Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Journal Track

Adversarial Framework with Certified Robustness for Time-Series Domain via Statistical Features (Extended Abstract)*

Taha Belkhouja[1] and **Janardhan Rao Doppa**[1]
[1]School of Electrical Engineering and Computer Science, Washington State University, USA
{taha.belkhouja, jana.doppa}@wsu.edu

Abstract

Time-series data arises in many real-world applications (e.g., mobile health) and deep neural networks (DNNs) have shown great success in solving them. Despite their success, little is known about their robustness to adversarial attacks. In this paper, we propose a novel adversarial framework referred to as **T***ime*-**S***eries* **A***ttacks via* **STAT***istical Features (TSA-STAT)*. To address the unique challenges of time-series domain, TSA-STAT employs constraints on statistical features of the time-series data to construct adversarial examples. Optimized polynomial transformations are used to create attacks that are more effective (in terms of successfully fooling DNNs) than those based on additive perturbations. We also provide certified bounds on the norm of the statistical features for constructing adversarial examples. Our experiments on diverse real-world benchmark datasets show the effectiveness of TSA-STAT in fooling DNNs for time-series domain and in improving their robustness.

1 Introduction

We are seeing a significant growth in the Internet of Things (IoT) and mobile applications which are based on predictive analytics over time-series data collected from various types of sensors [Belkhouja and Doppa, 2020]. Some important applications include smart home automation [Aminikhanghahi *et al.*, 2018], mobile health [Ignatov, 2018], smart grid management [Zheng *et al.*, 2017], and finance [Ozbayoglu *et al.*, 2020]. Deep neural networks (DNNs) have shown great success in learning accurate predictive models from time-series data [Wang *et al.*, 2017; Luo *et al.*, 2019; Luo *et al.*, 2018]. In spite of their success, very little is known about the adversarial robustness of DNNs for time-series domain. Most of the prior work on adversarial robustness for DNNs is focused on image domain [Kolter and Madry, 2018; Carlini and Wagner, 2017; Chen and Gu, 2020; Athalye *et al.*, 2018; Hosseini *et al.*, 2017; Xiao *et al.*, 2018; Laidlaw and Feizi, 2019; Chen *et al.*, 2020; Moosavi-Dezfooli *et al.*, 2017] and natural language domain [Wang *et al.*, 2019; Gao *et al.*, 2018; Samanta and Mehta, 2017]. Adversarial perturbations are constructed by bounding an l_p-norm and depend heavily on the input data space: they can be a small noise to individual pixels of an image or word substitutions in a sentence. Adversarial examples expose the brittleness of DNNs and motivate creation of training methods to improve model robustness.

Time-series domain poses unique challenges (e.g., sparse peaks, fast oscillations) that are not encountered in both image and natural language domains [Belkhouja *et al.*, 2022; Belkhouja *et al.*, 2023]. The standard approach of imposing an l_p-norm bound is not applicable as it does not capture the true similarity between time-series signals [Hussein *et al.*, 2022]. Consequently, l_p-norm constrained perturbations [Fawaz *et al.*, 2019; Kurakin *et al.*, 2016; Karim *et al.*, 2020; Siddiqui *et al.*, 2019] can potentially create adversarial examples which correspond to a different class label. There is no prior work on filtering methods in the signal processing literature to *automatically* identify such invalid adversarial candidates. Hence, adversarial examples from prior methods based on l_p-norm will confuse the learner when they are used to improve the robustness of DNNs via adversarial training [Tramer *et al.*, 2020; Tramèr *et al.*, 2018], i.e., augmenting the original training data with adversarial examples. In other words, the accuracy of DNNs can potentially degrade on real-world time-series data after adversarial training.

