MINIMAL ADVERSARIAL PERTURBATIONS IN MOBILE HEALTH APPLICATIONS: THE EPILEPTIC BRAIN ACTIVITY CASE STUDY

Amir Aminifar

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

ABSTRACT

Today, the security of wearable and mobile-health technologies represents one of the main challenges in the Internet of Things (IoT) era. 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 demonstrate the power of such adversarial attacks based on a real-world epileptic seizure detection problem. We identify the minimum perturbation required by the adversaries to declare a seizure (ictal) sample as non-seizure (inter-ictal) in emergency situations, i.e., minimal adversarial perturbation to fool the classification algorithm.

Index Terms— Adversarial Perturbation, Mobile Health, Privacy and Security, Epilepsy, Seizure Detection.

1. INTRODUCTION

Contrary to popular belief, the security and privacy of non-invasive mobile-health technologies are of paramount importance, even when there is no explicit close-loop intervention involved. Such vulnerabilities in mobile-health technologies are considered as security/privacy breaches and may even jeopardize the safety of patients, considering the inherent criticality of biomedical applications. However, due to the limited amount of resources (processing power, communication bandwidth, memory storage, and battery lifetime) available in such devices, security measures often serve as secondary design goals [1–3]. In fact, there are still a large proportion of Internet of Things (IoT) devices that do not even adopt encrypted communication or proper authentication [4].

The current poor security measures in the Internet of Things (IoT) devices allow a wide range of adversarial attacks. Adversarial attacks have been discussed in the literature, both in terms of theory [5,6] and applications [7–9]. The classical man-in-the-middle attack is perhaps one of the most common security threats in these scenarios. Such attacks become all the more important when considering wearable and mobile-health technologies for real-time health monitoring. The information acquired by these wearable sensors and mobile-health devices are often used for two main purposes: (1) to monitor the progress of the health pathology for better diagnosis, prognosis, and treatment and, (2) to inform accredited relatives and emergency units in real time for rescue. As a result, manipulation of such data/alerts, e.g., by intercepting/manipulating the communication between the wearable sensors and mobile devices or cloud engines, can have irreversible consequences and may even jeopardize patient's life.

In this article, we consider the case of epilepsy seizure detection problem as a real-world case study to demonstrate the importance of such adversarial attacks. Epilepsy is a chronic neurological disorder affecting more than 65 million people worldwide and manifested by recurrent unprovoked seizures [10]. The unpredictability of seizures not only degrades the quality of life of the patients, but it can also be life-threatening. Modern systems monitoring electroencephalography (EEG) signals are being currently developed with the view to detect epileptic seizures in order to alert caregivers and reduce the impact of seizures on patient's quality of life [11–14]. However, such seizures, if missed to be detected, e.g., due to adversarial man-in-the-middle attacks, may have irreversible consequences, involving patients lives. Here, we identify the minimum perturbation by the adversaries required to declare an ictal (seizure) sample as inter-ictal (non-seizure) in emergency situations, i.e., minimal adversarial perturbation to fool the classification algorithm, for the adversary to remain stealthy to ensure maximum consequences for a prolonged period of time.

2. ADVERSARIAL SCENARIOS AND MODEL

In this article, we assume wearable sensors acquiring biosignals and transmitting the raw signal to the smart phones or cloud engines for detection of health pathologies in real time. We assume stealthy adversaries, who would like to remain undetected, to maximize their impacts for an extended period.

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).

We assume unencrypted communication between the wearable sensor and the mobile phone or cloud engines. Almost two out of ten mobile applications for IoT devices use unencrypted communications to the cloud [4]. For communications in the local network, the number of unencrypted connections is even higher. Furthermore, we assume that the seizure detection machine-learning model is publicly available. We consider two man-in-the-middle attack scenarios, i.e., real-time alert-masking adversarial attack and data-manipulation adversarial attack, which are discussed in the following.

2.1. Real-Time Alert-Masking Adversarial Attack

Wearable and mobile-health technologies are often used for real-time monitoring of patients to raise an alert in the case of life-threating events. In the case of epilepsy, wearable technologies, e.g., the e-Glass sensor [11], monitor the brain activities of the patients in real time to inform family members, caregivers, and emergency units for rescue in case of seizures. If the adversary is able to minimally manipulate the signal that is sent between the wearable sensor and the mobile device for seizure detection, then it is possible to mask such seizures. As a result, the family members, caregivers, and emergency units will not be notified to rescue the patients during/after the seizures, which, in turn, may even jeopardize patient's life.

