Universal Adversarial Perturbations in Epileptic Seizure Detection

Amir Aminifar
Department of Information Technology
Uppsala University, Sweden
amir.aminifar@it.uu.se

***Abstract*—Adversarial examples have received a lot of attention over the past decade, particularly with the rise of deep neural networks. Adversarial manipulation of sensitive health-related information, e.g., if such information is used for prescribing medicine, may have irreversible consequences, involving patients' lives. In this article, we consider adversarial perturbations in the context of medical and health applications and focus on the epileptic seizure detection problem. We formulate an optimization problem for computing universal adversarial perturbations and show that such universal perturbations may be used to declare the majority of seizure samples as non-seizure, i.e., to fool the classification algorithm, while being imperceptible to the medical expert eye.**

***Index Terms*—Universal Adversarial Perturbation, Epileptic Seizure Detection, Epileptic Ictal Activity**

I. Introduction

Adversarial perturbations have been discussed in the literature over the past decade to demonstrate the vulnerabilities of modern machine-learning classifiers, both in terms of theory [1]–[3] and applications [4]–[6]. These vulnerabilities highlight that very small carefully-identified perturbation vectors exist and may cause data samples in many applications, e.g., image classification, to be misclassified. The adversarial perturbations are not only important in terms of reliability and robustness of the machine-learning classifiers, but also in terms of safety and security of their users. In line with this observation, several studies acknowledge such threats and attempt to detect/defend such adversarial perturbations [7]–[9].

Adversarial manipulation of sensitive health-related information, e.g., if used to prescribe medicine, may have irreversible consequences, involving patients' lives. In this article, we consider the problem of epileptic seizure detection as a real-world case study to demonstrate the importance of such adversarial perturbations. Epilepsy is a chronic neurological disorder affecting more than 50 million people worldwide [10] and is ranked number four after migraine, Alzheimers disease, and stroke [11]. Epilepsy is manifested by recurrent unprovoked seizures and the symptoms include behavioral arrest, rigid extension of limbs, automatic movements and severe body convulsions. The unpredictability of seizures not only degrades the quality of life of the patients, but can also be life-threatening. Several previous studies have considered the epileptic seizure detection and monitoring problem based on machine-learning techniques [12]–[21].

Modern systems monitoring the electroencephalography (EEG) signals are being currently developed with the view to detect epileptic seizures in order to alert caregivers in real time and reduce the impact of seizures on patients' quality of life. This is, for instance, possible using wearable and mobile-health technologies, e.g., e-Glass sensor [16], to monitor the brain activities of the patients in real time and inform family members, caregivers, and emergency units for rescue in case of seizures. However, such seizures, if missed to be detected, e.g., due to adversarial attacks, may even jeopardize patient's life.

At the same time, the information collected from each patient is used by the medical experts to develop a better understanding of such health pathologies and, in turn, diagnosis, prognosis, and treatment. In particular, epileptologists administer drugs based on the frequency and duration of the seizures. Therefore, if the adversary is able to manipulate the biosignals, e.g., by introducing adversarial perturbations to mask certain seizures, the medical experts will then prescribe according to this manipulated biosignals, which may clearly have irreversible consequences.

In this article, we consider adversarial perturbations in the context of medical and health applications for the epileptic seizure detection problem based on the EEG signals. We formulate an optimization problem to identify the minimum universal perturbations required by the adversaries to declare the majority of ictal (seizure) samples as non-ictal (non-seizure), i.e., to fool the classification algorithm, while being imperceptible to the medical expert eye. We evaluate our proposed universal perturbation scheme based on the CHB-MIT scalp EEG database [22] and demonstrate the effectiveness of the proposed scheme.

