1 Introduction

Functional near-infrared spectroscopy (fNIRS) is a brain activity measurement that has been employed extensively in the development of brain-computer interfaces (BCIs). Mental workload intensity level is one of the information that can be extracted from each individual's fNIRS signals with the use of machine learning. Traditionally, experiments are set up to collect a sufficient amount of data to build a customized classifier. However, the standard process usually takes over an hour for each participant, which also causes a lot of discomfort from wearing the sensor equipment. Moreover, since each person's fNIRS signals widely vary due to differences in brain structure and physiology of the hair, skin, and skull [1], the models trained on one subject struggles to make accurate predictions on new subjects. To reduce these efforts, many studies have been made on developing generic classifiers, as well as transfer learning models.

Huang and Wang performed a thorough evaluation of four different classification models, Random Forest, Logistic Regression, EEGNet and DeepConvNet, under both subject-specific and generic paradigm [1]. The outcome suggests many potentials in this research area. However, these generic classifiers still suffer from the cross-subject inherent variations as mentioned above. Thus, the study recommends domain adaptation in future study as a way to alleviate the gap in performance caused by these noises.

Similar to this problem, traditional machine learning and deep learning models often only work well in cases where test data comes from the same distribution as the training data. In practice, this assumption does not always hold. Hence, there is a growing interest in developing algorithms that give consistent performance despite this shift in covariate distributions. Li et al. introduced a deep learning method that extracts both invariant representations and risks, LIRR, in semi-supervised domain adaptation settings [2, 3]. The algorithm was tested on image classification tasks and traffic counting regression tasks and evaluated against other deep learning frameworks such as DANN, CDAN, and ADDA. LIRR was showed to outperform other methods on both the classification and regression task when there is a limited number of labeled target data.

Motivated by this finding, in this paper we will examine the effectiveness of LIRR mechanism on the fNIRS dataset in a cross-subject multi-classification task. We are going incorporate the LIRR framework, available through the open source code, into EEGNet. We treat EEGNet without LIRR model as the baseline and will use total loss as the stopping criteria during the training and accuracy scores as the evaluation metrics.

Description of the dataset: Here we describe the data collection procedure done by [1]. For each subject, the data collection produces an fNIRS recording lasting over 20 minutes (task duration). A sliding-window approach is applied to extract fixed-duration (2-40 sec.) windows. Experiments on our dataset suggest that 30 second windows give the highest accuracy. We extract overlapping windows with a stride of 0.6 seconds.

Thus, we have each window as one input unit to the deep learning model. The final dataset consists of 8 signal measurements that are used as the features for the model, and 1488 entries for each subject representing 1488 windows. There are 4 mental workload intensity levels that we want to predict $\{0, 1, 2, 3\}$, with a larger number representing more intensity.

In this study, we examine the transfer from subject coded number 7, to subject coded number 75 and the transfer from subject coded number 75 to subject coded number 36.

2 Contributions

If the LIRR model performs well in this setting, it can be employed for the development of efficient and precise BCIs by automating the process of producing predictions of mental workload

in real time. Additionally, good performance across different subjects means we can save costs and effort of setting up the experiment for new subjects in order to get enough data for training new subject-specific models.

3 Related work

Transfer learning for BCI: There has been a lot of progress on the research into transfer learning models in neuroscience settings using different kinds of brain signals data, contributing to the growing development of BCI. However, models on electroencephalogram (EEG) signals are more extensively studied than fNIRS. Wu et al. [2] compiled methodologies developed since 2016 for EEG-based domain adaptation techniques, ranging from statistical machine learning to deep neural networks. Out of the literature mentioned in this paper, Li et al. [4] most resembles the learning we are going to implement in this study in that both aim to use adversarial training to learn invariant representations while minimizing the classification error. However, the task in question of this study is on emotions classification, while our dependent variable is mental workload levels.

