Adversarial Attacks Against Acoustic Monitoring of Industrial Machines

Stavros Ntalampiras

***Abstract*—The recent rise of adversarial machine learning exposed the serious vulnerabilities existing in current frameworks depending on the smooth operation of such automated solutions. This article focuses on the critical field of monitoring the health of industrial machines based on the respective acoustic emissions. After building an audio-based monitoring solution using log-Mel spectrograms and convolutional neural networks, we systematically evaluate the applicability of four types of adversarial attacks: 1) fast gradient sign; 2) projected gradient descent; 3) Jacobian saliency map; and 4) Carlini and Wagner ℓ_∞. Seeing the problem from the attacker perspective, we designed two different attack types, aiming at inducing either false positives or false negatives. We define three figures of merit specifically designed to assess the performance of each attack type from diverse points of view. The experimental setup relies on a publicly available data set including acoustic emissions representing four industrial machines, i.e., *fan*, *pump*, *slider rail*, and *valve*.**

***Index Terms*—Acoustic monitoring, adversarial machine learning, audio signal processing, convolutional neural network, industrial machine health.**

I. Introduction

DESPITE the ever-increasing deployment of deep learning (DL)-based solutions on commercial products, the vulnerabilities of such technologies have not been thoroughly investigated. In fact, the amount of research revealing the importance of existing weaknesses is largely disproportionate with respect to research designing such solutions. That said, one of most relevant weaknesses is the susceptibility of DL modeling approaches to adversarial attacks [1]–[3], i.e., processes manipulating the testing data and create adversarial examples in order to drive the model toward a desired prediction [4]. Such purposefully generated samples constitute significant threats and question the reliability of DL-based algorithms. The problem becomes quite alarming when they address critical IoT applications, including healthcare [5], critical infrastructures [6], [7], etc. Furthermore, it is boosted by the poor interpretability of model predictions, which could potentially help in detecting such malicious threats [8]. As such, it is not an exaggeration to suggest that adversarial attacks comprise one of the principal obstacles toward wider adoption of DL-based solutions.

Manuscript received 29 August 2021; revised 7 March 2022 and 16 June 2022; accepted 26 July 2022. Date of publication 28 July 2022; date of current version 6 February 2023.

The author is with the Department of Computer Science, University of Milan, 20133 Milan, Italy (e-mail: stavros.ntalampiras@unimi.it).

Digital Object Identifier 10.1109/JIOT.2022.3194703

The aim of an attacker adopting the above described line of thought is to generate adversarial examples able to modify model's prediction by carefully altering the content of the input as little as possible [9]. Such attacks can be organised into two mail categories, i.e., *targeted* and *untargeted* [10]. The first class imposes constrains on the prediction that the attacker wishes to achieve, while the second merely aims at altering the model's prediction without requiring a specific predicted class. In addition, attacks can be distinguished by the knowledge regarding model characteristics, which are necessary to design the adversarial examples. As such, attacks assuming complete knowledge of the model (architecture, layers, weights, etc.) belong to the *white-box* category. On the contrary, when no such knowledge is available, the attack belongs to the *black-box* class. In such cases, a *surrogate* model is built to generate the attack, while considering transferability of the adversarial input [11]. Last but not least, in case partial information about the model is available, the attack can be categorized as a *gray-box* one.

This article is focused on adversarial attacks against DL models designed to acoustically monitor the health of industrial machines. Unfortunately, research regarding attacks on audio analysis solutions is still in its infancy, especially when the emphasis is placed on nonspeech audio [12]. There are sporadic attempts which either assume knowledge of the constructed model [13] or not [14]–[16]; however, these are designed to address very specific scenarios, while the generic applicability of the most prominent adversarial attacks existing in the literature on generalized audio classification tasks remains, at large, unknown.

On the other hand, there are solutions addressing audio-based diagnosis of machine health including drill bits [17], bearings [18], [19], wind turbine gearboxes [20], petrochemical units [21], etc. There, processing includes the extraction of handcrafted features, which are then classified by discriminative and nondiscriminative pattern recognition techniques [22]. A recent attempt stands out, addressing the unavailability of a public data set for such a research line, i.e., [23], where the authors describe an audio data set for malfunctioning industrial machine investigation and inspection, named MIMII. The specific data set is employed in this work in order to facilitate the reproducibility of the obtained results.

