Optimized preprocessing and Tiny ML for Attention State Classification

Yinghao Wang, Rémi Nahon, Enzo Tartaglione, Pavlo Mozharovskyi, and Van-Tam Nguyen

LTCI, Télécom Paris, Institut Polytechnique de Paris

yinghao.wang@telecom-paris.fr

Abstract—**Electroencephalography has been widely used to study mental processes such as attention, perception, and emotion. This is because mental state classification has important applications in many fields, including healthcare, human-computer interaction, and education.**

In this paper, we present a new approach to mental state classification from EEG signals by combining signal processing techniques and machine learning (ML) algorithms. We evaluate the performance of the proposed method on a dataset of EEG recordings collected during a cognitive load task. The results show that the proposed method achieves high accuracy in classifying mental states and outperforms state-of-the-art methods in terms of classification accuracy and computational efficiency.

Index Terms—**Tiny ML, Attention level detection, EEG, Brain-computer interface**

I. INTRODUCTION

Attention is a cognitive process that involves the selection and prioritization of sensory information to be processed [1]. The ability to accurately classify attention states using electroencephalography (EEG) could have important implications for a wide range of applications, from monitoring cognitive load in educational settings [2] to developing neurofeedback systems to improve cognitive performance [3] [4].

Due to the non-stationary nature of EEG signals, as well as their time-varying spectral characteristics [5], it is necessary to analyze EEG signals in the joint time-frequency domain rather than the time-only or the frequency-only domain only [6]. In recent research, machine learning-based classifiers have shown high potential for distinguishing mental states using spectrogram-based features. In [7], the Support Vector Machine (SVM) was used to classify three mental states (focused, unfocused, and drowsy) with 91.72% accuracy when training and testing were performed on a single subject with seven channels. In [8], an accuracy of 96.11% was achieved using an SVM for drowsiness detection on three-channel data from a single subject. Previous work also shown that, in some cases, deep learning-based algorithm is more efficient than conventional machine learning in this type of task [9] [10] [11], which motivates us to optimize the preprocessing process to make it more suitable for Tiny ML, in term of less computation, fewer engineers, and less data.

In this work, we explore the effect of different compositions of labeled EEG data on the classification results. In particular, for feature extraction, we conduct a quantitative study on the parameters of the spectrogram-based features extraction, and for the classification algorithm, we implement Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), and Random Forest (RF) along with two neural network models.

II. MATERIAL AND METHOD

Fig. 1 provides an overview of the approach undertaken in this work to go from raw EEG signal to the final mental state classification and the main hyperparameters involved. We use the dataset presented in Sec. II-A proposed by [7]. First, we proceed to a pre-selection of the data (Sec. II-B). Then we perform our feature extraction (Sec. II-C) that will improve the performance of the different classifiers we use for the mental classification task (Sec. II-D).

A. Raw dataset used

In the present study, we utilized a publicly accessible dataset, originally collected by [7] from five participants. The dataset comprises 25 hours of EEG recordings, with each trial lasting between 35 and 55 minutes. This dataset is the most scientific for long continuous acquisition despite difficulty in maintaining passive focus for ten minutes. It is also closely aligned with practical applications. The three mental states under investigation are as follows:

1) *Focused* but passive attention,
2) *Unfocused* or detached but awake,
3) *Drowsy* or on the verge of falling asleep.

The task consists of controlling a computer-simulated train using the program "Microsoft Train Simulator". During the first 10 minutes of each trial, participants are engaged in focused control of the simulated train, paying close attention to the simulator controls. For the next 10 minutes, the train moves on a flat road without any control from the participant, who entered a de-focused state. For the remainder of the experiment, participants are allowed to relax freely, close their eyes, and doze off, as they wish. This dataset provides EEG signals from 7 channels at a sampling rate of 128Hz, which are identified as F3, F4, Fz, C3, C4, Cz, and Pz in the standard 10-20 electrode system [12]. More detailed descriptions of the experimental paradigm were presented in [7].

