Towards Adversarial Robustness with Early Exit Ensembles

Lorena Qendro‡ and Cecilia Mascolo‡

***Abstract*— Deep learning techniques are increasingly used for decision-making in health applications, however, these can easily be manipulated by adversarial examples across different clinical domains. Their security and privacy vulnerabilities raise concerns about the practical deployment of these systems. The number and variety of the adversarial attacks grow continuously, making it difficult for mitigation approaches to provide effective solutions. Current mitigation techniques often rely on expensive re-training procedures as new attacks emerge. In this paper, we propose a novel adversarial mitigation technique for biosignal classification tasks. Our approach is based on recent findings interpreting early exit neural networks as an ensemble of weight sharing sub-networks. Our experiments on state-of-the-art deep learning models show that early exit ensembles can provide robustness generalizable to various white box and universal adversarial attacks. The approach increases the accuracy of vulnerable deep learning models up to 60 percentage points, while providing adversarial mitigation comparable to adversarial training. This is achieved without previous exposure to the adversarial perturbation or the computational burden of re-training.**

I. Introduction

The plethora of sensors embedded on wearable and mobile devices has enabled affordable biosignal collection delivering fast solutions for monitoring physical and mental health. Biosignals, such as brain electroencephalography (EEG) or heart electrocardiography (ECG), can be non-invasively measured, allowing for automatic detection of many health conditions such as epilepsy [1], depression [3] or heart failures [5]. Using this rich data, deep learning techniques are revolutionizing medicine by achieving state-of-the-art performance on a variety of tasks within the medical field. Recent work, however, has shown that deep learning models misbehave in the presence of malign adversarial perturbations [6]. They can be easily manipulated since current implementations overlook robustness towards adversarial attacks. Adversarial perturbations can very closely mimic physiologically plausible signal, and therefore, become virtually indistinguishable to human perception. Nevertheless, they can drastically decrease the deep learning model accuracy and introduce security and privacy vulnerabilities in safety-critical scenarios, such as health applications.

The multiplicity of attacks grows continuously and traditional approaches that require retraining and re-deployment of the network are a significant burden and often not feasible. Currently, the state-of-the-art mitigation approaches rely on adversarial training [11] to minimize the risk of misclassifying the perturbed signal. However, its side-effects are not negligible since classifiers trained with adversarial examples learn fundamentally different representations compared to standard classifiers reducing accuracy [16]; they also can cause disparity on accuracy between classes for both clean and adversarial samples [18]. In addition, they are (i) resource-consuming, (ii) mostly mitigate only against the attacks they have been exposed to during training, and (iii) cannot be applied to already deployed or trained networks, suggesting limited generalization and applicability in the real world.

To solve the aforementioned issues, we propose a mitigation technique which relies on recent findings interpreting early exit neural networks (NNs) as an implicit ensemble of models [13]. Early exit neural networks are a class of conditional computation models that exit once a criterion (e.g., sufficient accuracy) is satisfied in order to save on computation [14]. Under a different interpretation, the early exit paradigm can be used to mitigate the problem of model overthinking [10] or used as an ensemble for uncertainty quantification [13]. Our intuition on exploiting these networks for adversarial mitigation arises from the success of ensemble defense in improving both accuracy and robustness [15], however, without the computational burden of training and maintaining multiple single models.

We summarize the main contributions of our paper as follows:

- We introduce a novel adversarial mitigation technique which provides run-time robustness to adversarial attacks it has never been exposed to before. Our approach is generalizable to various adversarial attacks, while maintaining a low computational burden.
- We evaluate our method on state-of-the-art deep learning architectures applied to biosignal classification tasks. Our results show that we can increase the accuracy up to 60 percentage points (pp) compared to undefended deep learning models, as well as, provide well-calibrated adversarial mitigation comparable to adversarial training.
- Our experiment on four major adversarial attacks, show the potential of early exit ensembles in providing adversarial robustness and how we can exploit its orthogonality to adversarial training to build robust models we can trust.

