FEATURE IMITATING NETWORKS

Sari Saba-Sadiya[†] *Tuka Alhanai*[‡*] *Mohammad M Ghassemi*[†*]

[†] Department of Computer Science, Michigan State University, East Lansing, MI
[‡] Department of Computer Engineering, New York University, Abu Dhabi, UAE

ABSTRACT

We introduce a novel approach to neural learning: the Feature-Imitating-Network (FIN). A FIN is a neural network with weights that are initialized to reliably approximate one or more closed-form statistical features, such as Shannon's entropy. In this paper, we demonstrate that FINs (and FIN ensembles) provide best-in-class performance for a variety of downstream signal processing and inference tasks, while using less data and requiring less fine-tuning compared to other networks of similar (or even greater) representational power. We conclude that FINs can help bridge the gap between domain experts and machine learning practitioners by enabling researchers to harness insights from feature-engineering to enhance the performance of contemporary representation learning approaches.

1 Introduction

The successful application of deep learning to new problem domains has three conditions: (1) access to large data sets, (2) access to sufficient computing resources for hyper-parameter optimization and (3) modest expectations about model interpretability. Deep learning models require large data-sets to learn representations that generalize on future unseen data. Additionally, extensive exploration of the model topological space is often necessary to identify a network architecture with sufficient representational power for a given task. Lastly, despite ample recent work on interpretability of deep learning models, the community remains without normative standard for how deep networks should be interpreted; this is problematic for many problem domains (e.g. healthcare) where the importance of interpretability may supersede performance [1]. [2].

1.1 Contributions

In this paper we introduce Feature-Imitating-Networks (FINs): a FIN is a neural networks with weights that are initialized to approximate one or more closed-form statistical features. In this paper, we will demonstrate how this property of FINs improves their interpretability while also reducing data and hyper-parameter tuning requirements compared to other networks with similar or greater representational power. More specifically, we demonstrate how, when combined with a careful application of transfer-learning, and by taking into account expert knowledge, FINs can be used to quickly build and deploy robust and better performing models using less training epochs. Our validation of FINs focused on tasks involving biomedical signals; the data-sets in this domain are often smaller, and therefore stand to benefit the most from the introduction of our framework.

[*] These authors contributed equally to this work

1.2 Paper Organization

The remainder of the paper is organized as follows: First we review relevant literature regarding transfer learning. The *Related Work* section is followed by the *Methodology* section where we discuss how to build and design different FINs. The *Experiments* section contains three experiments - including a brief discussion of the data and results for each. Finally, the *Discussion* section examines all the results in aggregate and discusses how our framework might be expanded.

2 Related Work

Transfer learning is the application of a pre-trained model to tasks it was not originally intended to perform [3]. Transfer learning enabled researchers to make significant progress on various tasks in Machine Vision[4], Speech [5], and Natural Language Processing [6].

Most applications of transfer learning are *within-domain*; these involve fine-tuning a pre-trained model for new tasks. For instance, AlexNet [7], VGG [8], and ResNet [9] are computer vision models trained to classify the ImageNet data-set. The features learned by these models (in later layers), and their more fundamental image components (in earlier layers) can be re-purposed to solve other tasks using only a fraction of the training data required by original models.

Models that are trained on large heterogeneous data-sets are good candidates for "transfer". But for biomedical signal processing problems, there isn't a sufficiently large data-set to train such a model. Indeed, the largest publicly available biomedical signal archives contain only a few thousand subjects, which is too small by most data standards in other domains [10]. Consequently, transfer learning for biomedical signals is often performed *across-domains*. For instance, computer vision models such as VGG have been adapted to emotion recognition from speech [5], motor-imagery classification [11] and mental task classification [12], but these *cross-domain* transfers are not as effective as those performed *within-domain*. Finally, interpretability is greatly hindered when models are trained on broad data-sets with objective functions that differ from the final application [2].