Deep transfer learning on fNIRS data: Within the spectrum of fNIRS-based transfer learning, there has been a handful of studies focusing on using generative adversarial networks (GANs) to enhance their deep learning model on various classification tasks. Nagasawa et al. [5] used a GAN to produce fNIRS time series data that improves the accuracy of motor tasks. Wickramaratne and Mahmud [6] used a GAN as an augmentation tool for their finger- and foot-tapping tasks. In this analysis, we focus on the subclass of mental workload classification tasks.

Deep learning for mental workload classification: Eastmond et al. [7] reviewed all deep learning methods trained on fNIRS data, including a summary table of the architectures used in each study, along with other details of the tasks. Most studies on mental workload classification have only employed traditional black-box deep learning models in their experiments. Ho et al. [8] used CNN and Asgher et al. [9] used LSTM in their work. While greatly increasing the accuracy compared to traditional parametric machine learning methods, these models are still susceptible to shifts in input feature space.

4 Background

Neuroscience: For a comprehensive description of fNIRS data, see Quaresima and Ferrari [10] and Fantini et al. [11]. For the background on fNIRS is related to mental workload, see Huang and Zhang [1] description on the n-back experiment to collect fNIRS signals corresponding to each mental workload level.

Adversarial training: Normal model training involves minimizing a loss function defined for a particular problem. For instance, in classification tasks, our loss function can be cross-entropy loss. Adversarial training refers to the process of incorporating a secondary component (adversarial attacks) to the optimization function on top of the conventional loss function, making the training a two-step program. The model learns the parameters such that it minimizes the loss for our main task while maximizing some predefined secondary function. Zhao and Alwidian [12] gave a holistic discussion of the fundamentals of adversarial training, along with state-of-the-art methods that have been utilized in the industry and in academia.

Domain-Adversarial Neural Network (DANN): A representation learning technique for when training and test time come from similar but different distributions. The method aims to learn a set of features that can discriminate between the objective labels for the main clas-

sification task but cannot discriminate between the training (source) and test (target) domains i.e, given an input observation, the model cannot identify its domain of origin. At the end, we obtain features that have the same or very similar distributions in the source and the target domains, called invariant representation, and they allow us to make good prediction across all domains. To accomplish this task, beside the label predictor of the classification task, we also have a domain classifier that attempts to discriminate between the source and the target domains during training. We optimize the parameters to minimize the loss of the label classifier and to maximize the loss of the domain classifier. The latter update thus works adversarially to the domain classifier, and it encourages domain-invariant features to emerge in the course of the optimization [4].

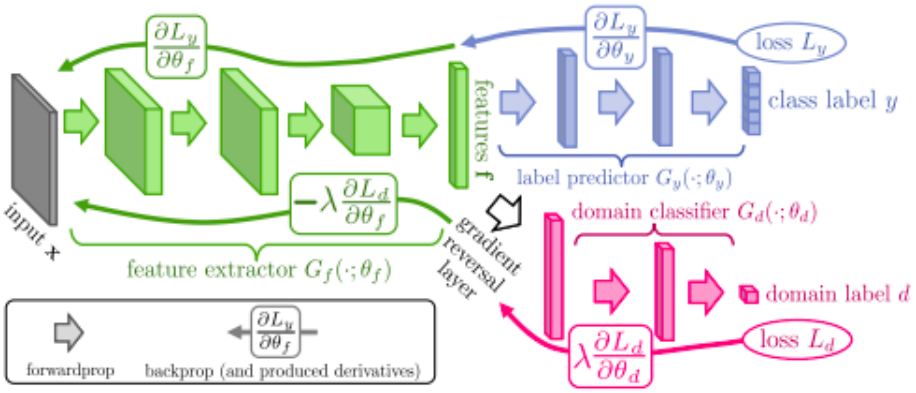

Figure 1: Architecture of DANN – Figure by author[13]

Invariant Risk Minimization (IRM): While machine learning are extracting complex prediction rules from data, they are inherently subject to biases and confounding factors. In trying to minimize training error, the algorithm can learn spurious correlations in the predictors. IRM thus aims to identify which properties of the training data describe spurious correlations and which properties represent the true causal relations of interest (domain-invariant predictors). However, IRM is not sufficient to ensure reduced accuracy discrepancy across domains, and thus we want to conduct representation learning simultaneously [14].