B. Dataset formation

Trial selection. Various factors such as motion artifacts or unfamiliarity with the experimental procedures can lead to mislabeling of the collected data. To minimize these risks,
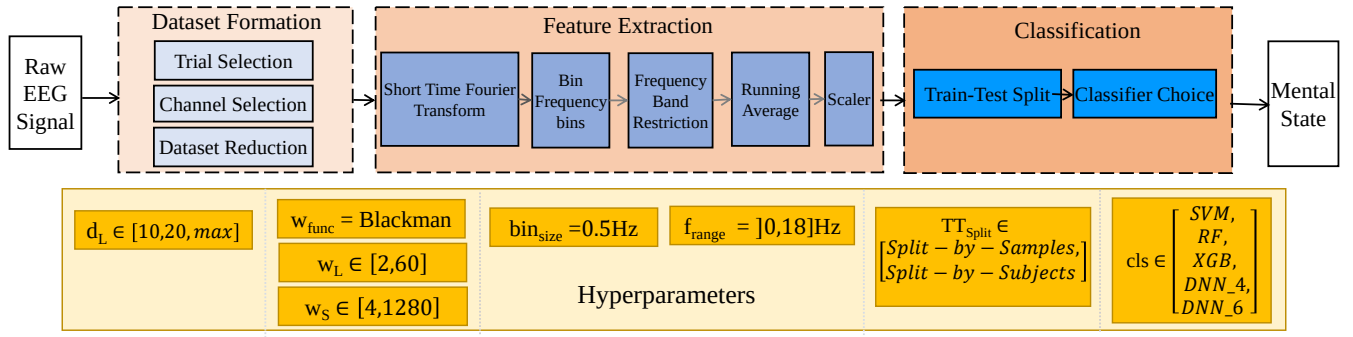
Fig. 1: Representation of the whole process undertaken from the raw signal to the mental state detection.

the choice made in [7] is strictly adhered to and the first two trials of each participant from the final dataset are deleted.

Dataset length reduction. Algorithmic fairness plays an important role in ML and imposing fairness constraints during learning is a common approach. Imbalanced datasets lead to a lack of generalizability not only of classification but also of fairness properties, especially in over-parameterized models [13]. To address this problem, we propose three different paradigms regarding the value of d_L, the length (in minutes) of the drowsy state portion of the trial that we retain:

- $d_L = 10$: the dataset is balanced, with ten minutes in each state for each trial, but the number of samples used for training is significantly reduced,
- $d_L = 20$: the dataset is balanced between drowsiness and non-drowsiness data and the duration of each trial is reduced to 40 minutes,
- $d_L = max$: the dataset is the original unbalanced one.

Channel selection. EEG headsets typically have dozens or even hundreds of electrodes that capture electrical signals from the brain. While using fewer electrodes can make the headset less cumbersome and more convenient in practice, it can also limit the amount and quality of data collected, which can affect the accuracy of the analysis. We performed most of our analysis with 7 electrodes, as in [7]. From the perspective of a wearable device, we examined how decreasing the number of EEG channels affects classification performance.

C. Feature extraction

STFT window length. We chose to build our feature vectors from spectrograms obtained by Short Term Fourier Transform (STFT). The choice of the STFT window length w_L is a trade-off between frequency resolution and temporal resolution. We explore $w_L \in [2, 60]$ (in seconds) because we want to achieve a frequency resolution of 0.5Hz (equivalent to $w_L = 2$) and we intend to use our method in a real-time perspective which makes windows bigger than 60 seconds intractable.

STFT window shift. Inspired by [14], we assume that the overlap between windows also has an impact on the efficiency of the feature extraction process. In particular, a small window shift w_S provides more samples, resulting in more training data, thus improving ML performance. However, it is also known that too much dependence between the input samples

leads to overfitting and loss of generalization of the models. We, therefore, perform a joint optimization of w_S and w_L, because the overlap can be calculated from them. We consider $w_S \in [4, 1280]$, ie. from the minimum accessible in terms of memory to ten seconds of shift between data points, which reduces too much the number of points for classification.

STFT window function. EEG signals often contain large spurious signals near the frequency of interest, so the *Blackman* function is used as in [7], because the resulting spectrum has a wide peak, but good side-lobe compression [15].

Frequency binning. The setting in [7] is used to bin the frequency bands using $bin_{size} = 0.5$ (in Hz), which can lead to higher accuracy while also reducing the ML training time.

