Amplifying pathological detection in EEG signaling pathways through cross-dataset transfer learning

Darvishi-Bayazi, Mohammad-Javad; Ghaemi, Mohammad Sajjad; Lesort, Timothee; Arefin, Md. Rifat; Faubert, Jocelyn; Rish, Irina

Publisher's version / Version de l'éditeur:
https://doi.org/10.1016/j.compbiomed.2023.107893
Computers in Biology and Medicine, 169, C, 2023-12-30

Computers in Biology and Medicine 169 (2024) 107893

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Amplifying pathological detection in EEG signaling pathways through cross-dataset transfer learning

Mohammad-Javad Darvishi-Bayazi [a,b,c,*], Mohammad Sajjad Ghaemi [d], Timothee Lesort [a,c], Md. Rifat Arefin [a,c], Jocelyn Faubert [b,c], Irina Rish [a,c]

[a] Mila, Québec AI Institute, Montréal, QC, Canada
[b] Faubert Lab, Montréal, QC, Canada
[c] Université de Montréal, Montréal, QC, Canada
[d] National Research Council Canada, Toronto, ON, Canada

ARTICLE INFO

Keywords:
EEG
Transfer learning
deep learning
Pathology Diagnosis
Data Model Scaling

ABSTRACT

Pathology diagnosis based on EEG signals and decoding brain activity holds immense importance in understanding neurological disorders. With the advancement of artificial intelligence methods and machine learning techniques, the potential for accurate data-driven diagnoses and effective treatments has grown significantly. However, applying machine learning algorithms to real-world datasets presents diverse challenges at multiple levels. The scarcity of labeled data, especially in low regime scenarios with limited availability of real patient cohorts due to high costs of recruitment, underscores the vital deployment of scaling and transfer learning techniques. In this study, we explore a real-world pathology classification task to highlight the effectiveness of data and model scaling and cross-dataset knowledge transfer. As such, we observe varying performance improvements through data scaling, indicating the need for careful evaluation and labeling. Additionally, we identify the challenges of possible *negative transfer* and emphasize the significance of some key components to overcome distribution shifts and potential spurious correlations and achieve positive transfer. We see improvement in the performance of the target model on the target (NMT) datasets by using the knowledge from the source dataset (TUAB) when a low amount of labeled data was available. Our findings demonstrated that a small and generic model (e.g. ShallowNet) performs well on a single dataset, however, a larger model (e.g. TCN) performs better in transfer learning when leveraging a larger and more diverse dataset.

1. Introduction

In recent years, significant strides have been made in the field of pathology diagnosis, propelled by advancements in brain imaging, particularly the analysis of electroencephalogram (EEG) signals and the decoding of brain activity [1,2]. The broad domain of brain science extensively employs EEG to scrutinize the electrical activities of the brain, achieved by strategically placing electrodes on the scalp's surface. These methodologies offer valuable insights into the intricacies of human brain function, furnishing critical information for the comprehension and diagnosis of diverse neurological disorders. Machine learning methodologies have played a central role in processing EEG data, fostering the creation of automated systems for pathology classification [3,4].

However, despite the remarkable progress made, the application of machine learning algorithms to real-world datasets presents a set of challenges [5]. One significant challenge arises from the scarcity of labeled data, particularly in scenarios where the number of instances is limited, commonly referred to as low regime data. In such cases, conventional machine learning techniques may struggle to achieve satisfactory performance due to the insufficient amount of data available for accurate model training [6]. The insufficient amount of data available is due to data privacy and security issues, especially when dealing with sensitive or personal data from multiple users or domains [7,8]. The scarcity of accessible EEG datasets has posed challenges in accurately estimating mental states using deep learning models [9].

To address this data scarcity issue, efforts have been made to create large-scale datasets by recording EEG signals directly from patients in hospitals [10,11]. These initiatives aim to gather diverse and comprehensive data, encompassing a wide range of neurological conditions and patient demographics. By collecting data in real-world clinical settings, these datasets capture the complexity and heterogeneity of

* Corresponding author at: Université de Montréal, Montréal, QC, Canada.
E-mail address: mohammad.bayazi@mila.quebec (M.-J. Darvishi-Bayazi).

https://doi.org/10.1016/j.compbiomed.2023.107893
Received 26 August 2023; Received in revised form 28 November 2023; Accepted 20 December 2023
Available online 30 December 2023

pathology cases encountered in actual practice. However, in some conditions, collecting more data directly from patients in hospital settings is often hindered by challenges such as high costs, time constraints, and logistical difficulties [12]. In light of these obstacles, transfer learning emerges as a compelling solution for addressing data scarcity in EEG signal analysis [13]. Leveraging pre-existing knowledge from large-scale datasets recorded in real-world clinical settings, transfer learning enhances training efficiency, accelerates learning convergence, and improves model adaptability to diverse neurological conditions.

Another challenge that arises from applying machine learning algorithms to real-world EEG datasets is the low generalization performance due to distribution shift [14] and noise [15]. Distribution shift refers to the phenomenon where the data distribution of the test set differs from that of the training set, which can lead to a degradation of the model's accuracy on unseen data [16]. Noise refers to the unwanted variations or distortions in the EEG signals, which can be caused by various factors, such as electrode placement, muscle artifacts, eye movements, or environmental interference [5]. These factors can affect the quality and reliability of the EEG data, making it difficult for machine learning models to extract meaningful features and patterns [17].

One possible solution to overcome these challenges is to use transfer learning, which is a technique that allows a model that has been trained on a source task or domain to be adapted to a new target task or domain by transferring some of its knowledge [13]. Transfer learning can leverage the information learned from a large and diverse source dataset to improve the performance of a model on a small and specific target dataset. Transfer learning significantly reduces the training time and enhances accuracy. Furthermore, this approach can effectively address the challenges of data insufficiency, variability, and inefficiency that are common in brain–computer interface applications [18]. However, transfer learning also poses some risks, such as negative transfer, which occurs when the transferred knowledge is irrelevant or harmful to the target task or domain [19]. Negative transfer can result in a worse performance than training from scratch, and it can be caused by various factors, such as domain discrepancy, task discrepancy, or label discrepancy [20]. Therefore, it is important to select an appropriate source dataset and a suitable transfer learning method for the target task or domain.

In this work, we investigated the knowledge transferability across two pathological detection datasets, The widely used Temple University Hospital Abnormal (TUAB) [11] and the new NUST-MH-TUKL (NMT) scalp EEG dataset [10], which contain thousands of recordings labeled as normal or abnormal by a team of qualified neurologists in different population and hospitals. We studied how and when a transfer of gained knowledge is possible. Furthermore, we shed light on the black box complex structure of deep neural networks via the Centered Kernel Alignment (CKA) similarity method. CKA metric evaluates the degree of similarity between the representations learned by different neural networks [21]. Our work improved the performance of the target model by leveraging knowledge from the source dataset (TUAB) to the target (NMT) datasets.

In this work:

- We investigated the knowledge transferability across two pathological detection datasets, The widely used Temple University Hospital Abnormal (TUAB) [11] and the new NUST-MH-TUKL (NMT) scalp EEG dataset [10], which contain thousands of recordings labeled as normal or abnormal by a team of qualified neurologists in different population and hospitals.
- We studied how and when a transfer of gained knowledge is possible
- We shed light on the black box complex structure of deep neural networks via the Centered Kernel Alignment (CKA) similarity method. CKA metric evaluates the degree of similarity between the representations learned by different neural networks [21].
- We improved the performance of the target model by leveraging knowledge from the source dataset to the target dataset, and showed that our model can generalize well to unseen data.

2. Materials and method

2.1. Datasets

Our experiment is based on the TUAB Abnormal EEG Corpus3 (v3.0.0) [11,22], which is a publicly available dataset of significant size that facilitates the task of decoding general pathology from EEGs. The corpus comprises 2993 recordings, each lasting at least 5 min, obtained from 2329 unique patients. It is divided into a training and a final test set. The corpus includes recordings of both male and female patients from a broad age range, spanning from 7 days to 96 years, covering infants, children, adolescents, adults, and seniors. The dataset includes various pathologies diagnosed in patients, such as epilepsy, strokes, depression, and Alzheimer's disease, among others. However, only binary labels are available (See Fig. 1A).