This article evaluates thoroughly the applicability of the most pertinent adversarial attacks against acoustic monitoring of industrial machines. To the best of our knowledge, this is the first time that such vulnerabilities are revealed in the specific application scenario characterized by crucial

relevance involving consequences on human lives, property loss/damage, production line problems, etc. Focusing on real-world conditions, we assume an audiostream composed of events representing both normal and faulty machine states coming from four machines, namely, *pump*, *valve*, *fan*, and *slider*. Following a malicious attacker line of thought, the considered attacks are targeted and aim at mispredicting sounds indicating normal operation as abnormal and *vice versa*. At the same time, it would be unrealistic to assume that the attacker has knowledge regarding the model, training data, etc., thus we consider the following black-box attack types: 1) fast gradient sign (FGS); 2) projected gradient descent (PGD); 3) Jacobian saliency map (JSM); and 4) Carlini and Wagner ℓ_∞ (C&W). We present extensive experiments to systematically assess the success of such attacks in the present scenario, while quantifying the applied perturbations and confidence classification levels. More precisely, we analyze the results on a per-machine basis and reveal the most widely applicable attack type.

The contributions of this work are the following.

1) For the first time, this work highlights relevant security issues for audio-based monitoring systems.
2) It systematically examines the applicability of adversarial attacks on state-of-the-art systems monitoring the health of industrial machines.
3) It considers a wide range of adversarial attacks and identifies of the most successful ones from the attacker point of view with respect to the present scenario.
4) It provides insights on how the attack mechanisms operate so that effective defence strategies can be derived.

The remainder of this article is organized as follows. Section II suitably formulates the present problem and Section III describes the audio classification system, including the feature extraction process as well as the DL model. Section IV explains the adversarial attacks against the previously mentioned system and Section V details the experimental set up and results assessing the efficiency of such attacks from multiple perspectives. Finally, Section VI concludes this work and identifies fruitful future research directions.

II. Problem Formulation

We denote the industrial machine health diagnosis task as $\mathcal{T}$ along with the a-priori known set of classes $\mathcal{C} = \{A, N\}$, where A represents abnormal states and N represents normal. Such health states are associated with characteristic sound emissions [23]. Without loss of generality, we assume inputs x of spectrogram form, i.e., $x \in \mathbb{R}^{d_1,d_2,d_3}$, where d_1 is the width, d_2 is the height, and d_3 is the number of color channels of the respective image. The ultimate goal of the present monitoring system is to build a model $\mathcal{M}$ identifying novel sounds while minimizing false positive and negative rates.

Unfortunately, such an automated system could be susceptible to malicious attacks applying an ideally imperceptible perturbation $\psi(x, y)$, where x is the input to be processed and y is the target class. Working against the classification system, an attacker may have a twofold goal as follows.

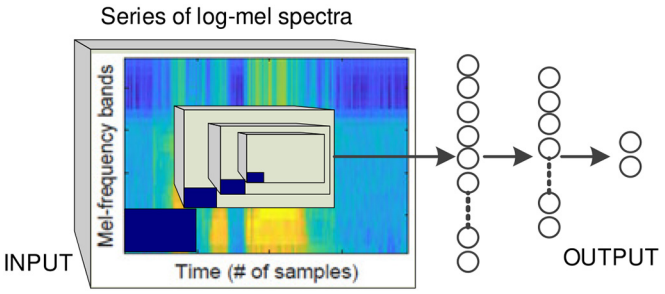

Fig. 1. CNN architecture for acoustic diagnosis of industrial machine health.

1) "Hiding" the existence of abnormal health states in A, i.e., increasing the false negative rate.
2) Force $\mathcal{M}$ to falsely classify a normal health state as an abnormal one, thus raising a false alarm.

It is assumed that the attacker is aware of the acoustic environment where the industrial machines operate, thus able to emit audio which serves his/her purposes. This work thoroughly assesses the efficiency of suitable perturbation functions ψ in the industrial machine health monitoring domain based on a state-of-the-art audio classification method.

III. Audio Classification System

Machine health diagnosis is carried out via an audio classification system based on a pipeline composed of two main stages: 1) extraction of log-Mel spectrograms and 2) distribution modeling by a convolutional neural network. Such a classification mechanism is well adopted in the related literature [24], and could potentially comprise the target of malicious attacks. The following sections explain briefly the two main stages, along with a suitable data set and the obtained results.