The remainder of this article is structured as follows. In Section II, we present the motivational example for adversarial perturbations in the case of epileptic seizure detection problem. In Section III, we formulate the problem as an

This work has been partially supported by the Swiss National Science Foundation (ML-edge project: Ref. 182009) and the Swedish Promobilia Foundation (WiTNESS project: Ref. 18079).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

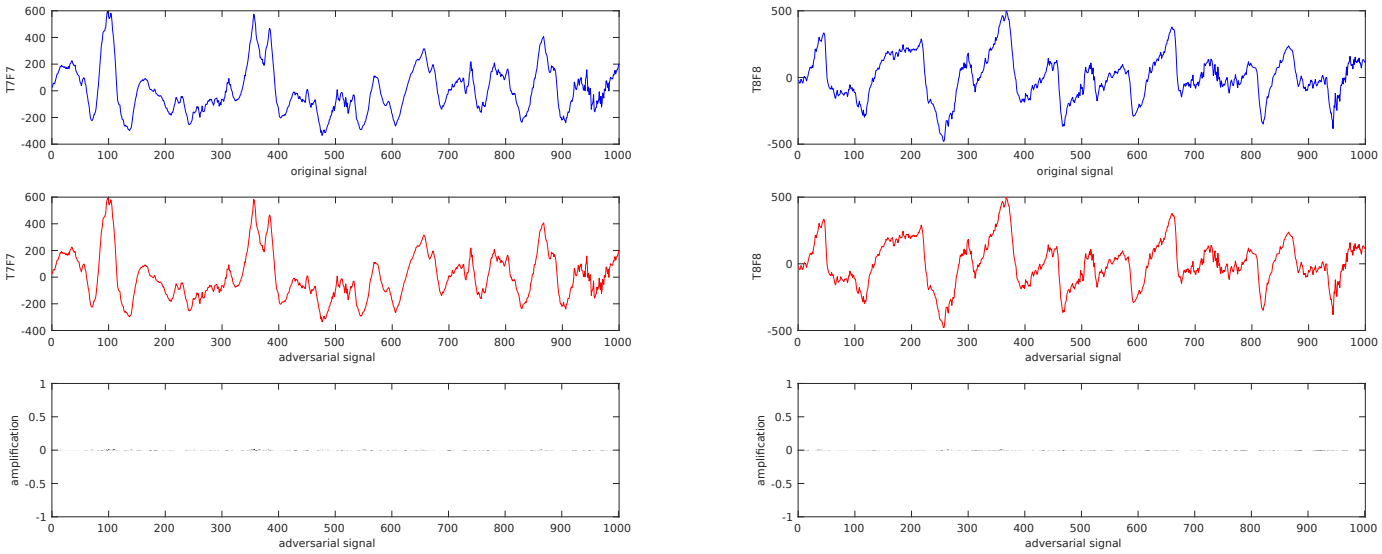

Fig. 1. The original signal (blue) and the adversarial signal (red) for two channels T7F7 and T8F8. The amplification factor per sample is shown in black. Notice that there are very small (but non-zero) amplification coefficients around 100 and just after 300 on the left side (T7F7), while the changes in the adversarial signal are imperceptible to the human eye.

optimization problem and describe in detail our proposed universal adversarial perturbations. In Section IV, we present the experimental setup and evaluate the proposed universal adversarial perturbation scheme. Finally, Section V serves as a conclusion.

II. Motivational Example

Let us consider the epileptic seizure detection problem, to identify seizures by monitoring the raw EEG signals for synchronous rhythmic activities. The standard 10–20 EEG acquisition system include as many as 19 electrodes on the scalp [23]. For the simplicity of the presentation, let us consider only two channels T7F7 and T8F8, which have been proven to be essential for epileptic seizure detection and are adopted in the state-of-the-art wearable sensors [16].

The original EEG signals and the corresponding minimal adversarial signals, based on our previous work in [24], for two channels T7F7 and T8F8 are shown in Figure 1. We define the minimal adversarial perturbation as the minimum manipulation required to declare a specific ictal sample as non-ictal. The presence of the well-known delta–theta rhythm, i.e., rhythmic slow activity with a frequency of oscillation in 0.5–4 or 4–7 hertz, is a clear indication of the ictal discharge and epileptic seizure in the EEG signals [25]. Notice that there are only slight differences between the original and the corresponding adversarial signals, e.g., the very small amplification coefficients around 100 and just after 300 on the left side (in T7F7), while the changes in the adversarial signal are imperceptible to the human eye. Nevertheless, the adversary is able to mask the seizure by slight manipulation of the original signals. The amplification factor for each sample is also shown Figure 1. We observe that the majority of the samples remain approximately the same.