In this paper, we propose a novel framework referred to as **T***ime*-**S***eries* **A***ttacks via* **STAT***istical Features (TSA-STAT)* and provide certified bounds on robustness. TSA-STAT relies on three key ideas. First, we create adversarial examples by imposing *constraints on statistical features* of the clean time-series signal. This is inspired by the observation that time-series data are comprehensible using multiple statistical tools rather than the raw data [Ignatov, 2018; Christ *et al.*, 2016; Ge and Ge, 2016]. The statistical constraints allow us to create valid adversarial examples that are more similar to the original time-series signal when compared to l_p-norm constrained perturbations. Second, we employ *polynomial transformations* to create adversarial time-series examples. We theoretically prove that polynomial transformations expand the space of valid adversarial examples over traditional additive perturbations, i.e., identify blind spots of additive perturbations. Our experiments demonstrate that polynomial

*Originally published at the Journal of Artificial Intelligence Research (JAIR) [Belkhouja and Doppa, 2022]

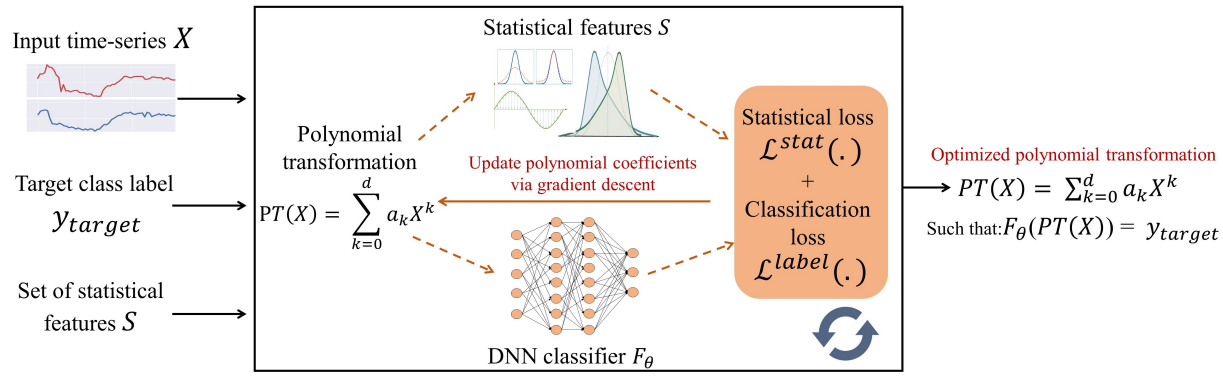

Figure 1: High-level overview of the TSA-STAT framework to create adversarial examples using optimized polynomial transformations. Given an input time-series signal X, a target label y_{target}, a DNN classifier F_θ, and a set of statistical features $\mathcal{S}$, TSA-STAT solves an optimization problem over two different losses to find the parameters of the polynomial transformation: 1) A statistical loss to ensure that original time-series signal X and the generated adversarial example X_{adv} are highly similar; and 2) A classification loss to make sure that the DNN classifier F_θ classifies X_{adv} with the target class label y_{target}.

transformation based attacks are more effective (in terms of successfully fooling time-series DNNs) than those based on additive perturbations. Third, to create attacks of different types, we solve an appropriate optimization problem to *identify the parameters of the polynomial transformation* via gradient descent. Figure 1 provides a conceptual illustration of the TSA-STAT algorithm. Certifiable robustness [Raghunathan *et al.*, 2018; Hein and Andriushchenko, 2017; Cohen *et al.*, 2019; Li *et al.*, 2019; Huang *et al.*, 2017] studies DNN classifiers whose prediction for any input X is verifiably constant within some neighborhood around X, e.g., l_p ball. We derive a certified bound for robustness of adversarial attacks using TSA-STAT. Our TSA-STAT framework provides certification guarantees that are applicable to DNNs for the time-series domain with different network structures.

The **key contributions** of this paper is the development, theoretical analysis, and experimental evaluation of the TSA-STAT framework [Belkhouja and Doppa, 2022]. Specific contributions include:

- Development of a principled approach to create targeted adversarial examples for the time-series domain using statistical constraints and polynomial transformations with theoretical analysis proving that polynomial transformations expand the space of valid adversarial examples over additive perturbations.
- Derivation of a certified bound for adversarial robustness of TSA-STAT that is applicable to any deep model for time-series domain.
- Comprehensive experimental evaluation of TSA-STAT on diverse real-world datasets and comparison with state-of-the-art baselines that demonstrate the practical benefits of extending the space of valid adversarial examples over those from prior l_p-norm based methods.