II. Method

Early exit ensembles are a collection of weight sharing sub-networks created by adding exit branches to any backbone neural network architecture. During inference, they provide an ensemble of predictions in a single forward pass

‡ Department of Computer Science and Technology, University of Cambridge, United Kingdom. lq223@cl.ac.cam.uk

which allows for efficient and robust adversarial mitigation by aggregating the predictions from individual exits.

Formally, any neural network $f_\theta(\cdot)$ can be decomposed into B blocks such that $f_\theta(\mathbf{x}) = (f^{(B)} \circ f^{(B-1)} \circ \cdots \circ f^{(1)})(\mathbf{x})$ where $(f^{(i)} \circ f^{(j)})(\mathbf{x}) = f_{\theta_i}(f_{\theta_j}(\mathbf{x}))$ denotes function composition for $i \neq j$ and $\theta = \cup_{i=1}^{B} \theta_i$, $\mathbf{x} \in R^D$ denotes a D-dimensional input. Let $\mathbf{h}^{(i)} \in R^{K_i \times D_i}$ denote the intermediary output of the i-th block having K_i features of dimension $D_i \leq D$ such that $\mathbf{h}^{(i)} = f_{\theta_i}(\mathbf{h}^{(i-1)})$ for $1 \leq i \leq B-1$, and $\mathbf{h}^{(0)} = \mathbf{x}$.

A. Early Exit Ensemble

An early exit block is defined as a NN $g_{\phi_i}(\cdot)$ which takes as input the intermediary output $\mathbf{h}^{(i)}$ from the i-th block of $f_\theta(\cdot)$, henceforth referred to as the backbone. Each exit block learns a predictive distribution $p_{\phi_i}(y|\mathbf{x}) = \sigma(g_{\phi_i}(\mathbf{h}^{(i)}))$ where $\sigma(\cdot)$ is the softmax transform and $y \in \{1, \ldots, C\}$ a corresponding discrete target taking one of C classes. As such, any NN is able to output a set $\mathcal{M} = \{p_{\phi_1}(y|\mathbf{x}), \ldots, p_{\phi_{B-1}}(y|\mathbf{x}), p_\theta(y|\mathbf{x})\}$ which represents an early exit ensemble. The ensemble $\mathcal{M}$ contains up to $B-1$ distributions from early exits blocks, in addition to the standard output from its final block. As such, ensemble size $|\mathcal{M}| = B$.

During training a weighted sum of each exits' individual predictive loss is optimized. This procedure allows the training of the whole ensemble jointly. More formally:

$$\mathcal{L} = L_{CE}(y, f_\theta(y|\mathbf{x})) + \sum_{i=1}^{B-1} \alpha_i L_{CE}(y, g_{\phi_i}(y|\mathbf{x})) \quad (1)$$

where $L_{CE}(\cdot, \cdot)$ is the cross-entropy loss function and $\alpha_i \in [0, 1]$ is a weight hyperparameter corresponding to the relative importance of each exit.

During inference, a single forward pass of a NN with early exits produces an ensemble $\mathcal{M}$ of predictions. The overall prediction from $\mathcal{M}$ can be computed as the mean of a categorical distribution obtained from averaging the predictions from the individual exits:

$$p_{\theta_\mathcal{M}}(y|\mathbf{x}) \approx \frac{1}{|\mathcal{M}|}\big(p_\theta(y|\mathbf{x}) + \sum_{i=1}^{B-1} p_{\phi_i}(y|\mathbf{x})\big). \quad (2)$$