The performance of transfer learning is proportional to the proximity of the domains across which knowledge is being transferred. Feature imitating networks were designed to address this limitation of current transfer learning paradigms; they provide the power and flexibility of transfer learning without the "Big data" and heavy computational requirements.

3 Methodology

A FIN is a neural network with weights that are initialized to approximate one or more closed-form statistical features. In this

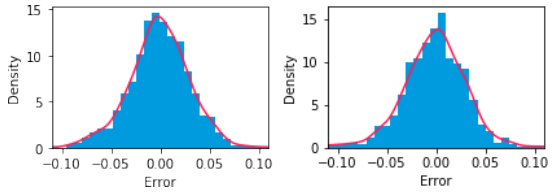

Fig. 1. Density plots for the errors for the entropy (left) and regularity (right) FIN reconstructions. The feature values were scaled and normalized, making the biggest possible error 1, as can be observed the FINs faithfully recreate the closed form equations.

paper, we train FINs that approximate five commonly used features in biomedical signal processing: Shannon's Entropy, kurtosis, skewness, fundamental frequency, Mel-frequency cepstral coefficients (mfcc), and regularity [13]. We evaluate the utility of the FINs on three biomedical signal processing experiments, which we describe in Section 4, below. The pre-trained FINs, and code to reproduce all experimental results are available online [1].

Network Construction For each feature, we used a simple gradient descent optimizer with mean square error (MSE) loss to train a simple dense network to approximate that feature on synthetic signals. The topological space explored for all the FINs was between 2 to 10 layers with the number of parameters in the $3-15$ million range. All best performing FINs used simple $relu$ and $tanh$ activation functions. See Figure 1 for density functions for the errors of the FINs reconstruction.

Input The input data for the FINs consisted of synthetic signals generated randomly (zero mean, unit variance) and converted to the time-frequency domain using the wavelett transform [5, 11].

Outcome The outcome data for the FINs consisted of closed-form feature values calculated on the synthetic signals using *SciPy* and *EEGExtract* packages [13].

Transfer When applying the models to new classification tasks, (i.e. Section 4) the very last layer was discarded in favor of a randomly initialized softmax layer with dimensions suitable to the task.

Baseline model The baseline in each experiment was the best performing neural network with similar (or greater) representational power to the corresponding FIN, trained using the same training data and schedule, but with weights that were randomly initialized; In total one hundred different topologies were explored for each baseline. The baseline with the best average performance on the validation data was retained for comparison against the FINs.

Training All models (both FIN and Baseline) were trained using early stopping and a simple gradient descent optimizer. Non topological hyper-parameters such as learning rate and momentum had minor effects in comparison and therefore will be omitted from future discussions. Training was conducted using *CUDA* on a *Tesla K80* GPU with 25GB of RAM.

4 Experiments

To evaluate our framework we ran three experiments on three different biomedical data-sets and tasks. The first experiment was an Electrocardiography (ECG) classification task; our goal was only to demonstrate that our FINs framework can successfully improve performance on small low-accuracy data-sets. The second experiment was an Electroencephalography (EEG) artifact detection task; our goal was to demonstrate the modular nature of FINs, and the potential benefits of using FIN ensembles. The third experiment was a drowsiness detection task using EEG; our goal was to compare both the performance and speed of FINs against state-of-the-art transfer-learning techniques under conditions of varying data scarcity.

For all three experiments, the data was regularized and transformed to the time-frequency domain as discussed in the Methods section. The data was partitioned into training, validation, and testing sets. In the first two experiments this was achieved by randomly partitioning the data $15\%-85\%$ for testing and training respectively, before repeating the same split for the training data to extract a validation subset. This was repeated for a 50-fold cross-validation. In the third experiment, where subject data is balanced, the partitioning was achieved by iteratively leaving two of the twelve subjects out for validation and testing. To compare training time we used similar instances of nodes with *Tesla K80* GPUs and 25GB of RAM, all training times are reported in seconds.