2 The TSA-STAT Framework

2.1 Key Elements

1) Statistical constraints. Time-series data is often analyzed using diverse statistical tools [Montgomery *et al.*, 2015]. Machine learning models have achieved good classification performance using statistical features of time-series data [Fulcher and Jones, 2014]. These prior studies motivate us to use statistical features of time-series data to develop adversarial algorithms. We propose a new definition to create adversarial examples for time-series signals. Let $\mathcal{S}^m(X) = \{S_1(X), S_2(X), \cdots, S_m(X)\}$ be the set of statistical features of a given time-series signal X (e.g., mean, standard deviation, kurtosis). We define an adversarial example X_{adv} derived from X as follows:

$$\begin{cases} \forall\ 1 \leq i \leq m,\ \|S_i(X_{adv}) - S_i(X)\|_\infty \leq \epsilon_i \\ \text{and}\ \ F_\theta(X) \neq F_\theta(X_{adv}) \end{cases} \tag{1}$$

where ϵ_i is the bound for the i^{th} statistical feature. Using this definition, we call to change the conventional l_p distance-based neighborhood-similarity to one based on statistical features for creating valid adversarial examples. We conjecture that this definition is better suited for adversarial examples in time-series domain and our experiments support this claim.

2) Polynomial transformation-based attacks. To explore a larger space of valid adversarial examples when compared to traditional additive perturbations, we propose polynomial transformation based attacks. The aim of this approach is to find a transformation over the input space that creates effective adversarial attacks. Hence, we define polynomial transformation $\mathcal{PT} : \mathbb{R}^{n\times T} \rightarrow \mathbb{R}^{n\times T}$ as follows: $X_{adv} = \mathcal{PT}(X) = \mathcal{PT}(X_{i,j})\ \ \forall (i,j) \in [n] \times [T]$ where $X \in \mathbb{R}^{n\times T}$ is the input time-series signal and X_{adv} is the corresponding adversarial example. The key idea is to create a threat model that does not require calling back the deep model for every new adversarial attack. Our goal is to preserve dependencies between features of the input space by having a transformation $\mathcal{PT}(\cdot)$ that depends on the input time-series X, unlike the standard additive perturbations. Inspired by power series [Drensky and Holtkamp, 2006], we approximate this transformation $\mathcal{PT}(\cdot)$ using a polynomial representation with a chosen degree d:

$$\mathcal{PT}(X) = \sum_{k=0}^{d} a_k\ X^k + \mathcal{O}(X^{d+1}), \tag{2}$$

where $a_k \in \mathbb{R}^{n\times T}$ denote the polynomial coefficients and $\mathcal{O}$ stands for Big O notation.

Theorem 1. *For a given input space $\mathbb{R}^{n\times T}$ and $d \geq 1$, polynomial transformations allow more candidate adversarial examples than additive perturbations in a constrained space. If $X \in \mathbb{R}^{n\times T}$ and $\mathcal{PT} : X \rightarrow \sum_{k=0}^{d} a_k\ X^k$, then $\forall X_{adv}$ s.t. $\|S_i(X_{adv}) - S_i(X)\|_\infty \leq \epsilon_i$:*

$$\left\{X_{adv} = \mathcal{PT}(X),\ \forall a_k\right\} \supsetneq \left\{X_{adv} = X + \delta,\ \forall \delta\right\}$$

, $S_i \in \mathcal{S}^m(X) \bigcup Identity$.

The above theorem states that polynomial transformations expand the space of valid adversarial examples and identify blind spots of additive perturbations. In other words, the theorem explains that *some* of the adversarial examples created using polynomial transformations are not possible using standard additive perturbations.