The NMT EEG dataset is an open-source annotated dataset of healthy and pathological EEG recordings [10]. The dataset consists of 2417 recordings from unique participants. Each recording is labeled as either normal or abnormal by a team of expert neurologists. The dataset also includes demographic information such as the gender and age of the patients. The dataset focuses on the South Asian population and can be used for training machine learning models to identify normal and abnormal EEG patterns. The dataset is intended to increase the diversity of EEG datasets and to overcome the scarcity of accurately annotated publicly available datasets for EEG research (See Fig. 1A).

The dataset is imbalanced in pathological and gender classes, in the training and statistics shift in the test sets, which may pose a challenge for machine learning algorithms. In the training set, there are 66.38% male and 33.62% female subjects, with 13.50% of the male subjects and 16.67% of the female subjects having abnormal EEG recordings. In the test set, there are 68.65% male and 31.35% female subjects, with 47.24% of the male subjects and 51.72% of the female subjects having abnormal EEG recordings.

Fig. 1A illustrates the challenge of generalization and transfer across different subsets and datasets, as the distribution of groups changes drastically in each case. For example, the NMT dataset has a lower proportion of abnormal recordings in the training than the TUAB dataset. Moreover, within each dataset, the train and evaluation subsets have different class balances, which can affect the performance and robustness of the models. Therefore, it is important to consider these factors when designing and evaluating the models for EEG analysis. Each dataset has a validation set, which we use for testing the models. We did not use any information from this set during training, to avoid data leakage. We split the training sets into 85% and 15% for training and validation, respectively. The validation set was used to select the best model based on the balanced accuracy metric. The test set was used to evaluate the final performance of the selected model on unseen data.

For the preprocessing step, We applied the same preprocessing steps to both datasets, which are TUAB and Temple. These steps are: We selected a subset of 21 electrode positions following the international 10–20 placement because these electrode positions were common to all the recordings in the datasets. We discarded the first 60 s of every recording, We used a maximum of 20 min of every recording, to avoid considerable feature generation and resampling times for exceptionally long recordings. We downsampled the EEG recordings to 100 Hz and clipped them at $\pm 800\ \mu V$ to reject unphysiologically extreme values and to ensure comparability to previous works [10,23]. We z-scored each channel of the recordings based on the channel mean and standard error, to normalize the data and reduce the effect of outliers. We also normalized each channel in each trial to range [0 1], to make the data more suitable for the neural network models.

We preprocessed both datasets in the same way, as follows. We chose 21 electrodes (Fig. 1B) that were common to all recordings in the dataset, following the international 10–20 system. We removed the first 60 s of each recording, as they contained many artifacts that could

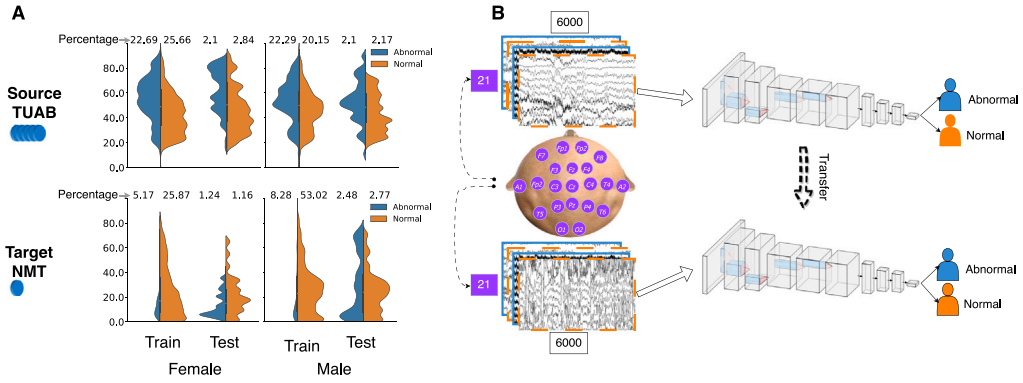

Fig. 1. Overview (A) Comparison of the recording sites (flags), number of recordings in minority groups (cylinders) and distribution with respect to age in two EEG datasets: TUAB and NMT. In the NMT dataset, the statistics between splits, gender and datasets vary drastically. The age distribution differs in data splits for gender and pathological states. The number of samples in different conditions is not balanced, which may affect the performance of the models. Cylinders are used to show that the number of recordings in minority groups in NMT is five times smaller than TUAB. (B) Our neural network model for pathological detection in EEG using transfer learning. The model transfers weights from a source domain (TUAB) to a target domain (NMT). The model leverages the knowledge from the source domain and applies it to the target domain, which may have different distributions or characteristics. The backbone could be TCN, Deep4Net, ShallowNet or EEGNet.

be due to adjusting the electrode cap or finding a comfortable position. We also limited each recording to 20 min at most, to reduce the computation time for feature extraction and resampling. We resampled the recordings to 100 Hz and clipped them at ±800 μV, following the same procedure as in Kiessner et al. [23] for TUAB and Khan et al. [10] for NMT. We applied Artifact Subspace Reconstruction (ASR) [24] to remove artifacts from the EEG recordings, based on principal component analysis (PCA) and robust statistics. We normalized each channel to its mean and standard deviation using z-score and scaled the data between zero and one.

2.2. Models

To evaluate the performance of different models, we trained several well-known EEG models while altering the number of recordings in the training set. It is worth noting that the test set remained fixed to the original split of the dataset. We experimented with five models with at least five different seeds for each of them. These experiments allowed us to thoroughly evaluate the effectiveness of data size and parameter size and select the best-performing model for decoding general pathology from EEGs. We used all models without making any changes to their original structure.

The four models that we experimented with are:

- EEGNet [25]: EEGNet is a compact convolutional neural network that can be used for EEG-based brain–computer interfaces. It is designed to work with different EEG paradigms, such as P300, ERN, MRCP, and SMR, and to be as efficient as possible. It has only three convolutional layers and 2018 trainable parameters, which makes it easy to train and deploy. It uses batch normalization and ELU activation to improve its performance and robustness.
- ShallowNet [26]: A shallow convolutional neural network that consists of two convolutional layers, each followed by a max-pooling layer and a dropout layer. Both layers use batch normalization and ELU activation. The final layer is a softmax layer that outputs the class probabilities. ShallowNet has 36 722 trainable parameters.
- Deep4Net [26]: A deep convolutional neural network that consists of four convolutional blocks, each followed by a max-pooling layer and a dropout layer. The convolutional blocks have different numbers of filters, kernel sizes, and strides, and use batch normalization and ELU activation. The final layer is a softmax layer that outputs the class probabilities. Deep4Net has 277 052 trainable parameters and was originally proposed for EEG-based emotion recognition.

Table 1
Models details.

Model name	# of parameters	Type
TCN [27]	456 502	Temporal ConvNet
Deep4Net [26]	277 052	4-layer ConvNet
ShallowNet [26]	36 722	1-layer ConvNet
EEGNet [25]	2018	Compact ConvNet

Table 2
Hyper-parameters.

Model	Drop prob	Batch size	lr	# of epochs	Weight decay
TCN	0.0527015	64	0.0011261	35	5.8373053e−07
Deep4Net	0.5	64	0.01	35	0.0005
ShallowNet	0.5	64	0.000625	35	0.0
EEGNet	0.25	64	0.001	35	0.0

- TCN-EEG [27]: A temporal convolutional network that consists of several residual blocks, each containing two dilated causal convolutional layers and different dilation rates. The residual blocks use weight normalization and ReLU activation. The final layer is a softmax layer that outputs the class probabilities. TCN-EEG has 456 502 trainable parameters.