Band restriction. We restricted the frequencies to a range $f_{range} = (0, 18]$ (in Hz) as it is known that the significant changes in EEG signals occur within that range [16].

Running average. The spectrograms are smoothed over time using a 15-second moving average to reduce noise. The final feature vector is then formed by converting the power values at each time point into decibels.

Scaling. To balance the influence of each feature on classification, regardless of the range of values it takes [17], we apply $scaler = standardization$ to the feature vector.

Train-test split. We use two distinct methods to divide the dataset into training and testing sets to fulfill different goals: The *Split-by-Samples* paradigm reflects the ability of the model to apply what it learns to the individual it has been training on. We divide it into two sub-paradigms regarding the number of subjects involved. With *Subject-Specific*, as we train and test on a single subject, we can efficiently identify changes in an individual's brain activity in response to various stimuli or tasks. With *Common-Subject*, we train and test on all subjects at once, which is widely used in the literature as it makes it possible to train with more variability in the data.

One of the major challenges we face in EEG classification lies in the generalization of what is learned from one set of individuals to another one, as the features picked up by the models tend to vary considerably from subject to subject. To test our method on this point, we propose the *Split-by-subject* paradigm that involves training and testing on different individuals. Its application in this paper referred to as *Leave-*

One-Out, consists of training the classifier on four subjects and testing on the fifth.

D. Classifiers used

In this work, we use four different families of algorithms to perform our EEG classification tasks:

RF is efficient to process noisy and unbalanced data such as those encountered in EEG signals [18].

SVM is used for its ability to handle high-dimensional data and its efficacy on both linear and non-linear problems [19].

XGB is known for its speed, accuracy, and ability to handle large and high-dimensional datasets [20].

Neural Networks (NN) are effective and popular for EEG classification tasks, especially when combined with appropriate feature extraction and preprocessing methods [9].*DNN_4* is a simple multilayer perceptron of four layers (with two hidden 64-neuron layers). We also use a deeper network, *DNN_6*, that consists of *DNN_4* with two supplementary internal layers (of 128 neurons) and two dropouts layers before the last hidden layer to address the increased risk of overfitting known to be inherent to larger neural networks [21].

III. EXPERIMENTS AND RESULTS

A. Data length study

In this section, we use the seven channels of EEG signal provided by the dataset, while keeping the hyperparameters introduced in Sec. II-B as presented in [7], without any modification. The *Leave-One-Out* paradigm is used here as it is the most demanding. Fig. 2 shows the results of the five classifiers presented in Sec. II-D. In addition, we chose to use a balanced accuracy instead of the traditional one that would be biased towards the more represented "drowsed" class.

Accuracy. As shown in the top part of Fig. 2, the classification accuracy is higher for $d_L = 20$ than for $d_L = max$, across the five subjects. However when comparing the paradigms $d_L = 20$ and $d_L = 10$, the performances are very close.

Recall. Regarding the recall (in the bottom part of Fig. 2), the best performances are obtained for $d_L = max$, and the results for $d_L = 20$ are very close to it. On the other hand, the results for $d_L = 10$ are much worse.

These results show that correctly balancing data is useful for ML training. However, finding the full balance of the dataset may significantly reduce the training samples, which will then hurt the training. As it seems to always yield good results, we use the $d_L = 20$ paradigm in all subsequent experiments.

B. STFT quantitative study

In order to study the effect of different STFT parameters on the performance of the classifier, the *Leave-One-Out* paradigm is chosen, due to the generalization perspective, but also due to the higher validity of the extracted features. Since large NN models contain huge amounts of parameters, the DNN_6 classifier is very sensitive to feature quality.