3) Optimization based adversarial attacks. To create powerful adversarial examples to fool the deep model $F_\theta(X)$, we need to find optimized coefficients a_k, $\forall$ k=0 to d, of the polynomial transformation $\mathcal{PT}(X)$. Our approach minimizes a *loss function* $\mathcal{L}$ using gradient descent that has two elements:

- **Classification loss.** To enforce an input signal X to be mis-classified to a target class y_t (different from true class label $y \in Y$), we employ the formulation of [Carlini and Wagner, 2017] to define a loss function:

$$\mathcal{L}^{label}(\mathcal{PT}, X) = \max\left[\max_{y\neq y_t}(\mathcal{Z}_y(\mathcal{PT}(X)) - \mathcal{Z}_{y_t}(\mathcal{PT}(X)),\ \rho\right]$$

  where $\rho < 0$. This loss function will ensure that the adversarial example will be moving towards the space where it will be classified by the DNN as class y_t with a confidence $|\rho|$ using the output of the pre-softmax layer $\{\mathcal{Z}_y\}_{y\in Y}$.
- **Statistical loss.** To preserve the similarity between an input and its adversarial example, we propose another loss function that controls the close proximity of statistical features in a given set $\mathcal{S}^m$. This loss function overcomes the impractical use of projection functions on the statistical feature space.

$$\mathcal{L}^{stat}(\mathcal{PT}, X, \mathcal{S}^m) \triangleq \sum_{S_i\in\mathcal{S}^m} \|S_i(\mathcal{PT}(X)) - S_i(X)\|_\infty$$

The final **combined loss** function $\mathcal{L}$ that we want to minimize to obtain coefficients a_k of the polynomial transformation $\mathcal{PT}(\cdot)$ is as follows:

$$\mathcal{L}(\mathcal{PT}, X, \mathcal{S}^m) = \beta_l \cdot \mathcal{L}^{label}(\mathcal{PT}, X) + \beta_s \cdot \mathcal{L}^{stat}(\mathcal{PT}, X, \mathcal{S}^m) \quad (3)$$

where β_l and β_s are hyper-parameters that can be used to change the trade-off between $\mathcal{L}^{label}$ and $\mathcal{L}^{stat}$. We note that experiments showed good results by simply using β_l=β_s=1.

2.2 Instantiations of TSA-STAT

Our goal is to create targeted adversarial attacks on a classifier F_θ. An adversarial example X_{adv} for a **single-instance** input X is defined as $X_{adv} = \mathcal{PT}_{y_t}(X) = \sum_{k=0}^{d} a_k\ X^k$ s.t.:

$$\begin{cases} \|S_i(X_{adv}) - S_i(X)\|_\infty \leq \epsilon_i\ \forall S_i \in \mathcal{S}^m \\ F_\theta(X_{adv}) = y_t \end{cases}$$

where y_t is the target class-label of the attack. This adversarial example is constructed by employing a gradient descent based optimizer to minimize the loss function in Equation 3 over $\{a_k\}_{0\leq k\leq d}$ defining $\mathcal{PT}$.

For **black-box attacks** where adversarial examples are created with no knowledge about the target deep model parameters θ, the attacker queries the target model to get the predicted label for any input time-series X. This allows the creation of a proxy deep model to mimic the behavior of the target model [Tramer *et al.*, 2020; Papernot *et al.*, 2017] and to be used to compute the adversarial examples. We can also employ TSA-STAT to create **universal perturbations**. A universal perturbation generates a single transformation that is applicable for any input $X \in \mathbb{R}^{n\times T}$. We introduce a targeted universal attack in this setting as:

$$X_{adv} = \mathcal{PT}_{y_t}(X) \ \text{ s.t. } \mathcal{PT}_{y_t}(F(X_{adv}) = y_t) > (1 - e_t) \quad (4)$$

where e_t represents the error probability of creating an adversarial example that F_θ would classify it with label $y \neq y_t$. Our proposed algorithm analyzes a given set of input time-series signals to find coefficients $\{a_k\}_{0\leq k\leq d}$ that would push the image of multiple inputs $\mathcal{PT}_{y_t}(X)$ to the decision boundary of a target class-label y_t defined by the classifier F_θ.

3 Certificates for Adversarial Robustness

We propose a novel certification approach for adversarial robustness of the TSA-STAT framework. Given a time-series input $X \in \mathbb{R}^{n\times T}$ and a classifier F_θ, our overall goal is to provide a certification bound δ on the $\|\cdot\|_\infty$-norm over the statistical features $\mathcal{S}^m(X)$ of the time-series signal X. This bound will guarantee the robustness of classifier F_θ in predicting $F_\theta(X_{adv}) = F_\theta(X)$ for any adversarial time-series X_{adv} such that $\sum_{S_i\in\mathcal{S}^m} \|S_i(X_{adv}) - S_i(X)\|_\infty \leq \delta$.