2.2. Data-Manipulation Adversarial Attack

The information collected by mobile-health and wearable technologies is often used by the medical experts to develop a better understanding of such health pathologies and, in turn, diagnosis, prognosis, and treatment. For instance, epileptologists administer drugs based on the frequency and duration of the seizures. Without proper security measures, the adversary is able to manipulate the biosignals acquired by wearable sensors, while its being transmitted to smart phones or cloud engines. The medical experts then will prescribe according to these manipulated biosignals, which may clearly have irreversible consequences.

3. MOTIVATIONAL EXAMPLE

Let us consider the real-time epileptic seizure detection problem using the e-Glass wearable sensor [11]. The raw EEG signals are acquired by the e-Glass sensor and transmitted to smart mobile phone. The seizure detection machine-learning algorithm runs on the mobile phone to notify the accredited relatives and emergency units for rescue in case of seizures. A man-in-the-middle attack scenario is shown Figure 1, where the adversary manipulates the EEG signals to mask the seizures. As a result, the seizures may go unnoticed, which could be life-threatening for the patients.

The original EEG signals and the corresponding adversarial signals for two leads T7F7 and T8F8, which are captured by the e-Glass sensor, are shown in Figure 2. 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 [15]. Notice that there are only slight differences between the original and the corresponding adversarial signals, e.g., the amplitude of the two large negative spikes just after 200 in T8F8. Nevertheless, the adversary is able to mask the seizure by minimal perturbation of the original signals. The amplification factor for each sample is also shown in Figure 2. We make two observation: first, the majority of the samples remain approximately the same and, second, all samples are amplified by a factor less than 20%.

Fig. 1. The overall scenario for the man-in-the-middle adversarial attack in mobile-health applications.

4. MINIMAL ADVERSARIAL PERTURBATION

In this section, we discuss the proposed adversarial attack, with minimal manipulation of the signal. In Section 4.1, we formulate the seizure detection problem as an optimization problem. In Section 4.2, we formulate a convex optimization problem to identify the minimum manipulation required for misclassification of an ictal (seizure) sample as inter-ictal (non-seizure).

4.1. Seizure Detection Classification

Let us consider the support vector machine (SVM) classification algorithm [16]. The original formulation for the SVM classification with soft-margin is as follows,