B. Exit Block Architecture

Exits from earlier blocks inherit intermediary outputs with weaker representational capacity, which negatively impacts ensemble accuracy. To address this issue, we design a conditional architecture for the i-th exit block as follows:

$$g_{\phi_i}(\mathbf{h}^{(i)}) = \begin{cases} \mathbf{W}_2^{(i)} \rho(\mathbf{W}_1^{(i)} s(\mathbf{h}^{(i)})), & \gamma > 0 \\ \mathbf{W}_3^{(i)} s(\mathbf{h}^{(i)}), & \gamma = 0 \end{cases} \quad (3)$$

where $s(\cdot)$ denotes global average pooling, $\rho(\cdot)$ is an activation function, $\mathbf{W}_1^{(i)} \in R^{K_\gamma \times K_i}$, $\mathbf{W}_2^{(i)} \in R^{C \times K_\gamma}$, $\mathbf{W}_3^{(i)} \in R^{C \times K_i}$ are the weights the of linear layers (biases are omitted to enhance notation clarity). The hyperparameter $\gamma \geq 0$ is a learning capacity factor used to increase the number of features from K_i to K_i^γ of the i-th intermediary output such that $K_i^\gamma = (\sqrt{1+\gamma})^{B-i}$ for $1 \leq i \leq B-1$ where K_B is the number of features in the last block defined by the backbone. Intuitively, when $\gamma > 0$ the number of features in each exit block is inversely proportional to the exit point i.e. earlier exits use additional parameters to learn more complex features.

III. Adversarial attacks

In this work, we consider 4 types of adversarial attacks: three white box attacks (PGD, PGD-AVG, and PGD-MAX) aimed at early exit neural networks [9] and one universal adversarial perturbation attack [4]. Given a clean signal $\mathbf{x}$, an adversarial attack introduces small perturbations ∇ such that the prediction for $\mathbf{x}$ and $\mathbf{x}^{adv}$ differs.

Projected Gradient Descent (PGD) in an iterative universal first-order attack which aims to find the adversarial perturbations by moving in the opposite direction to the gradient of the loss function $L(\mathbf{x}, y)$ w.r.t. the signal (∇):

$$\mathbf{x}_0^{adv} = \hat{\mathbf{x}}, \mathbf{x}_{n+1}^{adv} = clip_{\mathbf{x}}^{\epsilon}\{\mathbf{x}_n^{adv} + \beta sign\big(\nabla_{\mathbf{x}} L\big(\mathbf{x}_n^{adv}, y\big)\big)\} \quad (4)$$

where ϵ is the step size which restricts the l_∞ of the perturbation, and

$$\hat{\mathbf{x}} = \mathbf{x} + \epsilon_1 * sign\big(\mathcal{N}\big(\mathbf{0}^d, \mathbf{I}^d\big)\big) \quad (5)$$

(with parameters ϵ_1, ϵ such as $\epsilon_1 < \epsilon$) is an additional prepended random step which avoids going towards a false direction of ascent. This represents a single attack defined to fool $f_\theta(\cdot)$ only, without considering the intermediate exits $g_{\phi i}$, expressed as:

$$\mathbf{x}^{adv} = \underset{\mathbf{x} \in |\mathbf{X}' - \mathbf{x}|_\infty \leq \epsilon}{arg\,max} |L(f_\theta(\mathbf{x}', y))| \quad (6)$$

PGD Average Attack (PGD-AVG) considers all intermediate exits $g_{\phi i}$ by maximizing the average of all losses such that the adversarial sample $\mathbf{x}^{adv}$ can attack any exit in the ensemble:

$$\mathbf{x}_{avg}^{adv} = \underset{\mathbf{x} \in |\mathbf{x}' - \mathbf{x}|_\infty \leq \epsilon}{arg\,max} \big|\frac{1}{B}\big(L(f_\theta(\mathbf{x}', y)) + \sum_{i=1}^{B-1} L(g_{\phi_i}(\mathbf{x}', y))\big)\big| \quad (7)$$

PGD Max-Average Attack (PGD-MAX) emphasizes on the individual exits, not just at maximizing an all-averaged loss like in eq. 7. It creates M single attacks $\mathbf{x}^{adv}$ such as eq. 6 and denoting the collection of single attacks as Ω:

$$\mathbf{x}_{max}^{adv} \leftarrow \mathbf{x}_{j^*}^{adv}, where\, \mathbf{x}_{j^*}^{adv} \in \Omega$$
$$j^* = \underset{j}{arg\,max} \big|\frac{1}{B}\big(L(f_\theta(\mathbf{x}^{\mathbf{adv}}, y)) + \sum_{i=1}^{B-1} L(g_{\phi_i}(\mathbf{x}_\mathbf{j}^{\mathbf{adv}}, y))\big)\big| \quad (8)$$

This attack is the strongest of the three since $\mathbf{x}_{max}^{adv}$ not only maximally fools the individual exit but is also transferable between exits.