As shown in Fig. 3 the feature quality deteriorates with increasing window length and shift scale. In particular, we obtain the best features in the red triangular shape region.
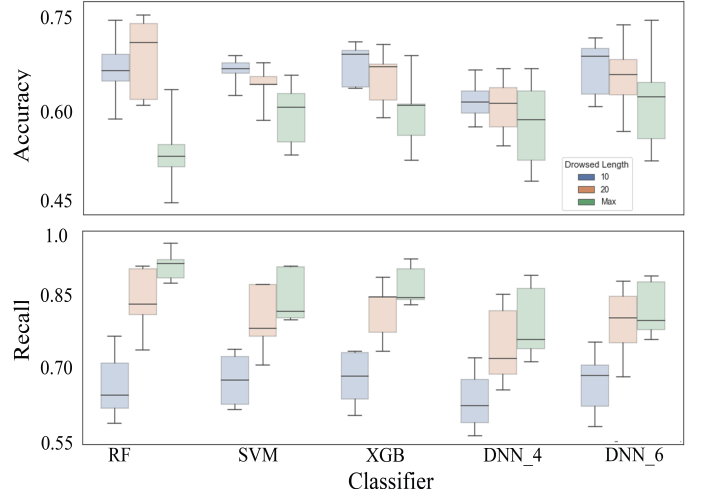

Fig. 2: Accuracy (top part) and recall on the "drowsy" state (bottom part) are evaluated for 5 classifiers for different d_L values.

TABLE I: Results on the *Subject-Specific* paradigm († indicates the results taken from [7]).

Classifier	Best accuracy	Average accuracy	Feature engineering	Fine tuning
RF	99.6%	99.3%	✓	✓
XGB	<u>99.8%</u>	<u>99.6%</u>	✓	✓
SVM	**99.9%**	**99.8%**	✓	✓
DNN_4	<u>99.8%</u>	99.6%	✓	✓
DNN_6	**99.9%**	<u>99.6%</u>	✓	✓
SVM	96.7%†	91.7%†	✓	✗

In order to study this area more thoroughly, we use sample-sliding instead of a window length ratio to avoid a decrease in accuracy caused by a lack of training data. In this way, we decouple the relationship between window length and the number of data points. As shown in Fig. 3, the best result is obtained when $w_S = 128$. The accuracy decreases as w_L increases, and the best result is obtained for $w_L = 4$.

Generally, different combinations of w_L and w_S affect the extracted features in two ways: the independence between the samples used to train the ML algorithm, and the number of samples. In the ideal case, we want each feature to provide unique information to the algorithm, and not have redundant or overlapping information. However, this requires a large amount of labeled data, and the cost of the data acquisition is a pressing problem for Tiny ML. In this situation, data enhancement is achieved by adding some overlap between samples, which improves accuracy by 6%.

C. Results of classification

In this section, we evaluate the classification performance on the three train-test split paradigms. In each case, we use the optimal hyperparameters found in Sec. III.

Feature engineering effects. To test the effects of our approach, TABLE I, performed in the *Subject-Specific* paradigm, shows that the detection of attentional states can be achieved
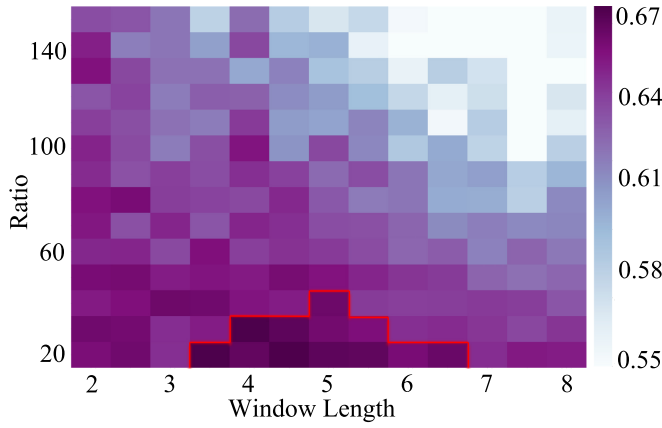
Fig. 3: Accuracy on the *Leave-One-Out* paradigm for different feature extraction configuration: Variation of the window length (in seconds) and of the ratio $\frac{w_S}{w_L}$. The red lining frames the best combinations.
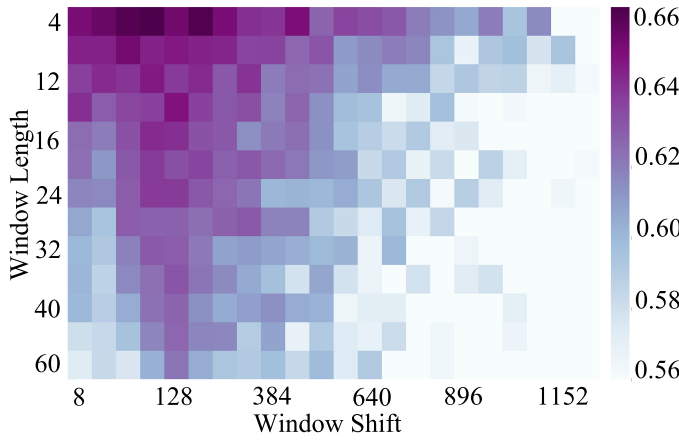