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

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

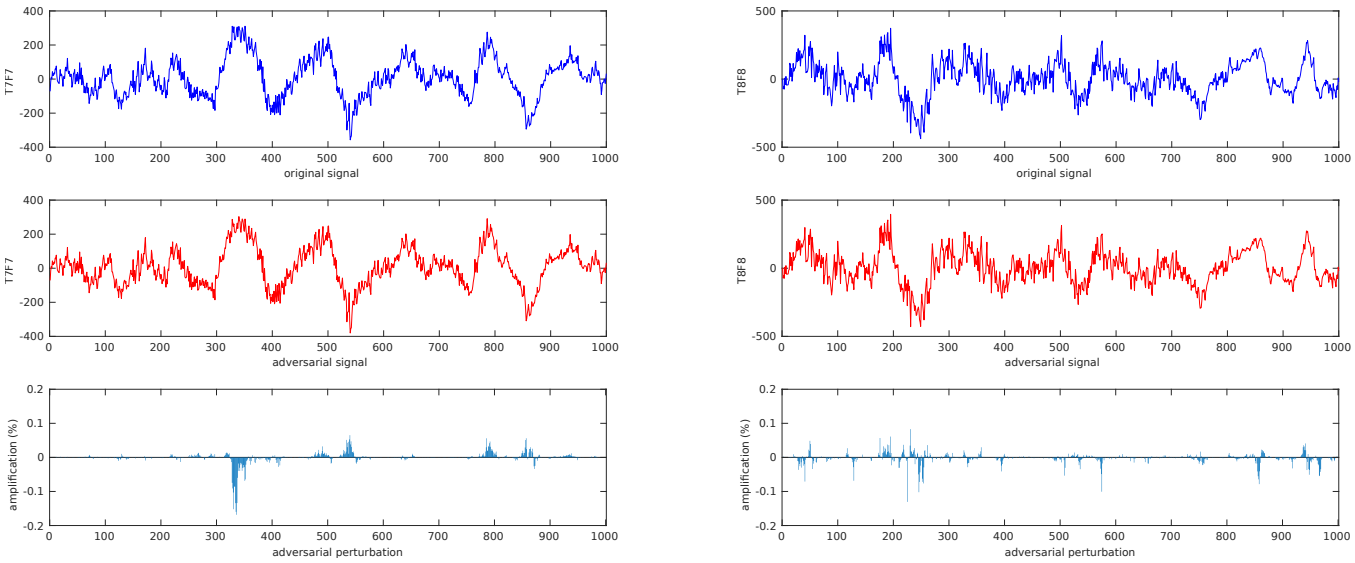

Fig. 2. The original signal (dark blue) and the adversarial signal (red) for two leads T7F7 and T8F8. The amplification factor per sample is shown also in light blue. Notice that there are slight differences between the original and adversarial signals, e.g., the amplitude of the two large negative spikes just after 200 in T8F8.

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

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

4.2. Minimal Adversarial Manipulation

Towards identifying the minimum adversarial perturbation for misclassification of seizure samples, we first consider the additive adversarial manipulation model, which is formulated as follows,

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

where $\boldsymbol{a}$ is the attack vector added with the data sample $\boldsymbol{x}_i$. Our objective is to minimize the adversarial perturbation, captured by $\|\boldsymbol{a}\|_2$. The constraint ensures that the data sample $\boldsymbol{x}_i$ is misclassified by adding the attack vector $\boldsymbol{a}$. Unfortunately, however, the above optimization problem is generally non-convex (because of the constraint) and not straightforward to solve efficiently.

Let us now consider the multiplicative adversarial perturbation model, which is formulated as follows,

$$\begin{aligned} \min_{\boldsymbol{a}} \quad & \|\boldsymbol{a} - \mathbf{1}\|_2^2 \\ \text{s.t.} \quad & y_i \cdot (\boldsymbol{w}^{\mathrm{T}}((\boldsymbol{x}_i \odot \boldsymbol{a}) \odot (\boldsymbol{x}_i \odot \boldsymbol{a})) + b) < 0. \end{aligned} \tag{4}$$

We reformulate the above optimization problem as follows,

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

which is a convex optimization problem and can be solved exactly to find the minimum adversarial effort, i.e., $\|\hat{\boldsymbol{a}} - \mathbf{1}\|_2$, required for the misclassification of 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 exact attack vector is captured by $\hat{\boldsymbol{a}}$.

5. 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.

5.1. Epilepsy Dataset

We consider the CHB-MIT database [17] 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. In total, these recordings include 198 seizures. These EEG signals are sampled at $F_s = 256$ Hz, with 16-bit resolution. We do not consider patients 6, 14, and 16 in this dataset since these patients generally suffer from short seizures [18]. We consider only two leads in the 10–20 EEG acquisition system [20], i.e., $T7F7$ and $T8F8$ in the

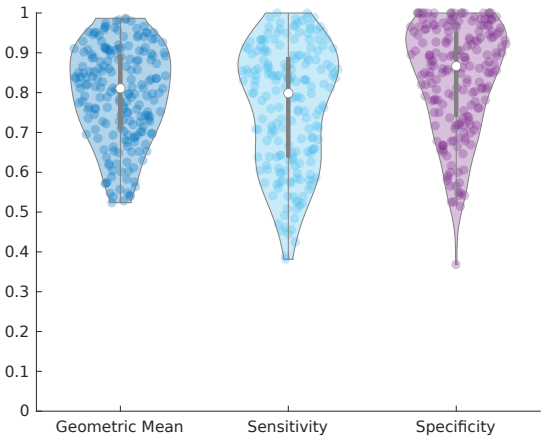

Fig. 3. The distribution of the geometric mean, the sensitivity, and the specificity for all subjects.