Universal Adversarial Perturbation (UAP) seeks a perturbation ∇ to fool $f_\theta(\cdot)$ on most data samples:

$$f_\theta(\mathbf{x} + \nabla) \neq f_\theta(\mathbf{x}) \quad (9)$$

where perturbations are constrained to $l_\infty \leq \epsilon$ to be visually imperceptible to humans. UAP aims at crafting a single perturbation for all data samples. The version used in this work is based on DeepFool [12].

	ECG - FCNet (F1/ECE)			**EEG - VGG16** (F1/ECE)		
Mitigation	**No attack**	**UAP**	**PGD**	**No attack**	**UAP**	**PGD**
No mitigation	0.98 / 0.01	0.70 / 0.16	0.33 / 0.27	0.81 / 0.11	0.76 / 0.05	0.18 / 0.66
PGD AT	0.92 / 0.07	0.93 / 0.03	0.73 / 0.13	0.80 / 0.03	0.80 / 0.03	0.82 / 0.05
Ours	0.99 / 0.01	0.89 / 0.15	0.77 / 0.11	0.85 / 0.03	0.80 / 0.06	0.81 / 0.04

TABLE I

ADVERSARIAL PERTURBATIONS' IMPACT ON F1 SCORE AND ECE FOR UNDEFENDED, PGD 6 ADVERSARIAL TRAINING (AT) AND OUR APPROACH ON THE TWO MODEL-DATASET COMBINATIONS. PGD: $\epsilon = 100mV$ AND $\epsilon = 10mV$ FOR EEG AND ECG, RESPECTIVELY, AND $\beta = \epsilon/4$ ($iterations = 20$, $AT\_iterations = 10$). UAP: $min\_fool\_rate = 0.8$ AND $sample\_size = 100$.

	ECG - FCNet (F1/ECE)			**EEG - VGG16 (F1 / ECE)**		
Mitigation	**No attack**	**PGD-AVG**	**PGD-MAX**	**No attack**	**PGD-AVG**	**PGD-MAX**
Ours	0.99 / 0.01	0.71 / 0.10	0.55 / 0.11	0.85 / 0.03	0.76 / 0.04	0.48 / 0.24
Ours + PGD AT	0.92 / 0.11	0.95 / 0.06	0.94 / 0.06	0.83 / 0.07	0.80 / 0.04	0.82 / 0.07
Ours + PGD-AVG AT	0.93 / 0.18	0.87 / 0.10	0.86 / 0.08	0.82 / 0.11	0.81 / 0.10	0.81 / 0.09
Ours + PGD-MAX AT	0.92 / 0.28	0.90 / 0.03	0.86 / 0.01	0.82 / 0.04	0.78 / 0.30	0.82 / 0.04

TABLE II

ADVERSARIAL PERTURBATIONS' IMPACT ON F1 SCORE AND ECE FOR OUR METHOD AND OR METHOD COMBINED WITH TRADITIONAL PGD ADVERSARARIAL TRAINING (AT) AND EXIT AWARE AT (PGD-AVG, PGD-MAX).

IV. EXPERIMENTS

Datasets & architectures. For the evaluation, two publicly-available biometric signal statasets are used: Electrocardiogram heart attack (ECG) [2], and Electroencephalogram artifacts (EEG) [7] where only eye-movement artifacts are considered. All datasets are split into $80\%/10\%/10\%$ train/validation/test maintaining class proportions. Each dataset is paired with a different architecture: FCNet [17] (fully-convolutional 5-layer network) for ECG and VGG16 for EEG.