A. Mel-Spectrogram

We employed 40 equal-width log energies with an overlap based on the Mel filterbank. The standard extraction method is followed based on short time Fourier transform. Mel-spectrograms have been proven to emphasize components playing an important role to human perception [25]. As such, the present framework does not rely of domain expert knowledge and handcrafted features but a completely standardized feature extraction process.

B. Convolutional Neural Network

The present CNN is composed of three convolutional layers using the standard ReLU activation function. These are followed by max-pooling layers (see Fig. 1) with a total number of parameters equal to 2 197 634. The network further includes a flattening, a dropout, and three densely connected layers, where the regression process is carried out. In order to avoid overfitting, a dropout layer was considered randomly removing 50% of the present hidden units. The specific process scales down the number of parameters to estimate during learning, while discarding insignificant relationships.

The kernel size is 3×3 with stride being equal to 2, while the number

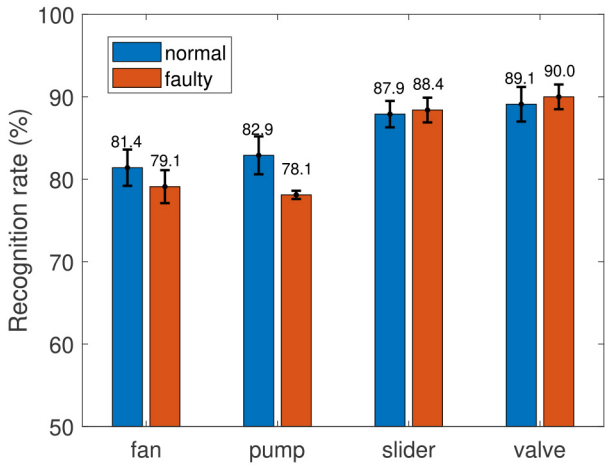

Fig. 2. Performance achieved by the considered modeling process with respect to data characterizing the available industrial machinery.

of output nodes represents the considered classes and is equal to 3 as shown in Fig. 1. It should be mentioned that the topology was optimized in terms of hyperparameters based on grid search including early stopping.

C. Data Set and Classification Results

Interestingly, a recent effort attempted to create a standardized data set facilitating the task of industrial machine health monitoring [23]. The data set resembles real-life scenarios and contains normal sounds recorded for different types of industrial machines (i.e., valves, pumps, fans, and slide rails), as well as anomalous sounds representing the faulty states of such machinery (e.g., contamination, leakage, rotating unbalance, and rail damage). There are 26 092 sound segments associated to normal conditions and 6065 sound segments of anomalous ones. These are mixed with the background noise recorded in multiple real factories in order to simulate real-world environments. The audio was captured with a quantization of 16 bit and sampling frequency 16 kHz. More information regarding the recording protocol (equipment, environment, etc.) along with guidelines to obtain it is available in [23].

We trained and optimized four different CNNs following the specifications of each industrial machine, i.e., valve, pump, fan, and slide rail. The achieved recognition rates (mean and standard deviation) following a ten-fold cross validation experimental protocol are shown in Fig. 2. In general, we observe that the considered audio classification pipeline, which represents well the typical line of thought nowadays, demonstrates quite promising classification performance ranging from 80% to over 90% for several machine types. This confirms the efficiency of the designed classification solution for the task-at-hand.

IV. Considered Adversarial Attacks

This section outlines the adversarial attacks fitting the requirements of the present application scenario as described in Section II. We evaluate the applicability of each attack against the audio-based machine health monitoring system.

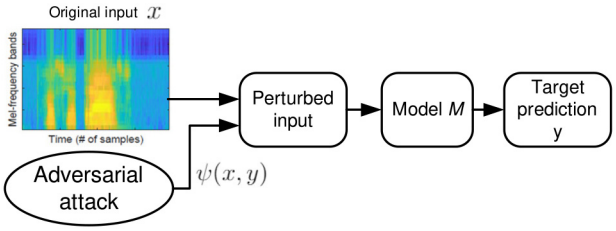

Fig. 3. Operation of the considered adversarial attacks.

As such, we carefully chose every attack available in the literature belonging to the targeted category (see Section I) without necessitating knowledge regarding the classification model, parameters, training data, etc. A block diagram demonstrating the way adversarial attacks operate is illustrated in Fig. 3.