Table 1 shows the summary of the models that we used in our experiments. It includes the name, the architecture, the number of parameters, and the original reference of each model. The models are Deep4Net, ShallowNet, TCN-EEG, and EEGNet. They are all convolutional neural networks that can be applied to EEG signal analysis. However, they differ in their structure, complexity, and specificity. Some of them are more generic and adaptable, while others are more tailored and efficient for specific EEG tasks or domains.

2.3. Training

We followed the cropped training approach proposed by [26], which divides the EEG signals into equally sized, maximally overlapping segments that match the receptive field of the networks. The receptive field is the number of signal samples that the networks can process at a time, which varies depending on the network architecture. All networks, except the TCN, had a receptive field of approximately 600 samples, while the TCN had a receptive field of approximately 900 samples. We used the AdamW optimizer [28] to optimize the categorical cross-entropy loss function, which measures the difference between the predicted and true class probabilities. AdamW is an improved version of Adam [29] that decouples the weight decay updates and the

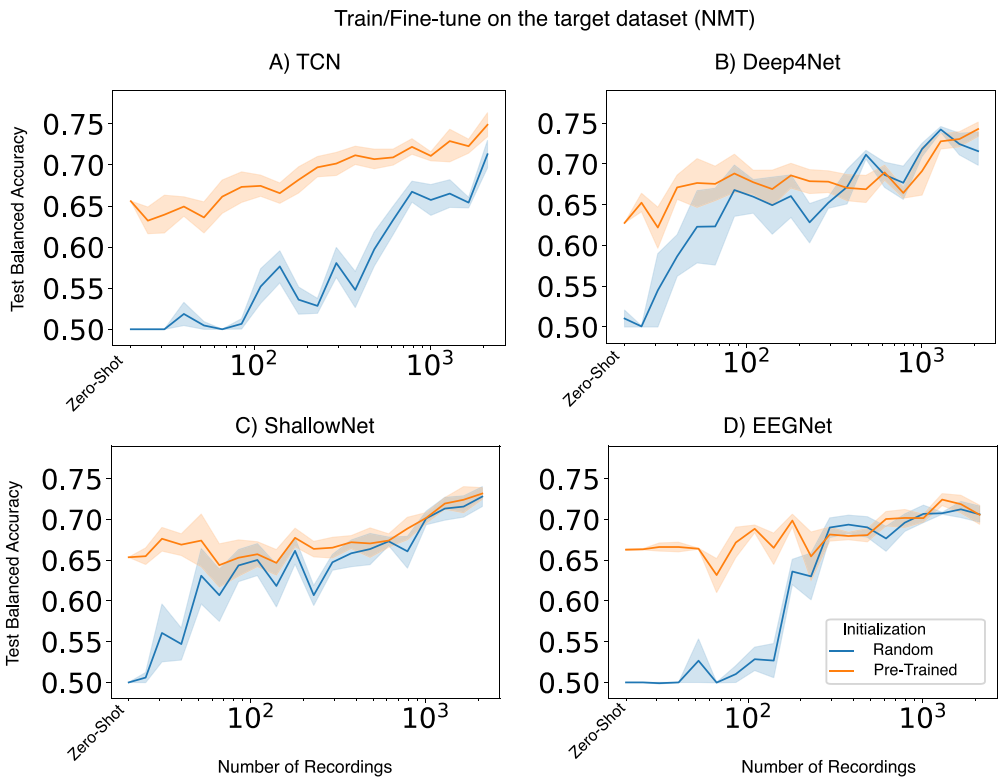

Fig. 2. The effect of pretraining on pathology detection balanced accuracy in the NMT dataset. The x-axis shows the number of training samples and the y-axis shows the test balanced-accuracy. The orange line represents the performance of fine-tuning a pretrained model and the blue line represents the baseline model trained from scratch. The error bars show the standard error of the mean across five random seeds.

loss function optimization, which leads to better generalization [28]. We used cosine annealing [30] to schedule the learning rates for both the gradient and weight decay updates. Cosine annealing is a technique that gradually reduces the learning rates according to a cosine function, which helps to avoid local minima and find better solutions. To have an unbiased selection of hyper-parameters (see Table 2), we followed the same experimental setup as in [23], which is a recent work on EEG pathology detection using deep learning. We compared four different neural network architectures: EEGNet, DeepConvNet, ShallowConvNet, and TCN. We trained each model from scratch on the source dataset and fine-tuned pretrained models on the target dataset. We evaluated each model on both TUAB and NMT datasets using balanced accuracy as the performance metric.

2.4. Evaluation

We evaluated our models on the predefined test set of each dataset, which was not used for training or validation. We used 15% of the training set for early stopping and model selection. We used balanced accuracy [31,32] as the evaluation metric, to compare the performance across datasets, as one of them was imbalanced. Balanced accuracy is defined as the average of the recall obtained in each class, and it is equal to the area under the ROC curve for a binary classification problem. It is a useful metric when the classes are imbalanced, as it gives equal weight to each class, regardless of its frequency.

Balanced accuracy is calculated as:

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{1}$$

where:

- Sensitivity: The "true positive rate"
- Specificity: The "true negative rate"

This formula can also be written as:

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \tag{2}$$

where:

- TP: True Positives
- FN: False Negatives
- TN: True Negatives
- FP: False Positives

2.5. Data augmentation

We used data augmentation for EEG based on this previous work [33], which proposed and compared different methods of augmenting EEG data to improve the performance and robustness of machine learning models. We used four methods of data augmentation, each with a probability of 0.1 in training all models:

- SignFlip: This method randomly flips the sign of the EEG signals, which simulates a change in the polarity of the electrodes.
- ChannelsDropout: This method randomly drops out some channels of the EEG signals, which simulates a loss of contact or a malfunction of the electrodes. We used a dropout probability of 0.2.
- FrequencyShift: This method randomly shifts the frequency spectrum of the EEG signals, which simulates a change in the sampling rate or drift in the frequency band. We used a maximum frequency shift of 2 Hz.
- SmoothTimeMask: This method randomly masks some time segments of the EEG signals with a smooth transition, which simulates a temporary occlusion or a distortion of the signals. We used a mask length of 600 samples.
- BandstopFilter: This method randomly applies a band-stop filter to the EEG signals, which removes a narrow frequency band from the signals. This simulates a noise reduction or a notch filter. We used a bandwidth of 1 Hz.

- ChannelsShuffle: This method randomly shuffles the order of the channels of the EEG signals, which changes the spatial configuration of the signals. This simulates a different electrode placement or a permutation of the channels. We used a shuffle probability of 0.2.

2.6. Transfer learning

Transfer learning is a technique that allows a model that has been trained on a source task or domain to be adapted to a new target task or domain by transferring some of its knowledge [34]. Fine-tuning is a common method of transfer learning, which involves adjusting the parameters of the model on the target data while keeping the initial weights from the source data. Transfer learning and fine-tuning can help to improve the performance and generalization of the model, especially when the target data is limited and similar to the source data [13].

In this work, we applied transfer learning and fine-tuning to our EEG models, which were trained on two different datasets: TUAB and NMT. TUAB is a large and diverse dataset of EEG recordings from normal and abnormal subjects, while NMT is a smaller and more specific dataset of EEG recordings from normal and pathological subjects. We used TUAB as the source dataset and NMT as the target dataset, and we experimented with four models: Deep4Net, ShallowNet, TCN-EEG and EEGNet.

We trained the models on the whole source dataset (TUAB) and used the weights of that model to initialize the network to train on the second dataset (NMT). This way, we hoped to leverage the knowledge learned from TUAB to improve the performance on NMT. To have a baseline, we also trained the same models with random initialization on the target dataset (NMT), without using any information from TUAB. We compared these two conditions (fine-tuned the pre-trained model vs. randomly initialized model) on the test set of NMT, using balanced accuracy as our evaluation metric.

2.7. Representation similarity

Cianfarani et al. [35] provides a novel perspective on how robust training methods affect the internal representations learned by neural networks. The authors use representation similarity metrics, such as centered kernel alignment (CKA) [21], to compare the representations of different layers of robust and non-robust networks on three vision datasets. They reveal some interesting properties of robust representations, such as less specialization, more homogeneity and higher overfitting. Inspired by this work, we experimented with different scenarios to assess the transferability between these two datasets and models trained on them. We wanted to investigate how the similarity of representations influences the ability of a model to adapt to a new domain or task.