[1] https://github.com/sari-saba-sadiya/Feature-Imitating-Networks

Model	Baseline	SVM	kNN	Fine-tuned FIN
Mean ($\pm$ std)	.443 ($\pm$0.174)	.5233 ($\pm$0.016)	.525 ($\pm$0.018)	.543 ($\pm$.0245)

Table 1. Mean and standard deviation of the accuracy for the experiment I classification task. As demonstrated the FIN based network out performs both randomly initialized neural networks and classical statistical approaches.

4.1 Experiment I

In this experiment, we explored the potential of FINs for the detection of artifact ridden ECG signals [14, 15].

Data and Prepossessing We used data made available by The *Brno University of Technology ECG Quality Database* [14]. ECG segments of variable lengths from 18 subjects were classified by experts into three categories according to signal quality. After standardizing the lengths we ended up with 2544 trials. Preprocessed data will be made available.

Models We hypothesized that a FIN trained to imitate Kurtosis might be useful in the context of this task [16]. The Kurtosis FIN was adapted for our classification task by replacing the very last layer with a softmax classification layer. In addition to the baseline neural network, we also compare against several non deep learning classification algorithms.

Results As can be seen in Table 1, The Kurtosis FIN consistently outperformed the baseline models. Moreover, the standard deviation in the performance of the FINs was an order of magnitude smaller than the deep network baseline models. A Levene's test indeed indicates a statistically significant ($p < .05$) difference in variance between the performance of the two methods throughout the iterations. This highlights the fact that our framework helps with the robustness of the models.

FIN	Regularity	Fundamental Frequency	Entropy+ Regularity	Kurtosis+ Regularity	Baseline	MFCC	Entropy	Kurtosis	Entropy+Kurtosis +Regularity
Mean (± std)	.6527 (±.1066)	.6825 (±.0591)	.6991 (±.0662)	.7134 (±.0807)	.7142 (±.0587)	.7167 (±.01411)	.7195 (±.03783)	.0.7214 (±.02397)	.0.724 (±.0451)

Table 2. Mean and std of the accuracy for experiment 2. Corrected one-tailed t-tests demonstrated that models imitating features known to be useful for EEG artifact detection (last three columns) significantly out-performed models imitating ill-suited features (first two columns).

4.2 Experiment II

In this experiment, we investigate how different FINs can be used in conjunction to build complex networks suited for EEG artifact detection. Moreover, we demonstrate how theoretical knowledge regarding the features and their relevancy to the task is helpful when using the FINs Framework.

Data and Prepossessing The data used in this experiment is from an EEG artifact detection data-set [13]. The data contains EEG segments from a $1kHz$ recording made using 32 electrodes during a passive viewing task. Each segment is a second long and was labeled as artifact ridden or clean by expert annotators. We re-sampled the data at $500Hz$ and converted the EEG setup to the international $10-20$ system that contains only 19 electrodes.

Models We evaluated individual FINs and FIN ensembles trained to imitate Kurtosis, Shannon's Entropy, Regularity, Fundamental Frequency, MFCCs, and ensemble combinations thereof. We expect some of these FINs to outperform others based on the task-relevance of the feature being imitated. For instance, the fundamental frequency of the signal, defined as the lowest periodic frequency of the waveform should be irrelevant, while the Kurtosis is highly relevant to the task [17, 18]. Similarly, we expect 'Complexity Features' such as the cepstrum coefficients and Shannon's entropy to outperform clinically grounded 'Continuity Features' such as the fundamental frequency or EEG regularity (burst suppression) [13]. Following these theoretical considerations we hypothesize that the MFCC, Entropy, and Kurtosis FINS, as well as a combination of these FINS will outperform the Regularity FINs. As we have multiple hypotheses, appropriate Bonferroni correction for multiple comparisons was used. To have enough statistical power after the correction we increased sample size by repeating each experiment 50 times.

Each FIN was applied on each electrode signal in parallel, the outputs were then concatenated and passed forward to a binary softmax classification layer. We compared the FINs against the best performing baseline dense neural network.