1) FGS [26] is capable of designing both targeted and untargeted attacks as it tries to control the ℓ_1, ℓ_2, or ℓ_∞ norm of the alteration injected by the adversary. In the present application scenarios, $\psi(x, y) = -\epsilon \times \text{sign}(\nabla_x(\mathcal{L}(x, y)))$, where $\epsilon > 0$ denotes the attack strength and y denotes the target class as defined in Section II. The aim of this attack is to apply changes on x so that the classifier's loss is minimized when its prediction is y. Several values of $\epsilon \in [0.1, 0.5]$ are considered and studied alongside the perturbations incurred by the log-Mel spectrograms.
2) PGD [27] is an iterative version of FGS since as the attack is applied in an iterative way. As such, similar to FGS, attack strength is again ϵ. Moreover, the attack is guided by an additional parameter denoted as ϵ_{step} determining the step size of each iteration. As each iteration is carried out, the resulting example is projected onto the ϵ-norm sphere, the center of which consists in the original input x.
3) JSM [28], different from the previous attacks types, tries to control the ℓ_0 norm of the alteration injected by the adversary. JSM changes a predetermined amount of x's coefficients when generating ψ. The specific amount is upper bounded by a limit δ. This process is executed in an iterative way until either: a) δ is reached or b) the targeted prediction by $\mathcal{M}$ is achieved.
4) C&W [29] is the specific attack type that concentrates on minimizing the ℓ_∞ norm of adversarial examples. To this end, it searches for the optimum tradeoff between achieving the target prediction, while keeping the perturbation $\psi(x, y)$ as small as possible. Its aim is to generate a perturbed input $\psi(x, y)$ with $\ell(\psi) = 0$ (ℓ being the same with the ℓ_2 version). In addition, the condition $||x - \psi||_\infty \leq \epsilon$, with $\epsilon > 0$ has to be satisfied. In essence, ϵ corresponds to the amount of allowed changes to inject.

Normal and attacked log-mel spectrograms with respect to every attack and machine type are demonstrated in Fig. 4. We observe that most differences are imperceptible when they are examined by the naked eye. A general observation is that the attacks tend to perturb higher frequency regions more significantly than the lower ones. Moreover, the attacks do not

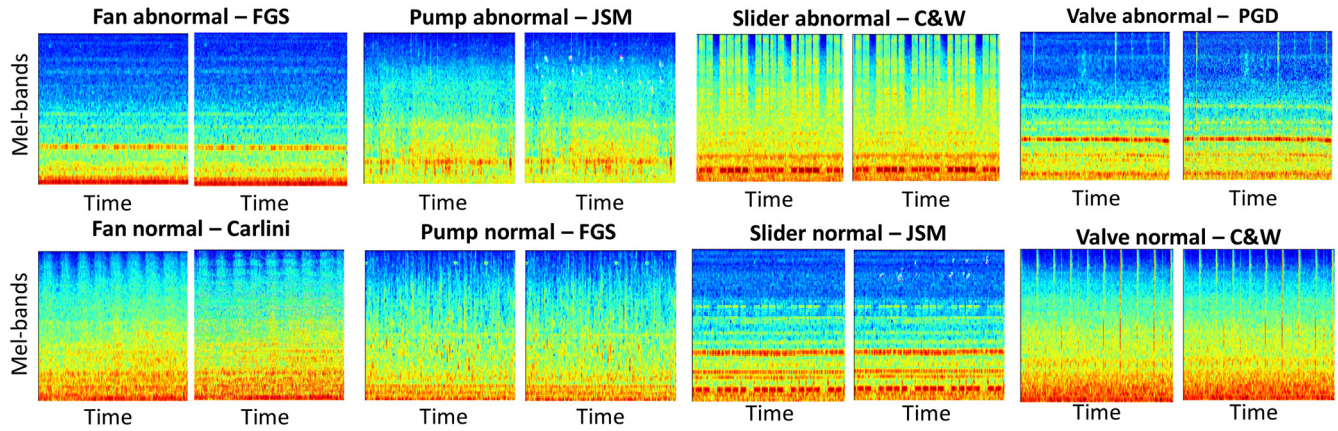

Fig. 4. Pairs of original (left) and attacked (right) log-Mel spectrograms with attack strength $\epsilon = 0.2$. Top row: diverse types of attack applied on abnormal samples. Bottom row: diverse types of attack applied on normal samples.

exhibit a long duration but prefer to alter the energy content of relatively small regions.

V. Experiments

This section describes: 1) the parameterization of the included modules; 2) the experimental protocol and figures of merit; 3) the obtained results for attack scenarios *AAS1* and *AAS2*; and 4) an in-depth analysis of the most successful adversarial attacks.