Electrocardiogram heart attack (ECG) [2] is a dataset of univariate timeseries of ECG signals of length 140 extracted from a single patient. Each signal falls into one of 5 classes which are combined to make two labels: Normal (N) and Abnormal (R-on-T, PVC, SP, UB).

Electroencephalogram eye movement artifact (EEG) [7] consists of univariate timeseries of length 2000 extracted from 213 patients from the Temple University Artifact Corpus (v2.0). A set of 21 EEG channels were retained from all patients and signals were resampled to 250 Hz. All EEG signals were bandpass filtered (0.3-40 Hz) using a second-degree Butterworth filter and notch filtered at the power lower frequency 60 Hz. Segments of clean and eye movement artifact signal were used for this specific task.

Metrics. Performance is evaluated using class weighted F1 and expected calibration error (ECE). While F1 indicates model accuracy, ECE measures model calibration as the expected difference between accuracy and predicted confidence. $\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$ where accuracy and confidence for bin B_m are

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{b \in B_m} 1(\hat{y}_b = y_b)$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{b \in B_m} \hat{p}_b$$

such that $\hat{y}_n = \arg\max_{c \in \{1,\dots,C\}} p_\theta(y_b = c|\mathbf{x}_b)$ is the predicted class. There are M bins of size $1/M$ and n samples, and bin B_m covers the interval $(\frac{m-1}{M}, \frac{m}{M}]$. Confidence $\hat{p}_b$ is the probability of the top model prediction for sample b.

Baselines. We compare early exit ensembles against the undefended backbone model without any exits (No mitigation) and backbone model trained completely with PGD adversarial attacks (PGD AT). Additionally, we provide analysis on adversarial training applied to early exit ensembles where the early exit ensemble unified model has been trained with adversarial examples and at test time the averaging of the predictions from the individual exits is applied. For (Ours + PGD AT) the model training procedure perturbed training samples having access to only the last output of the early exit model, while for (Ours + PGD-ADV AT) and (Ours + PGD-MAX AT) it had access to all individual exits, members of the ensemble. The latter two attacks are not applied to the backbone (without exit) model since they would be the same as a naive PGD since the ensemble size is 1.

Hyperparameters. All models are trained using the Adam optimizer and an optimally tuned learning rate, batch size, and epochs: FCNet ($1e^{-2}$, 200, 250), and VGG16 ($1e^{-4}$, 200, 200). All results for our approach are based on a loss with α_i =1 as well as a learning capacity factor and exit strategy optimally tuned as follows: FCNet (γ=0.0, *Block-wise*), and VGG16 (γ=0.5, *Semantic*) with an ensemble size $M = 5$ as suggested in [13]. To prevent overfitting, early-stopping is based on the best validation accuracy with a patience of 5.

For the PGD (incl. AVG and MAX) adversarial attacks 20 maximum iterations are used. The perturbation strength is set at $\epsilon = 100mV$ and $\epsilon = 10mV$ for EEG and ECG, respectively, while $\beta = \epsilon/4$ for both. PGD attacks (especially on EEG data) can bring strong and perceptible square-wave displacements, which would allow both expert and non-experts to distinguish them. With the aforementioned

values, we provide the strongest perturbations possible while keeping the attack imperceptible to at least a non-expert eye. For UAP, we used the default values in the original work [4] $min\_fool\_rate = 0.8$ and a $sample\_size = 100$. For adversarial training, we use the same configurations as the attacks with 10 iterations for the PGDs.