Results As summarized in Table 2, our experiment demonstrates how (when appropriately selected) FIN ensembles may be used in combination to further enhance task performance. We note here that deliberate consideration when combining FINs can improve task performance, while keeping the size of the ensemble small.

A corrected one tailed t-test showed that after correction the Kurtosis, Entropy, and Ensamble Network (last three columns in the table) performed significantly better than the Fundamental Frequency or Regularity FINs.

4.3 Experiment III

In this experiment, we compare the performance of FINs against state-of-the-art approaches for a fatigue and drowsiness detection from EEG task on a recently published data-set.

Data and Prepossessing We used data made available by [19]. This paper identified multiple subsets of electrodes as especially predictive. We tested separately on every subset. The data was then partitioned for six fold intra-subject cross validation. In other words, at each iteration ten out of twelve available subjects were used for training, one was used for validation, and the testing accuracy was calculated on the remaining subject. Finally, each cross-validation step was repeated 5 times. If only a fraction of the training data was being used different trials were picked at each of these iterations.

Models Prior work indicates that entropy is a useful feature for the prediction of drowsiness and fatigue from EEG data [19]. Thus, we compared a FIN trained to imitate Shannon's entropy against the baseline models (described in the methods), as well as a fine tuned VGG network pre-tarined on the ImageNet data-set [8]. The VGG model was similar to the 19 layer convolutional neural network introduced in [8] sans the last 3 dense layers and with the addition of a final softmax classification layer. This is the standard way in which the VGG model is used in biomedical tasks [12, 11].

Additionally, to test the performance of our pre-trained FIN when only very limited data is available we ran the same models but with varying fractions of the data being made available during training.

Results The pre-tuned Shannon entropy FIN outperformed the baseline and reinitialized FIN in each of the four subsets and at over 83% of the iterations of the cross validation. Additionally, as can be seen in Figure 2 (top), the pre-tuned FIN had lower loss at every epoch compared to the baseline. It is important to note that the FIN also beat the performance of classical classifiers with different entropy measures that was reported in the literature [19].

The performance improvements were even greater under limited data availability conditions As small training data sets are known to increase performance noise, we also repeated this process 10 times and reported the average accuracy; the difference in the average accuracy between our method and the baseline for each data percentage is plotted in figure 2 (bottom).

The VGG transfer learning network under-performed both the FINs and other baseline models and was particularly sensitive to small data sizes. When only a fraction of the data was available, VGG performed at close to chance.

5 Discussion

The feature imitating networks framework proposed in this paper is an innovative way to use transfer learning. Traditional transfer learning requires large, slow to train, black-box, networks such as *VGG* and *AlexNet* tuned on hundreds of thousands of labeled data. In contrast, FINs require no human labeling, are small and fast to train, and can be combined to create ensemble FIN networks in

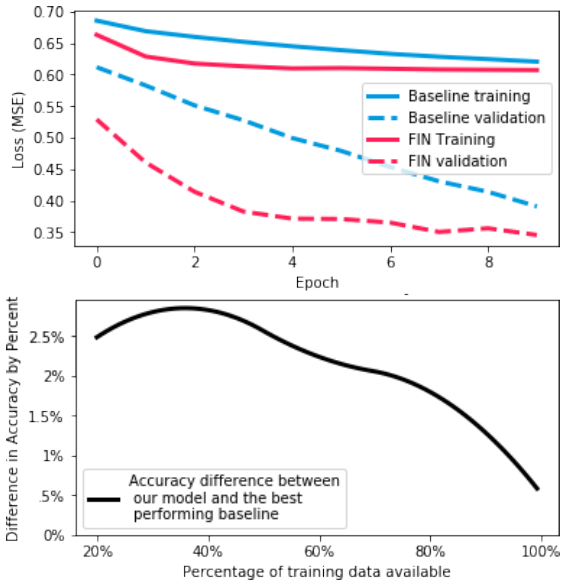

Fig. 2. Experiment 3 results. (top): Training and Validation loss for the baseline and pre-tuned entropy FIN. (Bottom): Difference in accuracy between our pre-trained FIN and the best performing baseline as a function of the percentage of data available.