A. Features and Model Parameterization

Following MPEG-7 standard recommendations [30], feature extraction frame is 30 ms with 20-ms overlap. Hamming windowing is applied, while the FFT size was 1024. In addition, standard normalization techniques, i.e., mean removal and variance scaling, were used. Attack strength ϵ was taken from the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ while we employed the implementation provided in [10]. Finally, the CNN learning process with respect to each machine type is bounded by 1000 epochs at a learning rate of 0.0001.

B. Experimental Protocol and Results

Following the motivation described in Sections I and II, we focus on two attack application scenarios.

1) *AAS1:* Targeted attacks forcing $\mathcal{M}$ to falsely classify an event as abnormal even though it is normal, i.e., increasing false positive rate.
2) *AAS2:* Targeted attacks aiming at "hiding" the existence of abnormal situations in $\mathcal{C}^A$, i.e., increasing the false negative rate.

We followed the tenfold cross-validation division as in Section III where the testing set of each fold was perturbed by the attack types described is Section IV. Thus, all samples available in the data set were employed for every attack and the results are averaged. The performance of each attack type is measured by means of three figures of merit seeing the problem from diverse perspectives.

1) Success rate s_r, which counts the times an attack was successful, i.e., the target class was predicted by the model (percentage).
2) Perturbation p, which is the absolute difference between the original x and the adversarial sample with alteration ψ.
3) Classification confidence $\alpha = p(y|\mathcal{M})/\text{sum}(p(*|\mathcal{M}))$, which is the probability of the predicted class divided by the sum of all probabilities.

C. Attack Application Scenario 1

The results obtained with respect to the application scenario *AAS1* are demonstrated in Fig. 5. The results are analyzed per industrial machine type as it is reasonable to assume that the attacker would be in a position to know the identity of her/his target machinery.

1) Fan: In the results of the specific machine, we observe that the attack with the highest s_r is JSM for every considered ϵ. It is followed by PGD, FGS, and *C&W*. As expected, s_r increases with ϵ. At the same time, JSM is associated with the lowest amount of perturbations, which are significantly lower than the rest of the attacks. More specifically, the attack with the largest perturbations is FGS followed by PGD and *C&W*. Even though p's depict a significant increase with ϵ, s_r does not increase in an analogous way. In other words, larger perturbations do not result to proportionally successful attacks or confidence levels. On top of that, as shown in the subfigure related to the confidence level, JSM presents the lowest values with respect to every ϵ value, while the attack the highest values is PGD, followed by FGS and CW. Given that the confidence level can be conveniently used as the basis of a defence mechanism, PGD outperforms the remaining types of the attack.

2) Pump: Focusing on attacks targeting sounds indicative of normal pump states, we observe a similar pattern with the results concerning data characterizing fan machines. In particular, JSM presents the highest s_r, while PGD and FGS demonstrate similar performance. Unfortunately, CW is associated with low s_r values across every different ϵ value. However, it presents the smallest amount of perturbations when $\epsilon = 0.1$. Otherwise, JSM alters the input with the smallest amount of perturbations, while FGS and PGD modify the inputs the most. Even if p's increase substantially,