The original EEG signals and the corresponding universal adversarial signals for two channels T7F7 and T8F8 are shown in Figure 2. We define universal adversarial perturbation as the minimum perturbation required to declare the majority of the ictal samples as non-ictal with high probability. Notice that there are slight differences between the original and adversarial signals, e.g., the peak just before 400 on the right side (in T8F8). Nevertheless, the well-known delta–theta rhythm, which is a clear indication of the ictal discharge and epileptic seizure in the EEG signals, is preserved in the adversarial signal. The amplification factor for each sample is also shown in Figure 2. Finally, we evaluate this universal perturbation for all (unseen) seizure signals available for this patient and observe a success rate of 90.20%. That is, considering this universal perturbation, 90.20% of the ictal samples are misclassified as non-ictal.

III. Universal Adversarial Perturbations

In this section, we discuss our proposed universal adversarial perturbation scheme. In Section III-A, we introduce the seizure detection problem and discuss the classification algorithm for the detection of epileptic seizures. In Section III-B, we formulate a convex optimization problem to identify the minimal perturbations required for the misclassification of one single seizure sample as non-seizure. We extend this formulation to identify the universal adversarial perturbations for the misclassification of the majority of seizure samples as non-seizure with high probability.

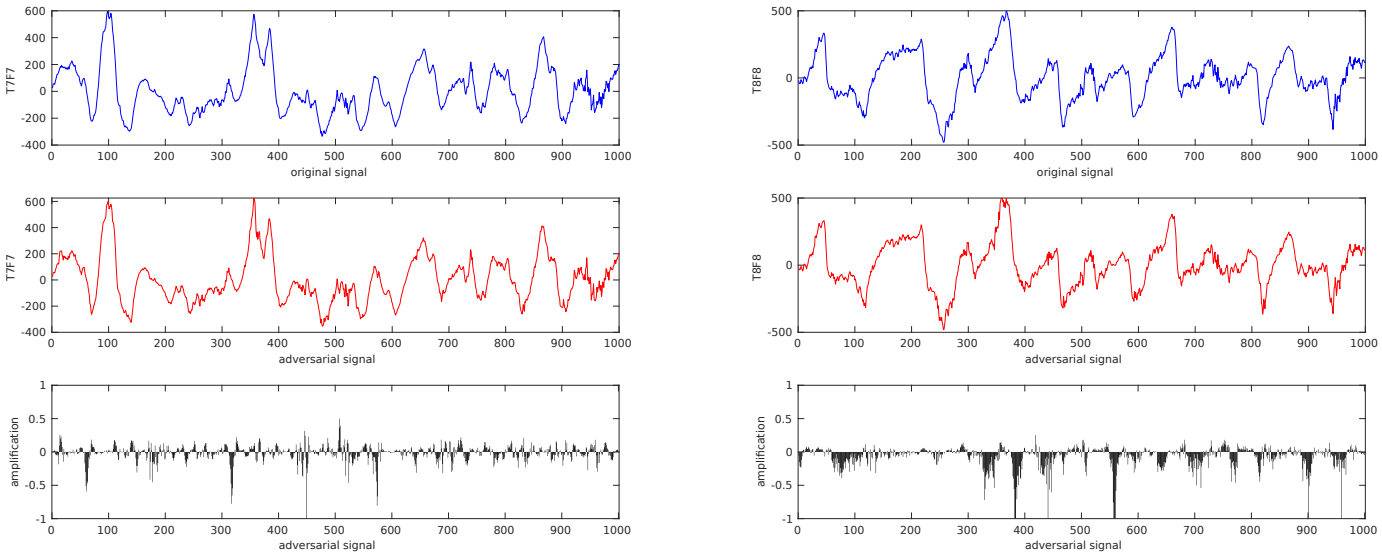

Fig. 2. The original signal (blue) and the universal adversarial signal (red) for two channels T7F7 and T8F8. The amplification factor per sample is shown in black. Notice that there are slight differences between the original and adversarial signals, e.g., the peak just before 400 on the right side (T8F8).

