.

Zhao, W., Zhao, W., Wang, W., Jiang, X., Zhang, X., Peng, Y., et al. (2020). A novel deep neural network for robust detection of seizures using EEG signals. *Comput. Math. Methods Med.* 2020:9689821. doi: 10.1155/2020/9689821

Zhao, Y., and He, L. (2015). "Deep learning in the EEG diagnosis of Alzheimer's disease," in *Computer Vision - ACCV 2014 Workshops, Lecture Notes in Computer Science*, eds C. Jawahar, and S. Shan (Cham: Springer International Publishing), 340–353.

Zhou, D., Liao, Z., and Chen, R. (2022). Deep learning enabled diagnosis of children's ADHD based on the big data of video screen long-range EEG. *J. Healthc. Eng.* 2022:5222136. doi: 10.1155/2022/5222136

Zhou, M., Tian, C., Cao, R., Wang, B., Niu, Y., Hu, T., et al. (2018). Epileptic seizure detection based on EEG signals and CNN. *Front. Neuroinform.* 12:95. doi: 10.3389/fninf.2018.00095