To derive the certification bound that is suitable for a time-series input $X \in \mathbb{R}^{n\times T}$ and a classifier F_θ, we employ two different noise distributions to generate two different noise samples that we denote $n_P \in \mathbb{R}^{n\times T}$ and $n_0 \in \mathbb{R}^{n\times T}$. $n_P \sim \mathcal{N}(\mu_P, \cdot)$ is generated to mimic the perturbation characterizing the robustness of classifier F_θ for predicting the same label for perturbed time-series. $n_0 \sim \mathcal{N}(0, \cdot)$ is generated as an arbitrary noise needed for the computation of the certification bound. If both perturbations result in the same classifier prediction, we compute the tolerable perturbation's upper bound $\delta = \max \|\mu_P\|_\infty$ as shown in Theorem 2.

Theorem 2. *Let $X \in \mathbb{R}^{n\times T}$ be an input time-series signal. Let $n_P \sim \mathcal{N}(\mu_P \in \mathbb{R}^n, \sum)$ and $n_0 \sim \mathcal{N}(0, \sum)$. Given a classifier $F_\theta : \mathbb{R}^{n\times T} \rightarrow Y$ that produces a probability distribution $(p_1, \cdots, p_k)$ over k labels for $F_\theta(X + n_P)$ and another probability distribution $(p_1^0, \cdots, p_k^0)$ for $F_\theta(X + n_0)$. To guarantee that $arg\max_{p_i}\ p_i = arg\max_{p_i^0}\ p_i^0$, the following condition must be satisfied:*

$$\|\mu_P\|_\infty^2 \leq \max_{\alpha\neq 1} \frac{2}{\alpha \cdot \sum^{(S)}} \cdot \left(-ln(1 - p_{(1)} - p_{(2)} + 2(\frac{1}{2}(p_{(1)}^{1-\alpha} + p_{(2)}^{1-\alpha}))^{\frac{1}{1-\alpha}})\right) \quad (5)$$

where $\sum^{(S)}$ *is the sum of all elements of* $\sum$.

Lemma 1. *If a certified bound* δ *is generated for the mean of an input signal* X *and a classifier* F_θ*, then certified bounds for other statistical features can be derived consequently.*

4 Experiments and Results

We briefly summarize the main results from [Belkhouja and Doppa, 2022].

4.1 Experimental Setup

To evaluate the proposed TSA-STAT framework, [Belkhouja and Doppa, 2022] employed diverse uni-variate and multi-variate time series benchmark datasets [Bagnall *et al.*, 2020; Dua and Graff, 2017; Kwapisz *et al.*, 2011]. For this extended abstract, we only show the performance over two datasets (WD and SC) noting that similar patterns are found on other datasets. Three different 1D-CNN architectures are used to create three deep models as target DNN classifiers: WB for white-box setting, and $BB1$ and $BB2$ for the black-box setting respectively. We employ models that are trained with clean data and others using augmented data from baselines attacks: Fast Gradient Sign method (FGS) [Fawaz *et al.*, 2019], Carlini & Wagner (CW) [Carlini and Wagner, 2017], and Projected Gradient Descent (PGD) [Madry *et al.*, 2017].

4.2 Results and Discussion

Spatial distribution of TSA-STAT outputs. We employ a t-Distributed Stochastic Neighbor Embedding (t-SNE) [Maaten and Hinton, 2008] technique to visualize the adversarial examples generated by TSA-STAT and PGD, an l_p-norm based attack. Figure 2 illustrates a representative example of the spatial distribution between same-class data of HAPT dataset, and their respective adversarial examples using TSA-STAT and PGD. We can see that TSA-STAT succeeds in preserving the similarity between the original and adversarial example pairs, and in most cases, better than PGD.

Effectiveness of adversarial examples from TSA-STAT. We show the effectiveness of generated adversarial examples to fool different deep models for time-series domain. We evaluate TSA-STAT using $\alpha_{Eff} \in [0, 1]$ (higher means better attacks) that measures the capability of targeted adversarial examples to fool a given classifier. Figure 3 shows the results for instance-specific and universal attacks under white-box and black-box settings on different deep models. Figure 4 shows the results of different deep models after adversarial training based on different methods including TSA-STAT.