Mean (± std) Training Seconds	Baseline	VGG	FIN
20% of data	.746(±0.019) 37.2(±8.9)	0.615(±0.122) 21911(±520)	**.771(±0.016)** 50.8(±5.7)
40% of data	.895(±0.04) 37.2(±0.2)	.644(±0.18) 10881(±1441)	**.922(±0.013)** 106.0(±23.3)
60% of data	.939(±.07) 189.8(±79.9)	.69(±.197) 18434(±1058)	**.962(±.006)** 322.4(±201.2)
80% of data	.983(±0.0203) 574.8(±132.3)	.802(±.103) 17141(±1838)	**.996(±.0013)** 441.6(±92.2)
100% of data	.993(±0.022) 645.7(±187.1)	.846(±.088) 19267.3(±2014)	**.998(±0.001)** 552.6(±129.7)

Table 3. Experiment 3 results. The models were trained using varying subsets of the data. We report the mean and standard deviation for both accuracy and training duration on a node with a *Tesla K80* GPU and 25GB of RAM running *CUDA*

accordance with insights from the literature surrounding the task being performed. Therefore, our network facilitates the integration of domain specific knowledge into modern data driven machine learning practices. Beyond these considerations there are several practical benefits to using our framework:

- **Robustness:** Our experiments indicated that FINs are more robust than other networks and techniques with similar representational power. This is evident in statistically significant differences in variations in accuracy when performing leave subject out and cross validation. Deep learning in general is sensitive to weight initialization randomness and data idiosyncrasies. Transfer learning of weights tuned to calculate task relevant features seems to guarantee we start at a 'neighborhood' of a good solution. Moreover, FINs expedite the hyper-parameter optimization step which remains resource-consuming despite recent research [20, 21].
- **Performance:** Data scarcity still plagues many domains. In the case of biomedical research data collection is especially costly and can prohibit researchers from applying deep learning to their tasks altogether [22]. Our experiments indicate that FINs are useful especially when only limited data is available. The intuition behind this is straightforward; pre-trained weights already extract useful task-relevant information, resulting in a better performance and lower loss when from the very first epochs of the training procedure, as can be seen in Figure 2.
- **Flexibility:** By tuning on task-specific data sets our framework also out performs methods that pass the calculated features as input to the classifiers. This is not surprising as our FINs are allowed to tune the extracted features to better suit the task (for instance by focusing on specific parts of the signal). Additionally, the modular nature of the FINs lends itself to easily building and testing ensembles networks.
- **Speed:** *VGG* and *AlexNet* are powerful networks that have been successfully applied in various domains. However, these architectures are extremely large. The *VGG* based network consists of at least 17 layers and contains over 20 million parameters. In contrast, FINs are simple shallow networks consisting of up to 4 layers and a quarter of that numbers of parameters. The shallowness of the models in particular guarantees that even when using an ensemble of FINs gradient descent calculations, and therefore training and inference times, remain simple and fast. This can be observed in the results of the third experiment presented in Table 3, training the VGG network was in some cases over 60 times slower than the FINs network training despite lower performance.

5.1 Future Directions

Designing the dense implementation of the FINs can be streamlined by considering the closed form equation of the signal and training layers to imitate each operator separately. For instance, the mathematical expression for Shannon's entropy requires discretization of the signal to create a histogram before averaging each bin. Partitioning the operations allows us to reuse pre-trained operation-specific layers to quickly construct FINs that are then fine tuned to mimic specific features.

6 Conclusion

In recent years, some have critiqued the current state of the machine learning community. These critiques often focus on disregard of traditional techniques in favor of data driven approaches [23] and the different ways deep learning have struggled to live up to it's promise, especially when it comes to real world applications [1]. In this paper, we presented *Feature-Imitating-Networks*, a variation over traditional transfer learning that uses networks trained to imitate simple closed form statistical features, that we believe elevates these concerns. We demonstrated that our framework is superior in both the speed and accuracy to deep and transfer learning techniques with similar (or greater) representational ability. Especially when only very limited data is available. The experiments were conducted on a variety of tasks and domains. Future work will extend this initial exploratory work. An extensive library of Feature Imitating Network benchmarked on many data-sets and other signal processing domains might be of particular use and interest to the research community.