Fig. 4: Accuracy on the *Leave-One-Out* paradigm for different feature extraction configuration: Variation of the window length (in seconds) and of window shift (in samples).

TABLE II: Accuracy on the *Common-Subject* paradigm for different numbers of channels.

Classifier	Number of channels used			
	7	3	2	1
RF	97.3%	96.6%	**95.8%**	**92.6%**
SVM	**99.2%**	**97.6%**	<u>94.0%</u>	82.8%
XGB	98.0%	95.0%	91.2%	81.9%
DNN_4	<u>98.3%</u>	96.7%	93.9%	<u>87.8%</u>
DNN_6	98.1%	<u>97.3%</u>	93.6%	87.4%

rithms, overfitting is also mitigated by adding a dropout layer, resulting in better accuracy. The SVM algorithm performs best when the amount of data is large. Overall, even though we use only two electrodes to collect EEG data, we achieve more than 95% accuracy, demonstrating the validity of our extracted features and is also promising for real-world applications.

Models generalization. Generalizing feature extraction for EEG signals has been an important problem for researchers due to the variation of EEG signals among individual subjects. Effective features and appropriate algorithms can help solve this problem, even in this very challenging paradigm. To test the performance of the models on that aspect, we used the *Leave-One-Out* paradigm. As shown in Fig. 5, we achieve the best accuracy of 71.1% using the DNN_6 classifier under this paradigm. In addition, the deep learning algorithm also shows good potential for more balanced performance across subjects.
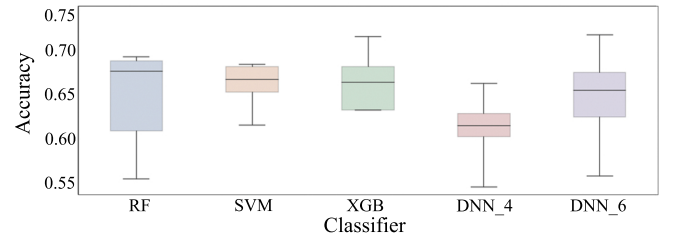

Fig. 5: Accuracy on the *Leave-One-Out* paradigm for five classifiers.

effectively using handmade features. Compared to the results obtained by [7], with the same classifier (SVM), our feature extraction with fine-tuned parameters significantly improves the yielded accuracy (from 96.7% to 99.9% for the best accuracy and from 91.7% to 99.8% for the average accuracy). This shows the efficacy of our feature extraction method.

Channel selection. We investigate the effect of the number of EEG electrodes used to acquire the signal on the prediction results. We perform this study in *Common-Subject* paradigm as it helps train on the most data. From the ablation study of the channels, we know that the three positions that have the greatest influence on classification accuracy are Fz, F3, and Pz. TABLE. II shows the accuracy obtained for the best channel combination for multiple numbers of channels. The RF algorithm can effectively avoid overfitting and obtain the best accuracy when the amount of data is small. This is especially evident when there is only one channel. In NN algo-

IV. CONCLUSION

In this paper, we presented in detail the feature engineering of EEG signals, verified the effectiveness and practicality of the extracted features through comprehensive experiments, and compared established classification algorithms. The results show that our careful choice of preprocessing parameters and their optimization significantly improves the final result compared to the state-of-the-art (over 3% for best accuracy and over 8% for average accuracy). This allows very high accuracy in *Subject-specific* and *Common-Subject* paradigms. At the same time, in the *Leave-One-Out* paradigm, the proposed algorithms are still far from the two other paradigms. This can be explained to a large extent by an insufficient number of individuals. However, the result from NN models are promising and motivates us to explore more efficient feature extraction methods and to design better tiny ML algorithms.