LIRR: combines representation learning and invariant risks learning (Figure 2). As a result, we are able to learn the common features that are causally related to our variable of interest across all domains and make better predictions [13].

The objective function for this learning task is the minimization of the total loss [3]:

$$\min_{g,f_i} \max_{\mathcal{C},f_d} \mathcal{L}_{\text{LIRR}} = \mathcal{L}_{\text{risk}} + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}}$$

The first term is the invariant risk objective which consists of the classification loss from both the domain-invariant and domain-dependent predictors. The second term is the invariant representation objective which we use to fool the domain classifier $\mathcal{C}$ by maximizing the domain classification loss.

EEGNet: Recently, a modified CNN architecture known as EEGNet has shown promise across multiple EEG-based BCIs tasks [1]. We utilized the architecture developed in this study [1] for our experiment.

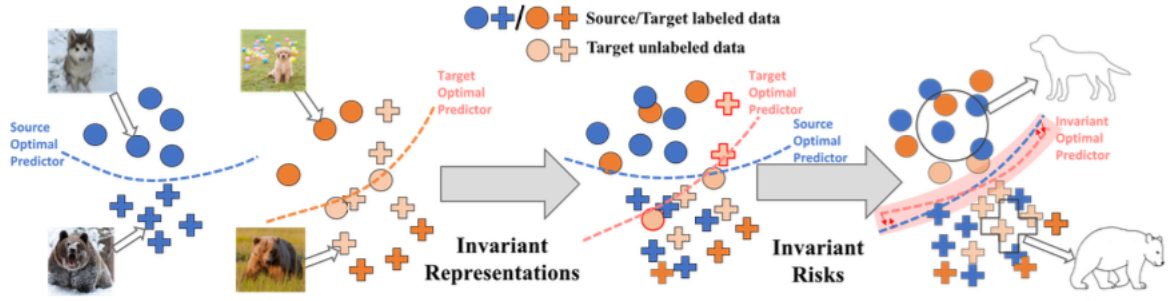

Figure 2: Overview of LIRR method. Learning invariant representations induces indistinguishable representations across domains, but there can still be mis-classified samples (as stated in red circle) due to misaligned optimal predictors. Besides learning invariant representations, LIRR model jointly across domains [3].

5 Method

From here onwards, to stay consistent with transfer learning literature, we refer to the subject we are learning the knowledge from as source subject, and the subject we are transferring the knowledge over to as the target subject.

Training models: In order to prepare the dataset for the model, we follow and combine the procedures of both [3] and [1]. We randomly break the dataset from the source subject and the dataset from the target subject into 5 different splits:

- The source training split: 4/5 of all data of the source subject. Forming part of the input to the model for training (along with the target labeled and unlabeled splits).
- The source validation split: 1/5 of all data of the source subject. Forming part of the validation split for the model (along with the target validation split)
- The target validation split: 1/5 of all data of the target subject. Forming part of the validation split for the model (along with the target validation split)
- The target labeled split: `target_labeled_portion` of the remaining data for the target subject used for model training, or `target_labeled_portion` $\times 4/5$ of all data of the target subject. Forming part of the input to the model.
- The target unlabeled split: $1 -$ `target_labeled_portion` of the remaining data for the target subject used for model training, or $(1 -$ `target_labeled_portion`$) \times 4/5$ of all data of the target subject. Forming part of the input to the model.

The random selection process involves shuffling the 1488 entries of each subject's data and index-subsetting based on the sizes of the training and the validation sets.

Then, the classification step of EEGNet from [1] is modified to match with LIRR framework by changing the design of the final classifier layer. The architecture of EEGNet, which consists of first 3 convolutional layers, namely a first convolutional layer, a depthwise layer, and a separable layer [15], are retained as the feature extractor of the network. The invariant risks learner (also called the environment predictor) and the domain classifier from LIRR will be piped into these layers. The total loss of the network will hence be calculated using the objective laid out in LIRR.