Figure 2: tSNE visualization showing the distribution of natural and adversarial examples from TSA-STAT and PGD on HAPT dataset.

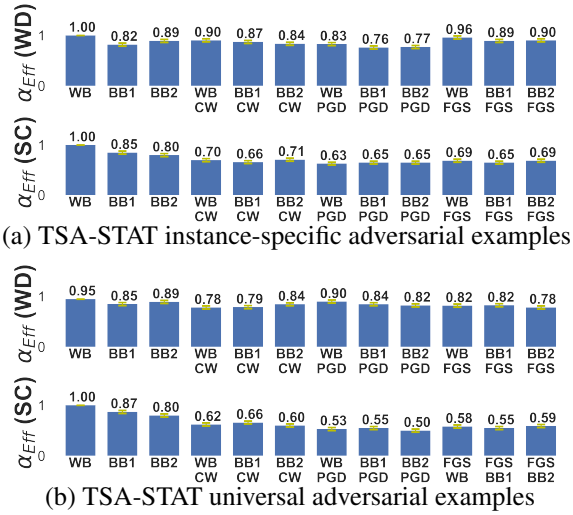

Figure 3: Results for TSA-STAT attack on different models.

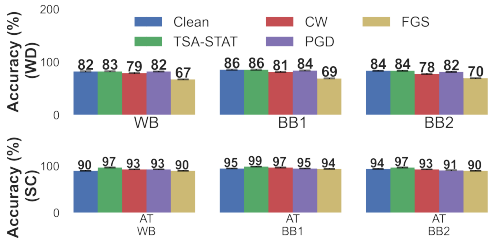

Figure 4: Results for adversarial training using adversarial examples from different methods for different deep models.

These experiments summarize the effectiveness of adversarial examples from TSA-STAT and show that statistical features are well-justified for time-series data. If the standard l_p-norm-based methods from the image domain were to be very effective for the time-series domain, 1) TSA-STAT's attacks would not be able to fool the models using baselines as a defense method, and 2) TSA-STAT's adversarial training would not outperform the baselines on clean accuracy.

4.3 Summary of Results

- The similarity measure based on statistical features is more effective for time-series data when compared to the standard l_p-norm based algorithms.
- Adversarial attacks created by TSA-STAT are very effective in fooling DNNs for time-series classification tasks and evading adversarial training based DNNs using adversarial examples created by prior methods.
- TSA-STAT provides better true-label guarantees (examples belonging to the semantic space of true label) than prior methods and can compute certification guarantees for robust classification as stated in **Theorem 2**.

Acknowledgements

This research is supported in part by the AgAID AI Institute, supported by the NSF and USDA-NIFA award #2021-67021-35344.

References

[Aminikhanghahi *et al.*, 2018] Samaneh Aminikhanghahi, Tinghui Wang, and Diane J Cook. Real-time change point detection with application to smart home time series data. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):1010–1023, 2018.

[Athalye *et al.*, 2018] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning (ICML*, 2018.

[Bagnall *et al.*, 2020] Anthony Bagnall, Jason Lines, William Vickers, and Eamonn Keogh. The UEA & UCR time series classification rep. www.timeseriesclassification.com, 2020. Accessed: 2020-01-01.

[Belkhouja and Doppa, 2020] Taha Belkhouja and Janardhan Rao Doppa. Analyzing deep learning for time-series data through adversarial lens in mobile and iot applications. *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems (TCAD)*, 39(11):3190–3201, 2020.

[Belkhouja and Doppa, 2022] Taha Belkhouja and Janardhan Rao Doppa. Adversarial framework with certified robustness for time-series domain via statistical features. *Journal of Artificial Intelligence Research*, 73:1435–1471, 2022.

[Belkhouja *et al.*, 2022] Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Training robust deep models for time-series domain: Novel algorithms and theoretical analysis. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[Belkhouja *et al.*, 2023] Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Dynamic time warping based adversarial framework for time-series domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

[Chen and Gu, 2020] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[Chen *et al.*, 2020] Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A frank-wolfe framework for efficient and effective adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[Christ *et al.*, 2016] Maximilian Christ, Andreas W Kempa-Liehr, and Michael Feindt. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717*, 2016.