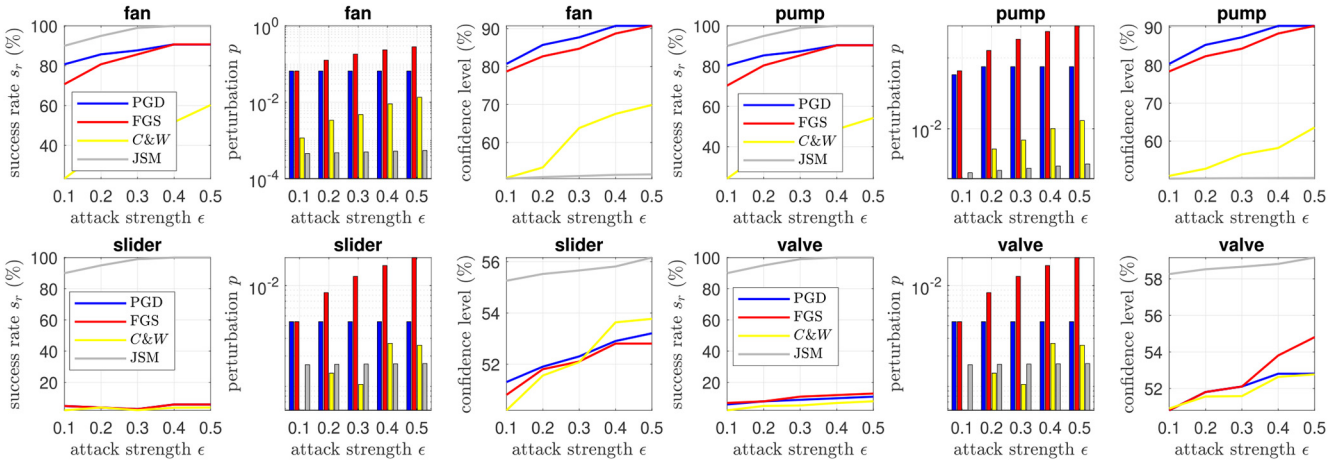

Fig. 5. *AAS*1 scenario perturbs sounds indicative of normal machine health targeting to have them misclassified as abnormal, i.e., inducing false alarms in every type of machine. Order of figures per machine: (a) success rate (%) versus attack strength ϵ, (b) normalized perturbation versus attack strength ϵ, and (c) confidence level α (%) for attacked samples versus attack strength ϵ.

s_r's present a moderate increase, which is true for the confidence levels too. In general, we observe that all figures of merit, i.e., s_r, p, and α, increase with ϵ. At the same time, JSM has the lowest confidence levels, while PGD and FGS have the highest. At the same time, CW offers mid-level α values. Similar to the fan case, PGD is the best performing attack.

3) Slider: With regards to slider, most attack types, i.e., PGD, FGS, and CW, fail to provide satisfactory s_r values, while JSM attacks are 100% successful for $\epsilon \geq 0.3$. Indeed, s_r remains practically constant with ϵ for PGD, FGS, and CW, which shows that the generated attacks are not suitable to mislead $\mathcal{M}$ to an erroneous prediction. At the same time, JSM presents relatively low perturbations, surpassed only by CW for $\epsilon \leq 0.2$. PGD and FGS are associated with higher perturbation values. There does not seem to exist a significant correlation between p and s_r, i.e., the amount of perturbations introduced to the signal do not necessarily render the attack successful nor alter α significantly. As regards to confidence levels, we observe the superiority of JSM, albeit the low values offered by all attack types with the highest being 56% for JSM and $\epsilon = 0.5$. Here, CW shows slightly higher α values than PGD and FGS. Overall, we see that α increases with ϵ for every attack type. Hence, we argue that JSM is the best-performing adversarial attack in the slider machinery case during AAS1.

4) Valve: For the specific machine type, the JSM attack achieved the highest s_r, which is equal to 100% for $\epsilon \geq 0.3$, while the rest attack types demonstrated much lower s_r's (<20%) for every considered ϵ value. s_r achieved by PGD, FGS, and CW do not alter with ϵ demonstrating their inability to deceive $\mathcal{M}$. Moreover, JSM presents relatively low perturbation values while the only attack changing the inputs less is CW for $\epsilon \leq 0.2$. Much higher perturbation values characterize PGD and FGS. However, there is no significant impact on s_r nor α. Furthermore, JSM demonstrates the highest confidence values ($\approx$ 59%), while the remaining attack types are less than 55%. At the same time, confidence levels α increase with ϵ. Thus, we can conclude that for the specific type of industrial machine and AAS1, JSM is the best-performing adversarial attack.

D. Attack Application Scenario 2

The results obtained with respect to application scenario *AAS2* are demonstrated in Fig. 6. Similarly to *AAS1*, they are organised per machine type.

1) Fan: As expected, we observe that every figure of merit, i.e., s_r, p, and α increase with ϵ. The only quantity that does not exhibit changes is the confidence level associated with JSM, which shows that achieved levels are invariant to ϵ. Essentially, this shows that JSM creates adversarial samples, which are on the edge of serving the attacker's goal. Moreover, we observe that the highest s_r is offered by PGD followed by FGS and JSM, which exhibit similar values. On the contrary, CW offers much lower s_r's. Interestingly, JSM is associated with quite low p's, while the largest perturbations are introduced by FGS, PGD, and CW. Furthermore, PGD and FGS achieve quite high confidence levels, succeeded by CW and JSM. It comes out that the amount of introduced perturbations is not directly related to success rate nor confidence level. Approaching the problem from the attacker's point of view, the best-performing attack is PGD since it offers the highest s_r and α. Even though, p is also high, it might not be evident to the defence mechanism especially when considering the highly noisy environment that these machines typically operate in.