REFERENCES

[1] Michael I Posner. Attention: the mechanisms of consciousness. *Proceedings of the National Academy of Sciences*, 91(16):7398–7403, 1994.

[2] Ning-Han Liu, Cheng-Yu Chiang, and Hsuan-Chin Chu. Recognizing the degree of human attention using eeg signals from mobile sensors. *sensors*, 13(8):10273–10286, 2013.

[3] Choon Guan Lim, Tih Shih Lee, Cuntai Guan, Daniel Shuen Sheng Fung, Yudong Zhao, Stephanie Sze Wei Teng, Haihong Zhang, and K Ranga Rama Krishnan. A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder. *PloS one*, 7(10):e46692, 2012.

[4] Stefanie Enriquez-Geppert, René J Huster, and Christoph S Herrmann. Eeg-neurofeedback as a tool to modulate cognition and behavior: a review tutorial. *Frontiers in human neuroscience*, 11:51, 2017.

[5] Boualem Boashash. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic press, 2015.

[6] Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

[7] Çiğdem İnan Acı, Murat Kaya, and Yuriy Mishchenko. Distinguishing mental attention states of humans via an eeg-based passive bci using machine learning methods. *Expert Systems with Applications*, 134:153–166, 2019.

[8] Mihong Yang, Huiyan Li, Xiaozhou Sun, Li Yang, Hailong Duan, Yanqiu Che, and Chunxiao Han. Toward wearable eeg-based alertness detection system using svm with optimal minimum channels. In *MATEC Web of Conferences*, volume 214, page 03009. EDP Sciences, 2018.

[9] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning robust features using deep learning for automatic seizure detection. In *Machine learning for healthcare conference*, pages 178–190. PMLR, 2016.

[10] Zhongke Gao, Weidong Dang, Xinmin Wang, Xiaolin Hong, Linhua Hou, Kai Ma, and Matjaž Perc. Complex networks and deep learning for eeg signal analysis. *Cognitive Neurodynamics*, 15:369–388, 2021.

[11] Miankuan Zhu, Jiangfan Chen, Haobo Li, Fujian Liang, Lei Han, and Zutao Zhang. Vehicle driver drowsiness detection method using wearable eeg based on convolution neural network. *Neural computing and applications*, 33(20):13965–13980, 2021.

[12] George H Klem. The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology. *Electroencephalogr. Clin. Neurophysiol. Suppl.*, 52:3–6, 1999.

[13] Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. Fifa: Making fairness more generalizable in classifiers trained on imbalanced data. *arXiv preprint arXiv:2206.02792*, 2022.

[14] Zhendong Mu, Jianfeng Hu, and Jianliang Min. Driver fatigue detection system using electroencephalography signals based on combined entropy features. *Applied Sciences*, 7(2):150, 2017.

[15] Raïs El'hadi Bekka and Djamel Chikouche. Effect of the window length on the emg spectral estimation through the blackman-tukey method. In *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, volume 2, pages 17–20. IEEE, 2003.

[16] Rami Alazrai, Motaz Abuhijleh, Mostafa Z Ali, and Mohammad I Daoud. A deep learning approach for decoding visually imagined digits and letters using time–frequency–spatial representation of eeg signals. *Expert Systems with Applications*, 203:117417, 2022.

[17] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, 16, 2003.

[18] Damodar Reddy Edla, Kunal Mangalorekar, Gauri Dhavalikar, and Shubham Dodia. Classification of eeg data for human mental state analysis using random forest classifier. *Procedia computer science*, 132:1523–1532, 2018.

[19] Jincai Zhang, Mei Wang, and Yuancheng Li. Support vector machine of particle swarm optimization for motor imagery recognition using feature fusion. In *2021 10th International Conference on Computing and Pattern Recognition*, pages 51–57, 2021.

[20] Anurag Tiwari and Amrita Chaturvedi. A multiclass eeg signal classification model using spatial feature extraction and xgboost algorithm. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4169–4175. IEEE, 2019.

[21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.