3. Results

3.1. Transfer learning

The results in Fig. 2 indicate that pretraining enhances pathology detection performance, particularly in the low data regime (<100 samples). This is crucial due to the low sample size in clinical studies, which poses a challenge for deep learning approaches. Our results suggest that the transfer is feasible if several criteria are satisfied, such as data normalization and adequate preprocessing.

The results in Fig. 3 show that the pretrained model does not maintain high performance on the source dataset while adapting to the target dataset. The performance of the baseline model on the source dataset increases a bit as it trains on the target dataset. The results also show a sign of catastrophic forgetting in these real-world datasets.

Table 3
Performance of models on different distributions. The highlighted numbers in bold indicate the best performance in each scenario and underlined text means that the performance of the model was very close to the best performance in that scenario (12 scenarios). PtTUAB > FtNMT means Pre-train a model on TUAB and Fine-tuned on NMT.

Train		Test on (BAC)		
Model	Train on	TUAB	NMT	TUAB+NMT
TCN	TUAB	81.69 ± 0.35	**63.04 ± 2.00**	74.71 ± 0.62
	NMT	58.16 ± 9.90	70.34 ± 4.79	63.42 ± 7.48
	TUAB+NMT	80.92 ± 0.73	**73.53 ± 1.92**	**78.04 ± 1.19**
	PtTUAB > FtNMT	65.92 ± 4.41	**73.70 ± 3.32**	69.35 ± 2.50
Deep4Net	TUAB	81.64 ± 0.52	62.77 ± 1.63	74.82 ± 0.81
	NMT	63.00 ± 5.40	71.56 ± 4.87	66.88 ± 5.17
	TUAB+NMT	80.90 ± 1.12	72.95 ± 1.34	77.51 ± 0.77
	PtTUAB > FtNMT	59.00 ± 5.06	73.34 ± 3.32	65.18 ± 3.81
ShallowNet	TUAB	**82.40 ± 0.74**	62.02 ± 2.01	**74.96 ± 1.18**
	NMT	**66.27 ± 1.60**	**72.81 ± 2.85**	**69.22 ± 0.92**
	TUAB+NMT	**80.98 ± 1.14**	71.52 ± 1.23	77.09 ± 0.55
	PtTUAB > FtNMT	**66.82 ± 0.37**	73.18 ± 1.61	**69.73 ± 0.73**
EEGNet	TUAB	81.40 ± 0.60	61.45 ± 4.22	74.19 ± 1.73
	NMT	62.09 ± 6.57	70.63 ± 2.36	66.02 ± 4.51
	TUAB+NMT	80.90 ± 1.28	71.52 ± 1.23	75.85 ± 1.52
	PtTUAB > FtNMT	65.82 ± 0.97	69.71 ± 2.98	67.82 ± 1.75

To enable a model to work on both datasets, we opt to train a model on the concatenate and train on both datasets as suggested in [19]. However, this approach faces the challenges of class imbalance and dataset bias, which affect the model's performance and generalization. To address these challenges, we designed a data sampler that samples uniformly from each dataset and each class, ensuring a balanced and fair representation of the data. The performance of the resulting network is reported as follows:

Table 3 shows the performance of the models under different training and test distributions. We can observe that the models trained on TUAB, which is a large and diverse dataset, have a high balanced accuracy (BAC) on TUAB itself, but a low BAC on NMT, which is a small and noisy dataset. This indicates that the models suffer from distribution shifts and overfitting when they encounter data from a different domain. For example, when we train on TUAB, the best performance on the TUAB dataset is 82.40 and it drops by 20.38 to 62.02 when we test on NMT, which is out of its distribution. On the other hand, the best BAC on NMT is 73.70, which we assume is in distribution accuracy. So the difference between a classifier that has not seen any data from the NMT dataset and one that has seen the whole NMT data is 11.68. This shows that there is still room for improvement in transferring knowledge from TUAB to NMT.

Table 3 also shows the performance of models that have been trained on the concatenation of both datasets, TUAB and NMT. We can see that these models have comparable BAC on both datasets compared to models trained on a single dataset. It was expected that combining data from different domains could improve the robustness and generalization of the models, as they can learn from a larger and more diverse set of examples. However, these models have a lower BAC than the models fine-tuned with our proposed framework, which indicates that simply concatenating data is not enough to achieve optimal transfer learning. We also see that as the model is bigger, it can get better BAC on the concatenated data, since the model has more capacity. For example, TCN, which has the largest number of parameters among the four architectures, has the highest BAC on both TUAB and NMT when trained on the concatenated data and ShallowNet which is a small and simple network that has the best accuracies in single datasets training. This implies that larger models can benefit more from combining data from different domains, as they can learn more complex and diverse features.

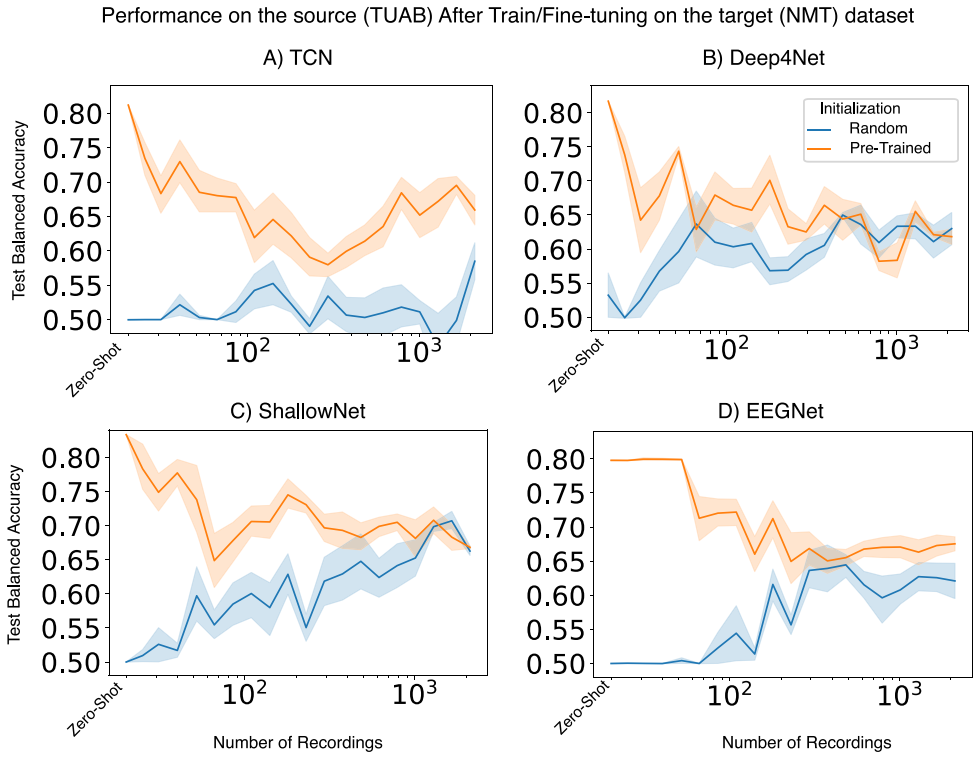

Fig. 3. The performance on the source dataset (TUAB) while finetuning or training on the target dataset (NMT). The x-axis shows the number of training samples from the target dataset and the y-axis shows the balanced accuracy of the source dataset. The orange line represents the pretrained model that is finetuned on the target dataset and the blue line represents the baseline model that is trained from scratch on the target dataset.