A. Seizure Detection Classification

Let us consider the Support Vector Machine (SVM) classification for the epileptic seizure detection problem [26]. The original formulation of the SVM classification with soft-margin is as follows,

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \lambda \sum_{i=1}^{n} |\xi_i| \\ \text{s.t.} \quad & y_i \cdot (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1 - \xi_i, \qquad i \in [1, n], \end{aligned} \tag{1}$$

where $\boldsymbol{x}_i$ is sample i and $y_i \in \{-1, +1\}$ is its corresponding label. The soft-margin slack variables are denoted by ξ_i for sample i. The total number of data samples is denoted by n. Finally, the hyperplane that separates the ictal and non-ictal samples is captured by $\boldsymbol{w}$ and b. The SVM classification algorithm in its original form is unable to capture the complexity of the seizure detection problem. Therefore, we consider a simple transformation and use $\boldsymbol{x}_i \odot \boldsymbol{x}_i$ instead of $\boldsymbol{x}_i$, as follows,

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \lambda \sum_{i=1}^{n} |\xi_i| \\ \text{s.t.} \quad & y_i \cdot (\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i \odot \boldsymbol{x}_i) + b) \geq 1 - \xi_i, \qquad i \in [1, n], \end{aligned} \tag{2}$$

where the operator $\odot$ is the element-wise multiplication between two vectors.

B. Universal Adversarial Manipulation

We shall first focus on the minimal adversarial perturbations for the misclassification of one single seizure sample. Let us consider the multiplicative adversarial manipulation model, which is formulated as follows,

$$\begin{aligned} \min_{\boldsymbol{a}} \quad & \|\boldsymbol{a} - \mathbf{1}\|_2^2 \\ \text{s.t.} \quad & y_i \cdot (\boldsymbol{w}^{\mathrm{T}}((\boldsymbol{x}_i \odot \boldsymbol{a}) \odot (\boldsymbol{x}_i \odot \boldsymbol{a})) + b) < 0, \end{aligned} \tag{3}$$

where $\boldsymbol{a}$ is the multiplicative perturbation vector. Our objective is to minimize the magnitude of the adversarial perturbation vector, which is capture by $\|\boldsymbol{a} - \mathbf{1}\|_2$. Unfortunately, however, the above optimization problem is not guaranteed to be convex because of its constraint.

To address this problem, we reformulate the above optimization problem as follows,

$$\begin{aligned} \min_{\hat{\boldsymbol{a}}} \quad & \|\hat{\boldsymbol{a}} - \mathbf{1}\|_2 \\ \text{s.t.} \quad & y_i \cdot (\boldsymbol{w}^{\mathrm{T}}((\boldsymbol{x}_i \odot \boldsymbol{x}_i) \odot \hat{\boldsymbol{a}}) + b) < 0, \end{aligned} \tag{4}$$

which is a convex optimization problem and can be solved to find the minimal adversarial perturbations required for the misclassification of seizure sample $\boldsymbol{x}_i$. Observe that the objective function is the classical L_2 norm, which is convex, and the constraints are linear with respect to variable $\hat{\boldsymbol{a}}$. The adversarial perturbation vector is captured by $\hat{\boldsymbol{a}}$.

We shall now extend this optimization problem to identify the universal adversarial perturbations, i.e., the adversarial perturbation vector that render the majority of the ictal samples to be misclassified as non-ictal samples. To achieve this, we introduce a constraint for each ictal sample in the training set that is used to learn this universal perturbation vector as follows,

$$\begin{aligned} \min_{\hat{\boldsymbol{a}}} \quad & \|\hat{\boldsymbol{a}} - \mathbf{1}\|_2 \\ \text{s.t.} \quad & y_i \cdot (\boldsymbol{w}^{\mathrm{T}}((\boldsymbol{x}_i \odot \boldsymbol{x}_i) \odot \hat{\boldsymbol{a}}) + b) < 0, \ \ i \in (n, m]. \end{aligned} \tag{5}$$

Note that the above optimization remains a convex optimization problem and can be solved to find the universal adversarial perturbations required for the misclassification of all seizure samples $\boldsymbol{x}_i$, where $i \in (n, m]$.