2) Pump: Overall, we observe that s_r, p, and α increase as ϵ varies from 0.1 to 0.5. Highest s_r's are achieved by JSM and PGD followed by FGS and CW. That said, JSM introduces quite small perturbations to the input log-Mel spectrograms, while FGS and PGD are associated with much higher p values. Moreover, PGD is able to mislead $\mathcal{M}$ with very high confidence scores, which even increase with ϵ to reach 100%. FGS follows, while CW and JSM demonstrate the lowest α values. Given the high s_r and alpha values, along with the implicit difficulty in quantifying perturbations on the defender side, PGD outperforms the remaining types of adversarial attacks.

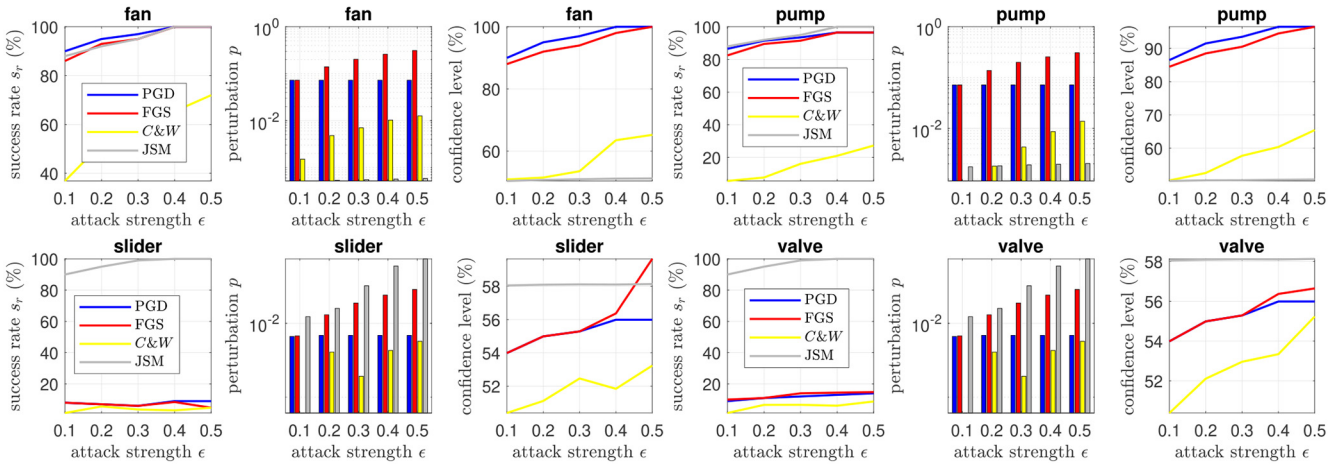

Fig. 6. *AAS2* scenario attacks sounds indicative of faulty states targeting to have them misclassified as normal, i.e., increasing false negatives w.r.t every machine. Order of figures per machine: (a) success rate (%) versus attack strength ϵ, (b) normalized perturbation versus attack strength ϵ, and (c) confidence level α (%) achieved for perturbed faulty samples versus attack strength ϵ.

TABLE I
BEST-PERFORMING ATTACKS PER INDUSTRIAL MACHINE TYPE AND APPLICATION SCENARIO

Industrial Machine	AAS1	AAS2
Fan	*PGD*	*PGD*
Pump	*PGD*	*PGD*
Slider	*JSM*	*JSM*
Valve	*JSM*	*JSM*

3) Slider: In this case, the figures of merit are not particularly influenced by the attack strength. *JSM* outperforms the remaining types of attack when it comes to s_r and α. At the same time PGD, FGS, and CW achieve low α and much lower s_r values. The best attack in terms of introduced amount of perturbations is CW followed by PGD, FGS, and finally JSM. We observe that even though p values increase substantially, it is not reflected in s_r and α values. Hence, the best-performing adversarial attack for slider in *AAS2* is JSM given the high s_r and α values.