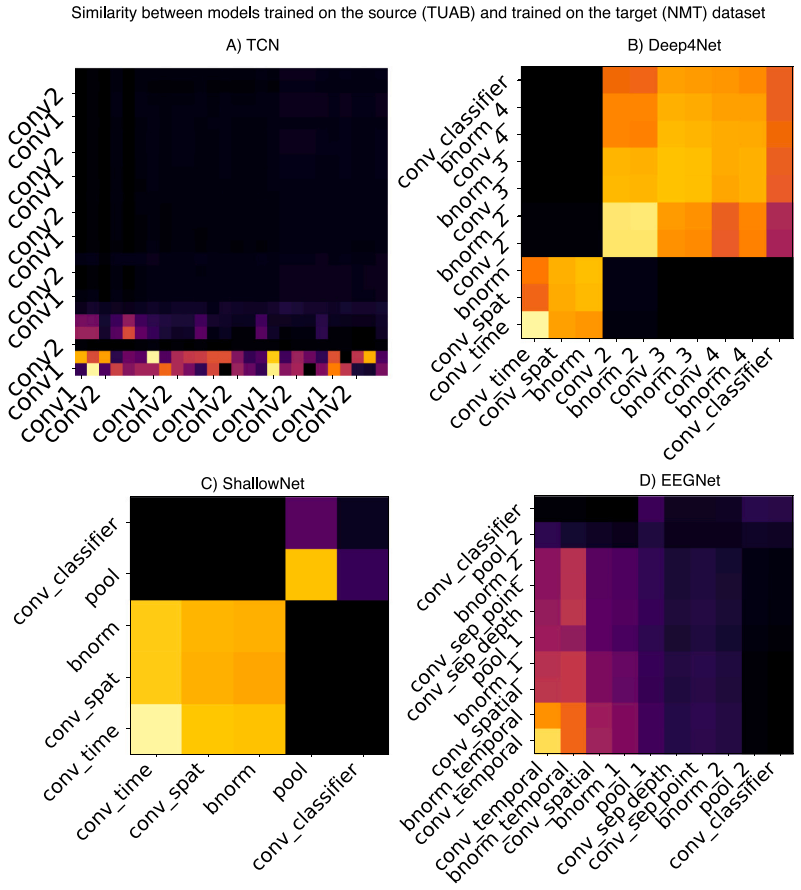

Fig. 4. Cross-model similarity between different layers of networks trained on the source (TUAB) and trained on the target (NMT) dataset.

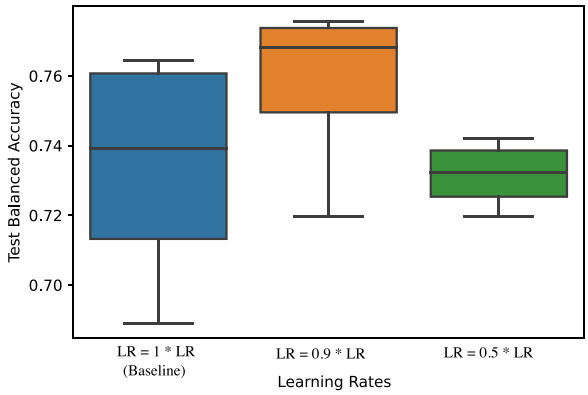

Fig. 5. The effect of discriminative fine-tuning on the balanced accuracy (BAC) of the model for the NMT dataset. The *x*-axis shows Discriminative Learning Rates, and the *y*-axis shows the BAC score. The figure shows that the model with discriminative fine-tuning achieves a higher BAC score than the baseline model, indicating that it can better adapt to the new domain and task.

3.2. Representation similarity

Fig. 4 shows the degree of similarity between different layers of networks trained on TUAB and NMT datasets. The similarity between TUAB and NMT networks decreases from the early layers to the last layers, as shown in Fig. 4, implying that the datasets have different representations at higher levels of abstraction. In the TCN, we see that the first block is similar but the other blocks are different which implies that features on the blocks other than the first block are very different between the two networks. These results demonstrate that the last layer of the network needs to have specific features, while the earlier layers can have general features.

3.3. Discriminative fine-tuning

Based on insights in Section 3.2, we hypothesize if we adjust the learning rate during the fine-tuning process from the last layer to the early layers, giving more weight to the lower layers that are more similar across datasets and less weight to the higher layers that are more divergent would improve the transfer performance. In the literature, it is called Discriminative Fine-Tuning. Discriminative Fine-Tuning is a technique for transferring knowledge from a pretrained model to a downstream task. It involves using different learning rates for different layers of the model, depending on how relevant they are to the task. The idea is to fine-tune the lower layers more slowly since they capture general features, and the higher layers more quickly since they capture task-specific features [36]. It has been shown that it works for MRI fine-tuning [37]. Here as our results confirm these assumptions, we fine-tune the model using the discriminative fine-tuning technique. This allows us to leverage the pretrained knowledge from the TUAB dataset and adapt it to the NMT dataset, which has a smaller size and a different domain. We compare the performance of our model with different learning rate schedules and report the results in Fig. 5.

Fig. 5 shows the impact of different learning rate decay factors on the performance of the model. The learning rate decay factor is the ratio between the learning rates of two consecutive layers. The *x*-axis shows the learning rate decay factor, and the *y*-axis shows the balanced accuracy (BAC) score. The figure shows that the model with discriminative fine-tuning achieves a higher BAC score than the baseline model when the learning rate decay factor is 0.9, indicating that it can better adapt to the new domain and task. However, when the learning rate decay factor is 0.5, the performance of the model with discriminative fine-tuning drops significantly, suggesting that it is too aggressive and causes instability in the training process. Therefore, choosing an appropriate learning rate decay factor is important for discriminative fine-tuning, as it adds a new hyperparameter to the model.

3.4. Pre-training (Why large neural networks?)

After examining the results in Fig. 6A, we observe that all models perform similarly and reach a consistent saturation accuracy of around 80% but showcase different slopes. For example, Shallownet excelled over others under the low data regime and with fewer than 200 recordings. However, in situations with a higher number of recordings, larger networks such as TCN outperformed other models. Regarding out-of-distribution data in Fig. 6B, Deep4Net and TCN outperform the smaller network EEGNet. These larger networks demonstrate superior generalization capabilities and increased capacity to adapt to new tasks, as illustrated in Fig. 2.

3.5. Model and data size

In this paper, we address the problem of EEG classification and transfer learning using several well-known architectures in the literature. For the ablation studies, we opt to explore temporal convolutional networks (TCN) that capture both spatial and temporal patterns because of their flexibility to change the number of blocks and filters and, therefore, the number of model parameters and depth. Also, TCN has been shown to achieve superior performance for EEG time series analysis compared to recurrent neural networks (RNN) and feed-forward networks (FFN) [38]. However, TCN also requires careful preprocessing and architecture design to improve its generalization ability and data efficiency.

In this section, we compare TCN models with different numbers of filters in one layer. We evaluate the models on in-distribution (TUAB) and out-of-distribution (NMT) test sets, using balanced accuracy as the metric. Fig. 7 shows the results of our comparison.

Fig. 7 illustrates that increasing the number of filters in one layer TCN model improves the performance on both test sets. The *x*-axis shows the number of recordings, the *y*-axis shows the balanced accuracy of our TCN model and the colors are the number of filters (4, 8, 16, 32, 64). The error bars are the standard error of five runs. We can see that more filters lead to higher accuracy on both test sets. This suggests that more filters can capture more spatial and temporal features from the EEG data, and enhance the representation learning ability of the TCN model. However, more filters also increase the model complexity and computational cost. Therefore, we need to find a trade-off between performance and efficiency.

In addition to the number of filters, we also explore the effect of the number of blocks in the TCN model. We fix the number of filters to 8 and vary the number of blocks from 1 to 6. We evaluate the models on in-distribution (TUAB) and out-of-distribution (NMT) test sets, using balanced accuracy as the metric. Fig. 8 shows the results of our exploration.

Fig. 8 demonstrates that deeper networks are not different from shallow networks in terms of performance on both test sets. The *x*-axis shows the number of recordings, the *y*-axis shows the balanced accuracy of our TCN model and the colors are the number of blocks (1, 2, 3, 4, 5 or 6). The error bars are the standard error of five runs. We can see that there is no notable difference between the models with different numbers of blocks on both test sets. This indicates that increasing the depth of the TCN model does not improve its representation learning ability or generalization ability. Therefore, we conclude that shallow networks are sufficient and efficient for the pathology detection task and width is more important than the depth of the network.