Such an optimization problem may lead to universal adversarial perturbations that are too aggressive, manipulating the

signals to the extent that is perceptible to the expert eye. To address this issue, we propose a soft optimization problem and introduce a penalty function for those data samples for which the universal perturbation vector is not effective, as follows,

$$\begin{aligned} \min_{\hat{\boldsymbol{a}},\boldsymbol{\epsilon}} \quad & \|\hat{\boldsymbol{a}} - \mathbf{1}\|_2 + \mu \sum_{i=n+1}^{m} |\epsilon_i| \\ \text{s.t.} \quad & y_i \cdot (\boldsymbol{w}^{\mathrm{T}} ((\boldsymbol{x}_i \odot \boldsymbol{x}_i) \odot \hat{\boldsymbol{a}}) + b) < \epsilon_i, \quad i \in (n, m] \end{aligned} \tag{6}$$

where ϵ_i is the penalty for the correct classification of ictal data sample x_i, when considering the universal perturbation. On the other hand, μ is the parameter which controls the balance between the magnitude of the universal adversarial perturbation vector and the penalty for the correct classification of ictal data samples when applying the universal adversarial perturbations.

IV. Experimental Setup and Results

In this section, we evaluate the power of our proposed adversarial perturbation scheme in the case of epileptic seizure detection problem. In Section IV-A, we discuss the dataset used for the evaluation of our proposed universal adversarial perturbations scheme. In Section IV-B, we evaluate our proposed universal adversarial schemes and evaluate the trade-off between the success rate in misclassification of the ictal samples and the magnitude of the adversarial perturbations.

A. Epilepsy Dataset

We consider the CHB-MIT database [22] that contains EEG signals from 23 epilepsy patients with intractable seizures. All recordings are collected from children and young adults in the 1.5–22 age range. The dataset is annotated by the medical experts and contains a total of 182 seizures. These EEG signals are sampled at $F_s = 256$ Hz, with 16-bit resolution. Several previous studies indicate that the majority of seizures last longer than one minute, with an expected value of median equal to 71.9 seconds [27]. However, patients 6, 14, and 16 in this dataset suffer from seizures lasting 15.30 ± 2.87, 21.13 ± 8.68, and 8.40 ± 2.27 seconds, respectively, hence are not considered in our analysis. We consider only the two channels $T7F7$ and $T8F8$ in the e-Glass wearable system [16], which have been shown to be important for the detection of epileptic seizures.

B. Universal Adversarial Manipulation

In this section, we evaluate our proposed adversarial perturbation scheme based on the epileptic seizure samples. We shall first evaluate the possibility of successful manipulation of each single seizure sample, such that the seizure samples are misclassified as non-seizure. We also evaluate the effort required for the misclassification of an ictal sample as non-ictal. We evaluate the performance of our adversarial perturbation scheme based on the median success rate and the median required perturbation.

The success rate captures the number of seizure samples that could be successfully perturbed to be misclassified as non-seizure over the total number of original seizure samples classified correctly. The results are shown in Figure 3. The median success rate among all patients is 70.3%. In addition, in all cases, the proposed scheme is able to successfully perturb more than 40.0% of the seizure samples such that the classification algorithm misclassifies these seizure samples as non-seizure.

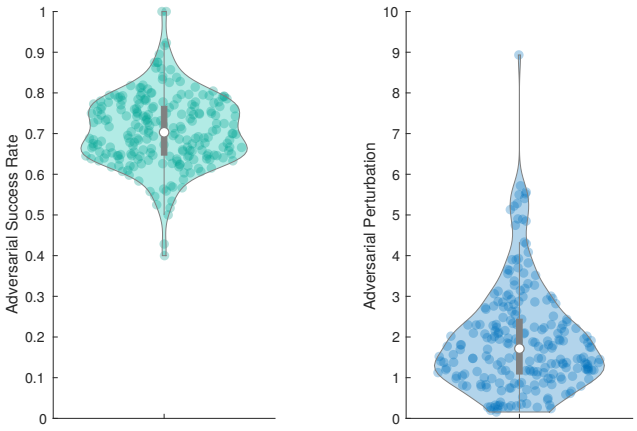

Fig. 3. The distribution of the success rate and minimum required perturbations for our proposed adversarial perturbations scheme when considering only one single ictal sample.