7 References

[1] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, James Rudd, Evis Sala, and Carola-Bibiane Schönlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," *Nature Machine Intelligence*, vol. 3, 03 2021.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[3] Andrew Ng, "Nips 2016 tutorial: Nuts and bolts of building ai applications using deep learning," 2016.

[4] Marcia Hon and Naimul Mefraz Khan, "Towards alzheimer's disease classification through transfer learning," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1166–1169.

[5] Margaret Lech, Melissa Stolar, Christopher Best, and Robert Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, vol. 2, pp. 14, 2020.

[6] Jeremy Howard and Sebastian Ruder, "Universal language model fine-tuning for text classification," 01 2018, pp. 328–339.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[10] Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford, "You snooze, you win: the physionet/computing in cardiology challenge 2018," in *2018 Computing in Cardiology Conference (CinC)*. IEEE, 2018, vol. 45, pp. 1–4.

[11] G. Xu, X. Shen, S. Chen, Y. Zong, C. Zhang, H. Yue, M. Liu, F. Chen, and W. Che, "A deep transfer convolutional neural network framework for eeg signal classification," *IEEE Access*, vol. 7, pp. 112767–112776, 2019.

[12] Sławomir Opałka, Bartłomiej Stasiak, Dominik Szajerman, and Adam Wojciechowski, "Multi-channel convolutional neural networks architecture feeding for effective eeg mental tasks classification," *Sensors*, vol. 18, no. 10, 2018.

[13] Sari Saba-Sadiya, Eric Chantland, Tuka Alhanai, Taosheng Liu, and Mohammad Mahdi Ghassemi, "Unsupervised eeg artifact detection and correction," *Frontiers in Digital Health*, vol. 2, pp. 57, 2020.

[14] Andrea Nemcova et al., "Brno university of technology ECG quality database (BUT QDB)," PhysioNet., 2021.

[15] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, and H. E Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals.," Circulation [Online], 2000.

[16] Zhidong Zhao and Yefei Zhang, "Sqi quality evaluation mechanism of single-lead ecg signal based on simple heuristic fusion and fuzzy comprehensive evaluation," *Frontiers in Physiology*, vol. 9, pp. 727, 2018.

[17] Arnaud Delorme, Scott Makeig, and Terrence Sejnowski, "Automatic artifact rejection for eeg data using high-order statistics and independant component analysis," in *Proceedings of the 3rd International Independant Component Analysis and Blind Source Decomposition Conference*, 01 2001.

[18] Soroush Javidi, Danilo Mandic, Clive Cheong Took, and Andrzej Cichocki, "Kurtosis based blind source extraction of complex noncircular signals with application in eeg artifact removal in real-time," *Frontiers in Neuroscience*, vol. 5, pp. 105, 2011.

[19] Jianliang Min, Ping Wang, and Jianfeng Hu, "Driver fatigue detection through multiple entropy fusion analysis in an eeg-based system," *PLOS ONE*, vol. 12, pp. e0188756, 12 2017.

[20] Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng, "Hyperparameter optimization for machine learning models based on bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[21] Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Daniel Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duffy, and Babak Hodjat, "Chapter 15 - evolving deep neural networks," in *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, Robert Kozma, Cesare Alippi, Yoonsuck Choe, and Francesco Carlo Morabito, Eds., pp. 293–312. Academic Press, 2019.

[22] Sari Sadiya, Tuka Alhanai, and Mohammad Ghassemi, "Artifact detection and correction in eeg data: A review," in *Proceedings of the 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 05 2021, pp. 495–498.

[23] Christopher D. Manning, "Computational linguistics and deep learning," *Computational Linguistics*, vol. 41, no. 4, pp. 701–707, 2015.