4. Discussion

The main contribution of our work is to demonstrate the effectiveness of transfer learning for EEG pathology detection, especially when the target dataset is small and specific. We showed that by leveraging a larger and more diverse source dataset, we can improve the performance of the target model on the NMT dataset. We also

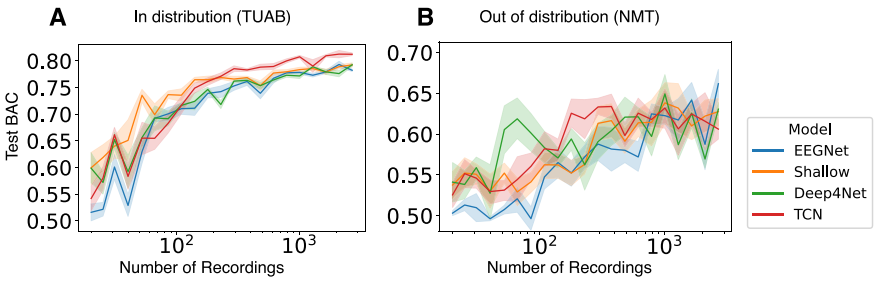

Fig. 6. Scaling behavior of different EEG models on the TUAB (Source) and NMT (Target) dataset. It illustrates the balanced accuracy plotted against the number of recordings in the training set for various models. The accuracy increases as the number of recordings increases, but the curve is saturated after reaching 1000 recordings. Larger models show better performance when more data is available. The error bars represent the standard error.

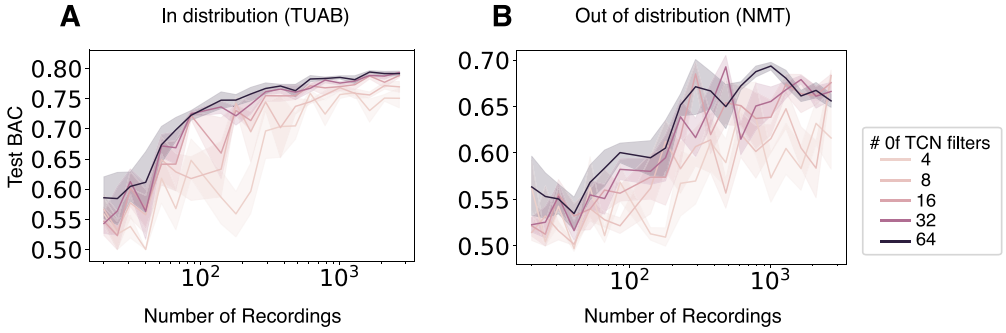

Fig. 7. Effect of the number of filters in the blocks of TCN architecture on in-distribution and out-of-distribution balanced accuracy. The x-axis shows the number of recordings, and the y-axis shows the balanced accuracy of our TCN model. Colors show the number of filters. The error bars show the standard error across five runs. We can see increasing the number of filters improves the performance of our model on both test sets.

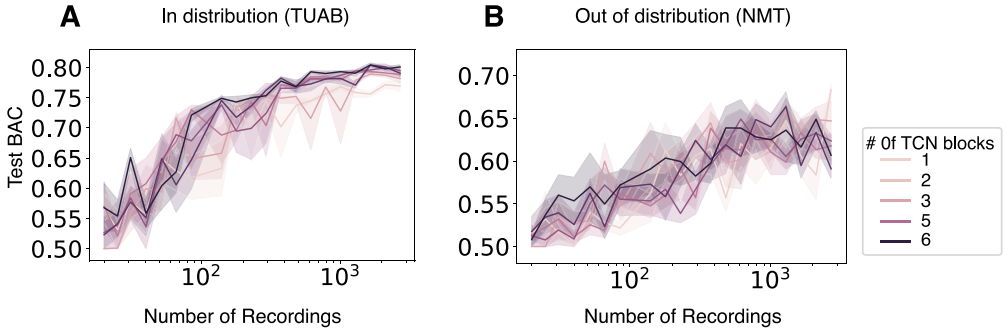

Fig. 8. Effect of the number of blocks in the blocks of TCN architecture on in-distribution and out-of-distribution balanced accuracy. The x-axis shows the number of recordings and the y-axis shows the balanced accuracy of our TCN model. Colors show the number of blocks. The error bars show the standard error across five runs. We can see increasing the number of filters improves the performance of our model on both test sets.

showed that the choice of the model architecture and the transfer learning method are crucial for the success of the transfer learning approach. We found that a larger model (TCN) can benefit more from transfer learning than a smaller model (ShallowNet). Our results in Table 3 and Fig. 2 show that transfer learning is more effective when the target dataset is small and specific, but its advantage diminishes when the full datasets are used.

In Fig. 6 all models show a saturation accuracy of around 80% in the source task (TUAB) which may have arisen from label noise or inter-rater agreement, as discussed in previous literature [23,39]. Based on our findings, we can conclude that approximately 500 recordings are sufficient for this task, and adding more recordings beyond 500 until the total of 2700 did not significantly improve the performance of the models. These results suggest that adding more data beyond a certain point may not always yield better results, and careful consideration of the amount of data used in training is essential for optimal model performance. Furthermore, we also identified that there are 54 common subjects in the two pathological groups (normal/abnormal) in the TUAB dataset. There are 149 out of 2993 rows that have their pathological status changed, some of them within a few days. This could be one of the possible sources of label noise that affects the model performance and generalization. This also might be due to noisy input data as it is a problem in other medical domains (e.g. magnetic resonance imaging [40]), which can be mitigated by transfer learning [37].

Gemein et al. [39] developed a feature-based decoding framework for analyzing EEG data and demonstrates that it achieves accuracies similar to state-of-the-art deep neural networks. The accuracies obtained ranged from 81 to 86% across both approaches, and analysis showed that they used similar aspects of the data, such as delta and theta band power at temporal electrode locations. The study argues that the accuracies of current binary EEG pathology decoders could saturate near 90% due to imperfect inter-rater agreement of clinical labels and that such decoders are already clinically useful where experts are rare. Our results confirm these findings, meaning that we also observe a saturation accuracy of around 80% for all the networks when trained on the TUAB dataset. However, we also notice that although

all the networks achieve similar accuracy when trained on all available recordings, their scaling and transfer behavior are different.

The difference between the reported accuracy in our results and the previous results comes from the fact that we chose the best model based on a validation set and reported results on the predefined test set, while they might have used Oracle model selection [16]. Oracle model selection is a method of choosing the best model based on the test set performance, which can lead to overfitting and optimistic results. By using a validation set, we avoid this problem and ensure that our results are more realistic and generalizable. We also report balanced accuracy (BAcc) as our evaluation metric, which is the average of the recall obtained in each class. BAcc is a suitable metric for ML when the goal is to minimize the overall classification error, especially when the classes are imbalanced [41].

It is worth noting that there have been several efforts to extend clinical EEG datasets with a larger number of recordings. For instance, a recent study proposed a dataset [23] with 15,300 automatically labeled EEG recordings and evaluated and improved the performance of machine learning algorithms on this dataset. However, our results in Section 3.4 suggest that for this specific task of general pathology decoding from EEGs, we do not need more than 500 recordings to achieve good model performance. This may be due to the richness and diversity of the TUAB Abnormal EEG Corpus3 (v3.0.0), which contains recordings from patients of various ages and with different pathologies. However, more recording in the source domain might improve the pretrained model. Further research could investigate whether this conclusion holds for other types of EEG analysis tasks or if more data is necessary for different contexts. Our findings suggest that we might be in the early stages of the broken scaling laws [42] where improvements are minimal, but significant improvements could be achieved beyond a certain point, which is consistent with the non-monotonic transitions observed in neural scaling behavior.