The median required perturbation captures the (median) effort required for the misclassification of one single seizure sample as non-seizure, which is captured by $\|\hat{\boldsymbol{a}} - \mathbf{1}\|_2$. The results are shown in Figure 3. The median value of $\|\hat{\boldsymbol{a}} - \mathbf{1}\|_2$ among all patients is 1.7, where the vector $\hat{\boldsymbol{a}}$ is of dimension 2000.

We shall now evaluate our proposed universal adversarial perturbation scheme and compare these results against the adversarial perturbations scheme when only considering one single seizure sample. We evaluate the success rate, i.e., the rate of successful perturbation of seizure samples to be misclassified as non-seizure samples considering our universal adversarial perturbation scheme. We also evaluate the effort required for the misclassification of ictal samples as non-ictal samples, which is captured by $\|\hat{\boldsymbol{a}} - \mathbf{1}\|_2$.

In Figure 4, we evaluate the median success rate and the median required perturbation versus the value of parameter μ. As discussed before, μ is the parameter which controls the balance between the magnitude of the universal adversarial perturbations and the penalty for the correct classification of ictal data samples under the universal adversarial perturbation. We observe that both the median success rate and the median required perturbation increase as we increase the value of μ. Note that, in Figure 4, we use a log-2 scale for the x-axis.

Let us now focus on $\mu = 0.01$. The median success rate among all patients is 76.9%. Observe that, in the worst-

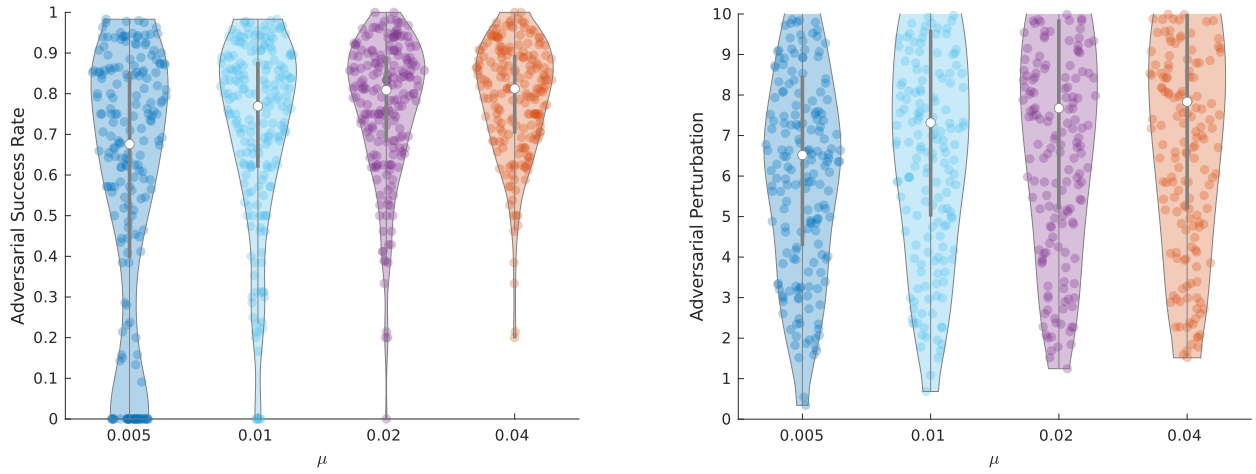

Fig. 4. The distribution of the adversarial success rate and minimum required perturbations for our universal adversarial perturbations scheme versus the value of μ in the optimization problem.

case scenario and for a minority of seizure samples, the universal adversarial perturbation might not be effective as shown in Figure 4. However, as we increase the value of μ, the number of seizure samples that are correctly classified under the universal adversarial perturbation reduces. In terms of the magnitude of our universal adversarial perturbation vector, the median value of $\|\hat{\boldsymbol{a}} - \mathbf{1}\|_2$ among all patients is 7.3, where the vector $\hat{\boldsymbol{a}}$ is of dimension 2000.