We use a set of fixed parameter to train the model, with learning rate 0.01, batch size of 32, 50 epochs. We keep track of the model that produces the lowest loss on the target test set, and obtain the accuracy score on the target test set using this model.

Evaluating models: Due to the randomness in splitting the data, in addition to the fact that there are overlapping windows, it may be the case that the training and the testing set of the target contain multiple overlaps and thus the true accuracy is overestimated. Therefore, we apply bootstrapping (30 times) to reduce this confounding factor. In each iteration, we reshuffles both the source and the target datasets to obtain different samples for each of the 5 splits.

6 Experiments

We select two case of transfer for the examination: (1) transfer from subject 7 to subject 75 and (2) transfer from subject 75 to subject 36. To compare the results of the two models, we use histograms and boxplots to visualize the distributions of the accuracy scores from all the bootstrap resamples for both cases. Our performance metric is the target validation set accuracy. We report the average and the standard deviation across all bootstrap runs to better understand the effectiveness of a model relative to another. We also provide a table that summarizes these statistics for our experiments (Table 1)

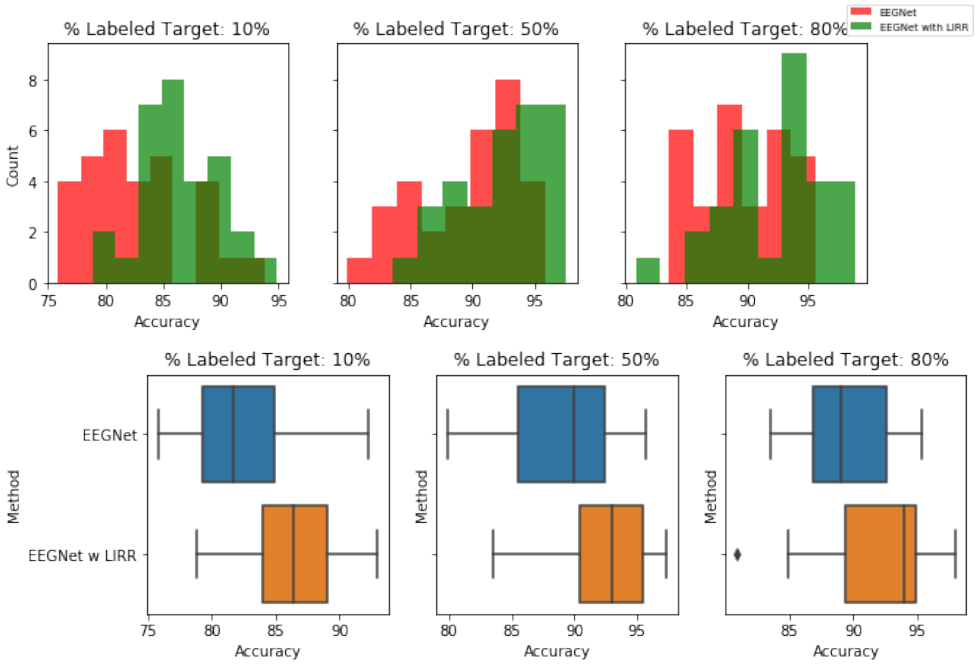

Figure 3: Accuracy (%) distribution of EEGNet model and EEGNet with LIRR model (Transfer from subject 7 to Subject 75) on 10%, 50%, 80% labeled target data after bootstrapping (n = 30)

Result: We make the following observations and takeaways following the outcomes.