In practical situations, any automated screening algorithm is highly likely to encounter data from sources it has not seen before. Therefore, it is crucial to assess how well different architectures can handle variations in the devices and sources used for data acquisition. Ideally, deep learning algorithms should consistently perform well across different datasets. However, this is not always the case. Evaluating the performance of abnormal EEG detection across datasets has been difficult due to the lack of publicly available datasets specific to this problem. To address this, In [10] they tested the generalization performance of several CNN-based architectures. They trained them on the TUAB dataset and then evaluated their performance on the NMT dataset. They reported that there was a noticeable decline in performance, with the Shallow CNN architecture achieving an accuracy of 45% and an AUC of 0.48, and the Deep CNN architecture achieving an accuracy of 48% and an AUC of 0.46. Similar performance degradation was observed when these architectures were trained on the NMT dataset and tested on the TUAB dataset. These results emphasize the importance of collecting diverse datasets from multiple sources, as algorithms trained on data from a single source struggle to generalize well in the case of EEG data. In contrast to the challenges mentioned earlier, our findings provide encouraging evidence that the two datasets, TUAB and NMT, exhibit transferability and perform well in the context of our study. Despite the anticipated performance degradation when applying the Deep and Shallow CNN architectures trained on one dataset to the other, we observed that these models still achieved satisfactory results. Specifically, the TCN architecture attained the best accuracy of $63.04\% \pm 2.00$ when tested on the NMT dataset, and the performance degradation was 18.65%. These findings suggest that, to some extent, the models trained on one dataset were able to generalize to the other dataset and improved with fine-tuning. However, finetuning may not always be an absolute requirement, and probing methods are sufficient when datasets are within the distribution [43]. The improvement after fine-tuning with discriminative learning rate in our results implies a divergence in distribution between the two datasets, an observation that might be unexpected for some other datasets.

The results in Fig. 3 indicate that the models tend to forget the previous knowledge when they are fine-tuned on a new task or domain. This is a common problem in ML known as catastrophic forgetting, which limits the ability of the models to learn from sequential or diverse data. This problem poses a challenge and an opportunity for the fields of continual learning and out-of-distribution generalization, which aim to develop methods and techniques that can enable the models to learn continuously from new data without forgetting the old ones and to generalize well to unseen or novel data that differ from the training data. These fields can benefit from exploring these real-world datasets, which can provide realistic and complex scenarios for testing and improving their methods and techniques.

Several studies investigated the importance of depth and width of neural networks [44–46]. Our results in Figs. 8 and 7 reveal that increasing the number of filters improves the performance of the TCN model on either test set but increasing the number of blocks does not improve. This suggests that the depth of the network is not a critical factor for the representation learning or generalization ability of the TCN model in this task. Therefore, we can conclude that shallow networks are sufficient and efficient for the pathology detection task and that the width of the network is more important than the depth. This might imply that the function between pathology and the EEG signals can be approximated better with one Block of TCN which includes two CNN layers followed by the Relu function and skip connections (See Fig. A.9 for more detail).

5. Conclusion and outlook

In this work, we presented an approach for EEG pathology detection based on transfer learning. Our methodology has the potential to improve the transferability performance of neural network models with noisy and small sample size datasets by leveraging knowledge from large and diverse datasets. We conducted extensive experiments on two real-world EEG datasets, TUAB and NMT. We provide insights into the transferability of neural network features for EEG pathology detection by analyzing and visualizing the CKA similarity scores of the feature maps. Our work has several implications and contributions to the field of EEG pathology detection and transfer learning. We demonstrated that transfer learning emerges as a potent methodology to tackle the challenge of data scarcity issue that hampers the application of machine learning algorithms to real-world medical datasets. We applied CKA as a novel and effective measure of similarity between neural network representations of EEG signals, that can serve as a guiding tool for the transfer learning procedure and for elucidating the interpretability and explainability of the learned features. We proposed a simple yet highly efficient method to identify the optimal transfer point through the CKA similarity scores. This method has a huge potential to enhance model generalization and robustness significantly. We provided insights into the scaling and transfer aspects of the networks, that enable researchers and practitioners in choosing the optimal network architecture and data size for their EEG analysis tasks.

In this work, we explored the use of transfer learning for EEG pathology detection, a challenging task that requires robust and generalizable neural network models. We proposed a methodology that leverages knowledge from large and diverse datasets to improve the transferability performance of neural network models on noisy and small sampler-size datasets. We evaluated our methodology on two real-world EEG datasets, TUAB and NMT, and obtained promising results.

Our work has several implications and contributions to the field of EEG pathology detection and transfer learning. We demonstrated that transfer learning is a powerful technique to overcome the data scarcity issue that limits the application of machine learning algorithms to real-world medical datasets. We introduced CKA as an effective measure

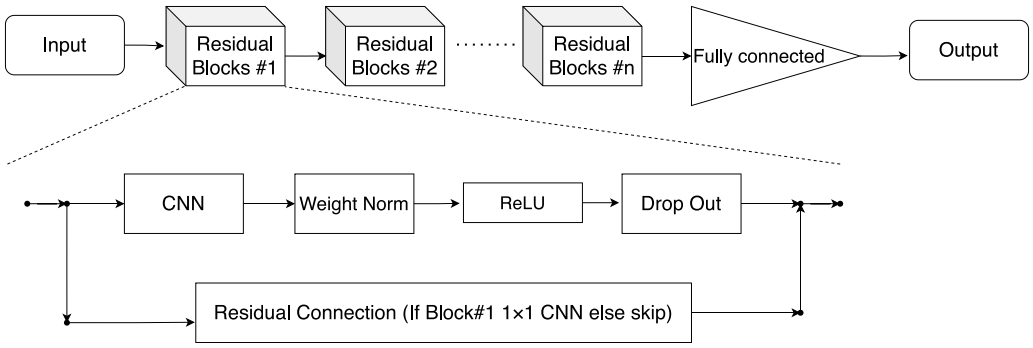

Fig. A.9. TCN architecture. A TCN network consists of several convolutional layers with residual connections. Each layer has the same input and output lengths.

of similarity between neural network representations of EEG signals, which can serve as a guiding tool for the transfer learning procedure and for enhancing the interpretability and explainability of the learned features. We proposed a simple yet highly efficient method to identify the optimal transfer point through the CKA similarity scores, which can significantly improve model generalization and robustness. We also provided insights into the scaling and transfer aspects of the networks, which can help researchers and practitioners in choosing the optimal network architecture and data size for their EEG analysis tasks.

We believe that our work opens up new avenues for further research and development in the field of EEG pathology detection and transfer learning. Future work could extend our methodology to other scenarios of EEG signals, such as single channels or methods that could handle different data properties. Future work could also explore more advanced techniques for feature extraction, representation learning, and domain adaptation, such as graph neural networks, self-supervised learning, or adversarial learning. We hope that our work will inspire more research on transfer learning for EEG pathology detection, and ultimately contribute to improving the diagnosis and treatment of neurological disorders.

CRediT authorship contribution statement

Mohammad-Javad Darvishi-Bayazi: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Mohammad Sajjad Ghaemi:** Conceptualization, Validation, Writing – review & editing. **Timothee Lesort:** Conceptualization, Investigation, Writing – review & editing. **Md. Rifat Arefin:** Conceptualization, Investigation, Validation, Writing – review & editing. **Jocelyn Faubert:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing. **Irina Rish:** Conceptualization, Funding acquisition, Investigation, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

None Declared.

Acknowledgments

This work was funded by Canada CIFAR AI Chair Program and from the Canada Excellence Research Chairs (CERC) program, National Research Council Canada, Natural Sciences and Engineering Research Council (NSERC-CAE-CRIAC-CARIQ, NSERC discovery grant RGPIN-2022-05122), Doctoral Research Microsoft Diversity Award (Microsoft-Mila), Faculty of medicine-UdeM, and Faculté des études supérieures et postdoctorales. We thank Compute Canada for providing computational resources.