[Cohen *et al.*, 2019] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

[Drensky and Holtkamp, 2006] Vesselin Drensky and Ralf Holtkamp. Constants of formal derivatives of non-associative algebras, taylor expansions and applications. *Rendiconti del Circolo Matematico di Palermo*, 55(3):369–384, 2006.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2017. Accessed: 2020-01-01.

[Fawaz *et al.*, 2019] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Adversarial attacks on deep neural networks for time series classification. *International Joint Conference on Neural Networks (IJCNN)*, 2019.

[Fulcher and Jones, 2014] Ben D Fulcher and Nick S Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 2014.

[Gao *et al.*, 2018] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *IEEE Security and Privacy Workshops (SPW)*, pages 50–56, 2018.

[Ge and Ge, 2016] Li Ge and Li-Juan Ge. Feature extraction of time series classification based on multi-method integration. *Optik*, 127(23):11070–11074, 2016.

[Hein and Andriushchenko, 2017] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2266–2276, 2017.

[Hosseini *et al.*, 2017] Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. On the limitation of convolutional neural networks in recognizing negative images. In *16th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017.

[Huang *et al.*, 2017] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In *International conference on computer aided verification*, pages 3–29. Springer, 2017.

[Hussein *et al.*, 2022] Dina Hussein, Taha Belkhouja, Ganapati Bhat, and Janardhan Rao Doppa. Reliable machine learning for wearable activity monitoring: Novel algorithms and theoretical guarantees. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 33:1–33:9. ACM, 2022.

[Ignatov, 2018] Andrey Ignatov. Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing*, 62:915–922, 2018.

[Karim *et al.*, 2020] Fazle Karim, Somshubra Majumdar, and Houshang Darabi. Adversarial attacks on time series. *IEEE Transactions on pattern analysis and machine intelligence*, 2020.

[Kolter and Madry, 2018] Zico Kolter and Aleksander Madry. Tutorial adversarial robustness: Theory and practice. *NeurIPS*, 2018.

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

[Kwapisz *et al.*, 2011] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[Laidlaw and Feizi, 2019] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *Advances in Neural Information Processing Systems (Neur'IPS)*, pages 10408–10418, 2019.

[Li *et al.*, 2019] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pages 9459–9469, 2019.

[Luo *et al.*, 2018] Yonghong Luo, Xiangrui Cai, Ying ZHANG, Jun Xu, and Yuan xiaojie. Multivariate time series imputation with generative adversarial networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2018.

[Luo *et al.*, 2019] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E²gan: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Montgomery *et al.*, 2015] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[Moosavi-Dezfooli *et al.*, 2017] 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 (CVPR)*, 2017.

[Ozbayoglu *et al.*, 2020] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *arXiv preprint arXiv:2002.05786*, 2020.

[Papernot *et al.*, 2017] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of Asia Conference on Computer and Communications Security (ASIACCS)*. ACM, 2017.

[Raghunathan *et al.*, 2018] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

[Samanta and Mehta, 2017] Suranjana Samanta and Sameep Mehta. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*, 2017.

[Siddiqui *et al.*, 2019] Shoaib Ahmed Siddiqui, Dominique Mercier, Mohsin Munir, Andreas Dengel, and Sheraz Ahmed. TSViz: Demystification of deep learning models for time-series analysis. *IEEE Access*, 7:67027–67040, 2019.

[Tramèr *et al.*, 2018] Florian Tramèr, Alex Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations (ICLR)*, 2018.

[Tramer *et al.*, 2020] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

[Wang *et al.*, 2017] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.

[Wang *et al.*, 2019] William Yang Wang, Sameer Singh, and Jiwei Li. Deep adversarial learning for nlp. In *Proceedings of the Conference of the NAACL: Tutorials*, pages 1–5, 2019.

[Xiao *et al.*, 2018] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

[Zheng *et al.*, 2017] Zibin Zheng, Yatao Yang, Xiangdong Niu, Hong-Ning Dai, and Yuren Zhou. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics*, 2017.