- **The model with LIRR is not significantly better**: The comparisons show that the mean accuracy scores are all higher with the model that uses LIRR learning technique, for both cases. However, from the bootstrapped distributions and the variances of the accuracy scores, the accuracy scores for the models with LIRR are not distinctly better than those of the baseline models. This suggests that LIRR can be applied on sequential data to make some improvements cross-domain classification. Running the simulation

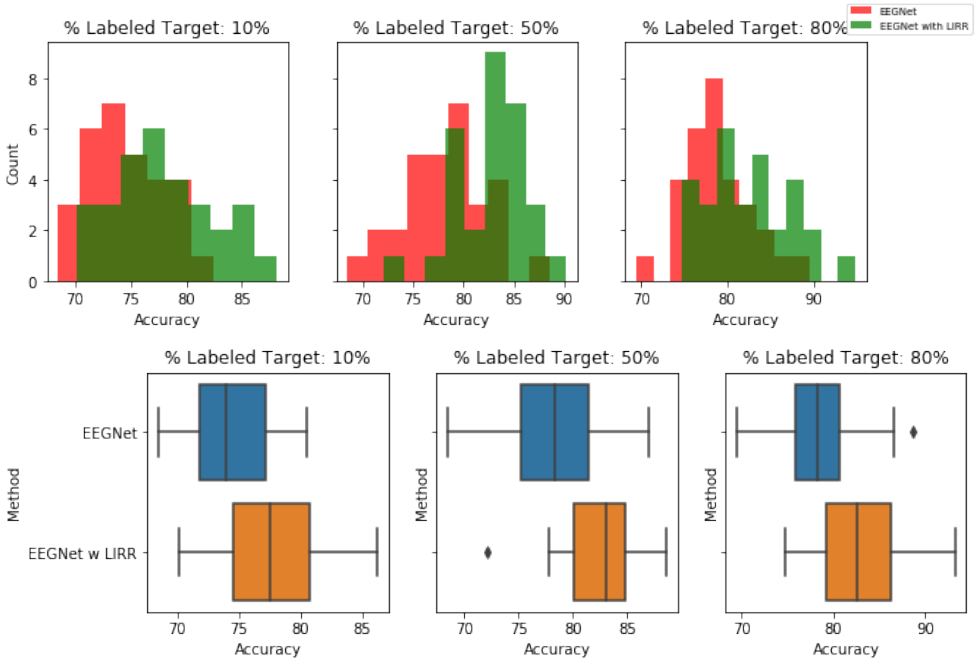

Figure 4: Accuracy (%) distribution of EEGNet model and EEGNet with LIRR model (Transfer from subject 75 to Subject 36) on 10%, 50%, 80% labeled target data after bootstrapping (n = 30)

Method	10% labeled target	50% labeled target	80% labeled target
7 to 75			
EEGNet	82.57 ± 4.44	89.17 ± 4.06	89.50 ± 3.68
EEGNet with LIRR	86.27± 3.37	92.60 ± 3.40	92.27 ± 4.13
75 to 36			
EEGNet	74.35 ± 3.33	78.12 ± 4.20	78.75 ± 4.01
EEGNet with LIRR	77.73 ± 4.38	82.67 ± 3.47	82.57 ± 4.71

Table 1: Accuracy (%) comparison (higher means better) of EEGNet model and EEGNet with LIRR model on 10%, 50%, 80% labeled target data (mean ± std).

for more iterations will can give us a narrower intervals, for which we can determine the quantitative amount of improvement LIRR makes to the classifier.

- **The difference in the average accuracy is greatest when we have less labeled target data**: We see that the more labeled target is present, the better the learning model is, which is aligned with the intuition for semi-supervised domain adaptation. LIRR produces the most improvement from the baseline model when there are less labeled target data. This gap decreases as we have more labels in the target domain, since the baseline method approaches a supervised training in such scenarios. More experiments with even fewer labeled data (5% and 1%) can help us further testify this.

7 Conclusion

In this study, we apply a semi-domain adaptation technique, LIRR, that has been experimented on image classification task, to the fNIRS signal data. We present the preliminary results from bootstrapping EEGNet with and without LIRR on two cases of model transfer, each with different amount of labeled target data. Following these findings, here we also discuss some possible next steps to produce a more rigorous, applicable results:

- Include more sources: [1] and many other studies utilize data from multiple different individuals to increase the training set size, which is especially important in this study setting given that data collection is a significant barrier that has motivated many research in the field. The current LIRR algorithm is currently only tested on transferring from one domain to another. We can expand upon this method to include more subjects in a multi-source transfer learning model.
- Cosine: [2] suggests the use of cosine module to improve the performance of semi-supervised domain adaptation. This is showed in the same study to even further improve the results of both the image classification task and the regression task. We can apply this method on our fNIRS dataset to examine if this holds in the case of sequential data.
- Having recognized the potential of LIRR in improving the mental workload classification task, a next step will be to carry out an extensive tuning of the models to develop a more rigorous model that can be deployed in practice. In addition to the parameters already present in EEGNet and in the LIRR framework, this model tuning also includes considerations for other networks for the individual classifiers in both representation learning and risk invariant learning tasks.

References

[1] Zhe Huang and Liang Wang. *The Tufts fNIRS to Mental Workload Dataset: Toward Brain-Computer Interfaces that Generalize*. 2021. URL: https://openreview.net/forum?id=QzNHE7QHhut.

[2] Xu Y Wu D and Lu B.-L. "Transfer learning for EEG-based braincomputer interfaces: A review of progress made since 2016". In: *IEEE Transactions on Cognitive and Developmental Systems* 14 (2022), pp. 4–19.

[3] Bo Li et al. "Learning Invariant Representations and Risks for Semi-supervised Domain Adaptation". In: *CoRR* abs/2010.04647 (2020). arXiv: 2010.04647. URL: https://arxiv.org/abs/2010.04647.

[4] Li J et al. "Domain adaptation for EEG emotion recognition based on latent representation similarity". In: *IEEE Transactions on Cognitive and Developmental Systems* 12 (2020), pp. 344–353. DOI: 10.1109/tcds.2019.2949306.

[5] Nagasawa T et al. "FNIRS-Gans: Data augmentation using generative adversarial networks for classifying motor tasks from functional near-infrared spectroscopy". In: *Journal of Neural Engineering* 17 (2020), p. 16068. DOI: 10.1088/1741-2552/ab6cb9.

[6] Mahmud M Wickramaratne S. "Conditional-gan based data augmentation for deep learning task classifier improvement using fNIRS Data. Frontiers in Big Data". In: (2021). DOI: 10.3389/fdata.2021.659146.

[7] Subedi A Eastmond C. "Deep learning in fNIRS: A Review". In: *Neurophotonics* (2022). DOI: 10.1117/1.nph.9.4.041411.

[8] Gwak J Ho T. "Discrimination of mental workload levels from multi-channel fNIRS using deep leaning-based approaches". In: *IEEE Access* 7 (2019). DOI: 10.1109/access.2019.2900127.

[9] Khalil K Asgher U. "Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain-computer interface". In: *Frontiers in Neuroscience* 14 (2020). DOI: 10.3389/fnins.2020.00584.

[10] Ferrari Quaresima. "A Mini-Review on functional near-infrared spectroscopy (FNIRS): Where do we stand, and where should we go?" In: *Photonics* 6 (2019). DOI: 10.3390/photonics6030087.

[11] Sassaroli A Fantini S Frederick B. "Perspective: Prospects of non-invasive sensing of the human brain with diffuse optical imaging". In: *APL Photonics* 3 (2018). DOI: 10.1063/1.5038571.

[12] Alwidian S Zhao W and Mahmoud Q. "Adversarial training methods for Deep learning: A systematic review". In: *Algorithms* 15 (2022). DOI: 10.3390/a15080283.

[13] Yaroslav Ganin et al. "Domain-Adversarial Training of Neural Networks". In: *Journal of Machine Learning Research* 17.59 (2016), pp. 1–35. URL: http://jmlr.org/papers/v17/15-239.html.

[14] Martin Arjovsky et al. "Invariant Risk Minimization". In: (2019). DOI: 10.48550/ARXIV.1907.02893. URL: https://arxiv.org/abs/1907.02893.

[15] Vernon J Lawhern et al. "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces". In: *Journal of Neural Engineering* 15.5 (July 2018), p. 056013. DOI: 10.1088/1741-2552/aace8c. URL: https://doi.org/10.1088%2F1741-2552%2Faace8c.