Note that the median success rate in the case of universal perturbation scheme with $\mu = 0.01$ is 76.9%, which is sightly larger than the median success rate in the adversarial perturbation scheme when considering only one single seizure sample, i.e., 70.3%. This is essentially because the median required perturbation in the case of universal perturbation scheme is 7.3, which is more than three times the median required perturbation in the adversarial perturbation scheme which considers only one single seizure sample, i.e., 1.7. This is due to the fact that the latter perturbation is specifically optimized for one single ictal sample, but the universal perturbation is carefully crafted to ensure the misclassification of the majority of the ictal samples. Therefore, the effort required in the universal adversarial perturbations is beyond that of the adversarial perturbation for one single ictal sample.

V. Conclusion

Machine-learning models have been proven to be vulnerable to adversarial perturbations of their input data and, in turn, misclassify such perturbed input data. The adversarial perturbations of sensitive health-related information, e.g., if such health-related data are used for prescribing medicine, may have irreversible consequences, involving patients' lives. In this article, we demonstrated the power of universal adversarial perturbations based on a real-world epileptic seizure detection problem. We formulated the problem as a convex optimization problem to quantify the effort required to declare the majority of seizure samples as non-seizure, i.e., the universal adversarial perturbations to fool the classification algorithm.

References

[1] 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, pp. 2574–2582.

[2] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[3] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[4] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

[5] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[6] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[7] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[8] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff, "On detecting adversarial perturbations," *arXiv preprint arXiv:1702.04267*, 2017.

[9] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 582–597.

[10] World Health Organization, "Epilepsy," 2020.

[11] D Hirtz, DJ Thurman, K Gwinn-Hardy, M Mohamed, AR Chaudhuri, and R Zalutsky, "How common are the common neurologic disorders?," *Neurology*, vol. 68, no. 5, pp. 326–337, 2007.

[12] Samanwoy Ghosh-Dastidar, Hojjat Adeli, and Nahid Dadmehr, "Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 2, pp. 512–518, 2008.

[13] Alexandros T Tzallas, Markos G Tsipouras, and Dimitrios I Fotiadis, "Epileptic seizure detection in eegs using time–frequency analysis," *IEEE transactions on information technology in biomedicine*, vol. 13, no. 5, pp. 703–710, 2009.

[14] Hasan Ocak, "Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2027–2036, 2009.

[15] Larbi Boubchir, Somaya Al-Maadeed, and Ahmed Bouridane, "On the use of time-frequency features for detecting and classifying epileptic seizure activities in non-stationary eeg signals," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5889–5893.

[16] D. Sopic, A. Aminifar, and D. Atienza, "e-Glass: a wearable system for real-time detection of epileptic seizures," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018.

[17] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.

[18] Damián Pascual, Amir Aminifar, and David Atienza, "A self-learning methodology for epileptic seizure detection with minimally-supervised edge labeling," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 764–769.

[19] Farnaz Forooghifar, Amir Aminifar, and David Atienza, "Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 6, pp. 1338–1350, 2019.

[20] Farnaz Forooghifar, Amir Aminifar, Leila Cammoun, Ilona Wisniewski, Carolina Ciumas, Philippe Ryvlin, and David Atienza, "A self-aware epilepsy monitoring system for real-time epileptic seizure detection," *Mobile Networks and Applications*, pp. 1–14.

[21] A. Thomas, A. Aminifar, and D. Atienza, "Noise-resilient and interpretable epileptic seizure detection," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020.

[22] Ali Hossam Shoeb, *Application of machine learning to epileptic seizure onset detection and treatment*, Ph.D. thesis, Massachusetts Institute of Technology, 2009.

[23] George H Klem, Hans Otto Lüders, HH Jasper, C Elger, et al., "The ten-twenty electrode system of the international federation," *Electroencephalogr Clin Neurophysiol*, vol. 52, no. 3, pp. 3–6, 1999.

[24] A. Aminifar, "Minimal adversarial perturbations in mobile health applications: The epileptic brain activity case study," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1205–1209.

[25] Jeffrey W Britton, Lauren C Frey, JL Hopp, P Korb, MZ Koubeissi, WE Lievens, EM Pestana-Knight, and EK Louis St, *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants*, American Epilepsy Society, Chicago, 2016.

[26] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[27] Sigmund Jenssen, Edward J Gracely, and Michael R Sperling, "How long do most seizures last? a systematic comparison of seizures recorded in the epilepsy monitoring unit," *Epilepsia*, vol. 47, no. 9, pp. 1499–1503, 2006.