V. Results

Table I summarizes the accuracy (as measured by the F1 score) and the expected calibration error (ECE) results when the models are under adversarial attack. As expected, our technique provides better accuracy on clean data (no attack) compared to both the undefended (no mitigation) model provided by the implicit ensemble paradigm. Additionally, PGD AT shows a lower accuracy on clean data demonstrating one of the disadvantages of adversarial training as mentioned in Section I. As an universal adversarial attack, UAP, degrades accuracy less than PGD, however, early exit ensembles can significant mitigation on it with a 19pp improvement on ECG-FCNet. The PGD attack applied in this scenario, is a white box attack which can see only the final output, allowing for the intermediate exits in the ensemble to disagree with the attacked exit and provide the desired mitigation. Compared to the undefended model, our technique improves accuracy by 44pp and 63pp for ECG-FCNet and EEG-VGG16, respectively. These results show that early exit ensembles can provide comparable (even better in some instances) mitigation to the state-of-the-art adversarial training without any previous exposure to the adversarial attack presenting a general (non attack-specific) approach.

In a second scenario, Table II, shows the impact of adversarial attacks which are aware of the intermediate exits of the model and maximize the overall loss accordingly. Here, early exit ensembles can still provide adversarial robustness, although the more the attacker aims the individual exits weaker the mitigation. The strength of our approach still relies on the fact that it does not depend on a specific attack or training procedure, making it a perfect technique to combine with adversarial training. In a real world deployment, early exit ensembles have the potential to have a high level of robustness provided by adversarial training of known attacks as well as robustness towards new and unknown attacks. As dicussed in Section III, PGD-MAX is the strongest attack given it transfereability characteristics. We can see this in our results too, where early exit ensembles benefit more from the PGDs adversarial training to compensate the accuracy loss caused by PGD-MAX. However, contrary to the lightweight approach provided by early exit ensembles, PGD-MAX is very expensive to produce, circa 10x more expensive than PGD and PGD-AVG. Crafting a PGD-MAX attack for training purposes on a VGG16 model (batch size of 250 samples), costs 475 seconds on an Nvidia Tesla-V100 GPU, while PGD and PGD-AVG cost 38 and 48 seconds, respectively [8]. Early exit ensembles, instead, only add a slight memory overhead of 16% and computation overhead of 12% in the worse case scenario (VGG16).

VI. Conclusions

In this paper, we propose a novel adversarial mitigation technique which provides robustness exploiting early exit ensembles. Our approach achieves remarkable performance and efficiency trade-offs, comparable to the state-of-the-art adversarial training on various biosignal classification tasks. We believe that the ease of implementation of our method, its orthogonality to adversarial training and the promising results offer the foundations for further exploring early exit ensembles in the adversarial machine learning context.

References

[1] U Rajendra Acharya et al. Automated eeg analysis of epilepsy: a review. *Knowledge-Based Systems*, 45, 2013.

[2] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6, 2019.

[3] Fernando Soares de Aguiar Neto et al. Depression biomarkers using non-invasive eeg: A review. *Neuroscience & Biobehavioral Reviews*, 105:83–93, 2019.

[4] Moosavi-Dezfooli et al. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[5] Siontis C. Konstantinos et al. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7), 2021.

[6] Samuel G Finlayson et al. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[7] Ahmed Hamid, Katherine Gagliano, Safwanur Rahman, Nikita Tulin, Vincent Tchiong, Iyad Obeid, and Joseph Picone. The temple university artifact corpus: An annotated corpus of eeg artifacts. In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 2020.

[8] Sanghyun Hong et al. A panda? no, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference. *arXiv preprint arXiv:2010.02432*, 2020.

[9] Ting-Kuei Hu, Tianlong Chen, Haotao Wang, and Zhangyang Wang. Triple wins: Boosting accuracy, robustness and efficiency together by enabling input-adaptive inference. *arXiv preprint arXiv:2002.10025*, 2020.

[10] Yigitcan Kaya et al. Shallow-deep networks: Understanding and mitigating network overthinking. In *International Conference on Machine Learning*. PMLR, 2019.

[11] Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[12] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[13] Lorena Qendro, Alexander Campbell, Pietro Lio, and Cecilia Mascolo. Early exit ensembles for uncertainty quantification. In *Machine Learning for Health*.

[14] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016.

[15] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[16] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[17] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017.

[18] Han Xu, Xiaorui Liu, Yaxin Li, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. *arXiv preprint arXiv:2010.06121*, 2020.