Appendix. TCN model

In this paper, we use TCN-EEG, A temporal convolutional network that consists of several residual blocks, each containing two dilated causal convolutional layers and different dilation rates. The residual blocks use weight normalization and ReLU activation. The final layer is a softmax layer that outputs the class probabilities. TCN-EEG has 456502 trainable parameters. See Fig. A.9 for more detail. This architecture allows us to change the number of blocks and the number of filters in each block. .

References

[1] H. Albaqami, G.M. Hassan, A. Datta, Automatic detection of abnormal EEG signals using WaveNet and LSTM, Sensors 23 (13) (2023) 5960.
[2] Y. Zhong, H. Wei, L. Chen, T. Wu, Automated EEG pathology detection based on significant feature extraction and selection, Mathematics 11 (7) (2023) 1619.
[3] A. Craik, Y. He, J.L. Contreras-Vidal, Deep learning for electroencephalogram (EEG) classification tasks: a review, J. Neural Eng. 16 (3) (2019) 031001.
[4] S. Nakagome, A. Craik, A. Sujatha Ravindran, Y. He, J.G. Cruz-Garza, J.L. Contreras-Vidal, Deep learning methods for EEG neural classification, in: Handbook of Neuroengineering, Springer, 2022, pp. 1–39.
[5] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural Eng. 16 (5) (2019) 051001.
[6] A. Adadi, A survey on data-efficient algorithms in big data era, J. Big Data 8 (1) (2021) 24.
[7] W. Zhang, X. Li, Data privacy preserving federated transfer learning in machinery fault diagnostics using prior distributions, Struct. Health Monit. 21 (4) (2022) 1329–1344.
[8] X. Li, S. Yu, Y. Lei, N. Li, B. Yang, Intelligent machinery fault diagnosis with event-based camera, IEEE Trans. Ind. Inform. (2023).
[9] S. Dutta, A. Nandy, Data augmentation for ambulatory EEG based cognitive state taxonomy system with RNN-LSTM, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2019, pp. 468–473.
[10] H.A. Khan, R. Ul Ain, A.M. Kamboh, H.T. Butt, S. Shafait, W. Alamgir, D. Stricker, F. Shafait, The NMT scalp EEG dataset: an open-source annotated dataset of healthy and pathological EEG recordings for predictive modeling, Front. Neurosci. 15 (2022) 755817.
[11] I. Obeid, J. Picone, The temple university hospital EEG data corpus, Front. Neurosci. 10 (2016) 196.
[12] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big data 3 (1) (2016) 1–40.
[13] Z. Wan, R. Yang, M. Huang, N. Zeng, X. Liu, A review on transfer learning in EEG signal analysis, Neurocomputing 421 (2021) 1–14.
[14] Z. Lan, O. Sourina, L. Wang, R. Scherer, G.R. Müller-Putz, Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets, IEEE Trans. Cogn. Dev. Syst. 11 (1) (2018) 85–94.
[15] C. He, J. Liu, Y. Zhu, W. Du, Data augmentation for deep neural networks model in EEG classification task: a review, Front. Hum. Neurosci. 15 (2021) 765525.
[16] I. Gulrajani, D. Lopez-Paz, In search of lost domain generalization, in: International Conference on Learning Representations, 2020.
[17] H. Banville, S.U. Wood, C. Aimone, D.-A. Engemann, A. Gramfort, Robust learning from corrupted EEG with dynamic spatial filtering, NeuroImage 251 (2022) 118994.
[18] C.-Y. Yang, P.-C. Chen, W.-C. Huang, Cross-domain transfer of EEG to EEG or ECG learning for cnn classification models, Sensors 23 (5) (2023) 2458.

[19] W. Zhang, L. Deng, L. Zhang, D. Wu, A survey on negative transfer, IEEE/CAA J. Autom. Sin. 10 (2) (2022) 305–329.
[20] D. Tran, J. Liu, M.W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu, et al., Plex: Towards reliability using pretrained large model extensions, 2022, arXiv preprint arXiv:2207.07411.
[21] S. Kornblith, M. Norouzi, H. Lee, G. Hinton, Similarity of neural network representations revisited, in: International Conference on Machine Learning, PMLR, 2019, pp. 3519–3529.
[22] N. Shawki, M.G. Shadin, T. Elseify, L. Jakielaszek, T. Farkas, Y. Persidsky, N. Jhala, I. Obeid, J. Picone, Correction to: The temple university hospital digital pathology corpus, in: Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications, Springer, 2022, p. C1.
[23] A.-K. Kiessner, R.T. Schirrmeister, L.A. Gemein, J. Boedecker, T. Ball, An extended clinical EEG dataset with 15,300 automatically labelled recordings for pathology decoding, NeuroImage: Clin. (2023) 103482.
[24] S. Blum, N.S. Jacobsen, M.G. Bleichner, S. Debener, A Riemannian modification of artifact subspace reconstruction for eeg artifact handling, Front. Hum. Neurosci. 13 (2019) 141.
[25] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, Eegnet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.
[26] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420.
[27] T. Elsken, J.H. Metzen, F. Hutter, Neural architecture search: A survey, J. Mach. Learn. Res. 20 (1) (2019) 1997–2017.
[28] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv:1711.05101.
[29] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[30] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, 2016, arXiv preprint arXiv:1608.03983.
[31] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 3121–3124.
[32] J.D. Kelleher, B. Mac Namee, A. D'arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies, MIT Press, 2020.
[33] C. Rommel, J. Paillard, T. Moreau, A. Gramfort, Data augmentation for learning predictive models on EEG: a systematic comparison, J. Neural Eng. 19 (6) (2022) 066020.
[34] H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Transfer learning for time series classification, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 1367–1376.
[35] C. Cianfarani, A.N. Bhagoji, V. Sehwag, B. Zhao, H. Zheng, P. Mittal, Understanding robust learning through the lens of representation similarities, Adv. Neural Inf. Process. Syst. 35 (2022) 34912–34925.
[36] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, 2018, arXiv preprint arXiv:1801.06146.
[37] N. Malik, D. Bzdok, From YouTube to the brain: Transfer learning can improve brain-imaging predictions with deep learning, Neural Netw. 153 (2022) 325–338.
[38] D. Walther, J. Viehweg, J. Haueisen, P. Mäder, A systematic comparison of deep learning methods for EEG time series analysis, Front. Neuroinform. 17 (2023) 1067095.
[39] L.A. Gemein, R.T. Schirrmeister, P. Chrabąszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, T. Ball, Machine-learning-based diagnostics of EEG pathology, NeuroImage 220 (2020) 117021.
[40] M.-A. Schulz, B.T. Yeo, J.T. Vogelstein, J. Mourao-Miranada, J.N. Kather, K. Kording, B. Richards, D. Bzdok, Different scaling of linear models and deep learning in UKbiobank brain images versus machine-learning datasets, Nat. Commun. 11 (1) (2020) 4238.
[41] P. Thölke, Y.-J. Mantilla-Ramos, H. Abdelhedi, C. Maschke, A. Dehgan, Y. Harel, A. Kemtur, L.M. Berrada, M. Sahraoui, T. Young, et al., Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data, NeuroImage 277 (2023) 120253.
[42] E. Caballero, K. Gupta, I. Rish, D. Krueger, Broken neural scaling laws, in: The Eleventh International Conference on Learning Representations, 2022.
[43] O. Ostapenko, T. Lesort, P. Rodríguez, M.R. Arefin, A. Douillard, I. Rish, L. Charlin, Continual learning with foundation models: An empirical study of latent replay, 2022, arXiv preprint arXiv:2205.00329.
[44] T. Nguyen, M. Raghu, S. Kornblith, Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth, 2020, arXiv preprint arXiv:2010.15327.
[45] G. Vardi, G. Yehudai, O. Shamir, Width is less important than depth in relu neural networks, in: Conference on Learning Theory, PMLR, 2022, pp. 1249–1281.
[46] R. Eldan, O. Shamir, The power of depth for feedforward neural networks, in: Conference on Learning Theory, PMLR, 2016, pp. 907–940.