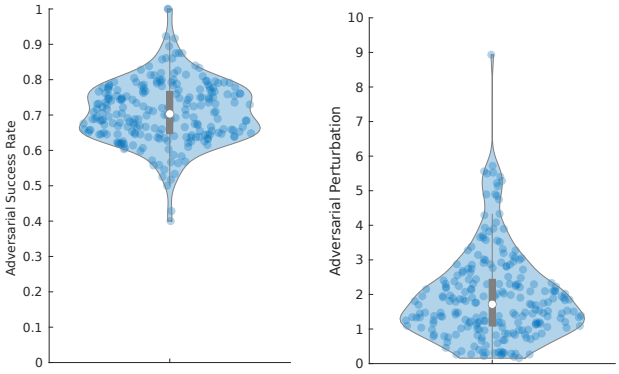

Fig. 4. The distribution of the adversarial success rate and minimum required perturbation.

e-Glass wearable system [11] which have been shown to be important for the detection of epileptic seizure.

5.2. Quality of Seizure Detection

In this section, we evaluate the seizure detection classification algorithm discussed in Section 4.1. We evaluate the performance of this algorithm based on the sensitivity (Sen), specificity ($Spec$), and geometric mean ($Gmean$) of the sensitivity and specificity, defined as follows,

$$Spec = \frac{TN}{FP + TN}, \tag{6}$$

$$Sen = \frac{TP}{TP + FN}, \tag{7}$$

$$Gmean = \sqrt{Spec \cdot Sen}, \tag{8}$$

where FP, TN, TP and FN capture the number of false-positive samples, true-negative samples, true-positive samples, and false-negative samples, respectively. We consider the geometric mean, since it is the only correct average of normalized measurements [19].

We split the entire dataset into 70% training and 30% test. We perform a 10-fold cross-validation and summarize the results. The geometric mean, sensitivity, and specificity are shown in Figure 3. The median value of the geometric mean among all subjects is 81.1%. Note that since we limit our analysis to only two leads, i.e., $T7F7$ and $T8F8$ in the e-Glass wearable system [11], the algorithm is not able to accurately detect all seizures, as it can be seen in Figure 3.

5.3. Minimal Adversarial Manipulation

In this section, we consider the epileptic seizure samples and evaluate the possibility of successful manipulation of these samples, such that seizure samples are misclassified as non-seizure. We evaluate the performance of our adversarial manipulation 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 4. The median success rate among all subjects is 70.3%. More importantly, for all patients, the proposed scheme is able to successfully perturb more than 40.0% of the seizure samples such that the classification algorithm misclassifies these samples as non-seizure.

The median required perturbation captures the effort required for misclassification of seizure samples as non-seizure. The results are shown in Figure 4. The median value of $\|\hat{\boldsymbol{a}} - \mathbf{1}\|_2$ among all subjects is 1.7, where vector $\hat{\boldsymbol{a}}$ is of dimension 2000. This demonstrates the possibility of such attacks based on minimal perturbation of epileptic brain activities, which may have irreversible consequences if not detected.

6. CONCLUSION

The security of Internet of things (IoT) and mobile-health technologies represents one of today's main challenges. Adversarial manipulation of sensitive health-related information, e.g., if used for prescribing medicine, may have irreversible consequences, involving patients' lives. In this article, we demonstrated the power of such adversarial attacks based on a real-world epileptic seizure detection problem. We formulated this problem as a convex optimization problem to identify the minimum effort required by stealthy adversaries to declare seizure samples as non-seizure, i.e., minimal adversarial perturbation to fool the classification algorithm.

7. REFERENCES

[1] Zhi-Kai Zhang, Michael Cheng Yi Cho, Chia-Wei Wang, Chia-Wei Hsu, Chong-Kuan Chen, and Shiuhpyng Shieh, "IoT security: ongoing challenges and research opportunities," in *2014 IEEE 7th international conference on service-oriented computing and applications*. IEEE, 2014, pp. 230–234.

[2] Minhaj Ahmad Khan and Khaled Salah, "IoT security: Review, blockchain solutions, and open challenges," *Future Generation Computer Systems*, vol. 82, pp. 395–411, 2018.

[3] Amir Aminifar, Petru Eles, and Zebo Peng, "Optimization of message encryption for real-time applications in embedded systems," *IEEE Transactions on Computers*, 2018.

[4] Mario Ballano Barcena and Candid Wueest, "Insecurity in the internet of things," *Security Response, Symantec*, 2015.

[5] 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.

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

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

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

[15] 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.

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

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

[18] 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.

[19] Philip J Fleming and John J Wallace, "How not to lie with statistics: the correct way to summarize benchmark results," *Communications of the ACM*, vol. 29, no. 3, pp. 218–221, 1986.

[20] 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.