4) Valve: When processing valve sounds, we see that JSM provides the highest success rates ($\approx$ 100), which are increased with ϵ. On the contrary, the remaining attack types, i.e., PGD, FGS, and CW, offer poor s_r's (<20%). Perturbations are increased as ϵ varies from 0.1 to 0.5. However, this is not directly reflected in s_r values. Moreover, CW is associated with the smallest amount of perturbations followed by PGD, FGS, and JSM. With regards to confidence, *JSM* offers the highest values across every ϵ ($\approx$ 58%), while the remaining attacks lie under 57%. Last but not least, it should be mentioned that α's associated with JSM is invariant to changes in ϵ. Ultimately, taking into account s_r and α, we can argue that JSM outperforms the rest of the attacks in *AAS2*.

E. Collective Analysis of the Experimental Results

Table I tabulates the best-performing attack types according to industrial machine and attack application scenario based on the achieved results. Such an analysis is based on two reasonable assumptions, i.e.: 1) $\mathcal{M}$ is able to classify original unperturbed samples with very high confidence levels and 2) it is particularly difficult to quantify the amount of introduced perturbation on the defender side due to the adverse, noisy environments that industrial machines typically operate in. As we can see, the best performing adversarial attack type of both fan and pump, during both *AAS1* and *AAS2*, is PGD. Furthermore, JSM outperforms the remaining attack types both for slider and valve during both *AAS1* and *AAS2*. As such, attack type selection is based solely on the machine that the attacker targets regardless of the attack application scenario.

In the following, we provide representative adversarial examples in order to understand how the attacks alter the log-Mel spectrograms toward misleading $\mathcal{M}$ to predict the target class. Fig. 7 demonstrate examples of attacks on normal sounds with $\epsilon = 0.5$, i.e., forcing $\mathcal{M}$ to predict false alarms. The top two rows concern PGD applied on fan and pump machine types. As we can see, the attack alters heavily the spectral content across every frequency bands with a slight preference in mid and low frequency regions. The two bottom rows represent JSM attacks on slider and valve machinery. In these cases, the alterations are less evident, while very low and very high frequency regions remain unaffected. Mid and low regions present only slight perturbations, while several small high frequency parts with strong changes can be distinguished. It comes out that JSM carefully selects the spectrogram parts to be altered, while PGD tries to change the input in a more uniform way.

Fig. 8 is analogous to Fig. 7 while the focus is placed on attacking sounds indicative of abnormal machine conditions, i.e., the attacker aims at increasing the false negative rate. Similar attack behavior to Fig. 5 can be observed with PGD heavily altering the nearly the entire spectral content, while JSM carries out a filtering process. Different to Fig. 5, JSM attacks against slider machines affect less mid and low spectrogram parts. Interestingly, JSM emphasizes on frequency parts where energy is already large and increases further the respective values. Overall, we argue that a relatively small amount

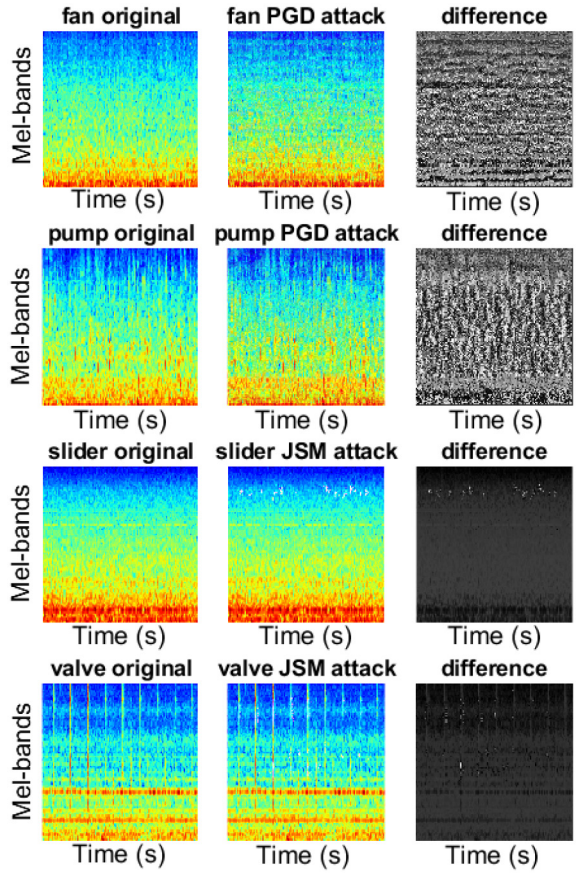

Fig. 7. Examples of how the best-performing attacks (PGD and JSM) perturb the spectral content of sounds indicative of normal machine state for creating adversarial samples. The log-Mel spectrogram difference is depicted as a third subfigure.

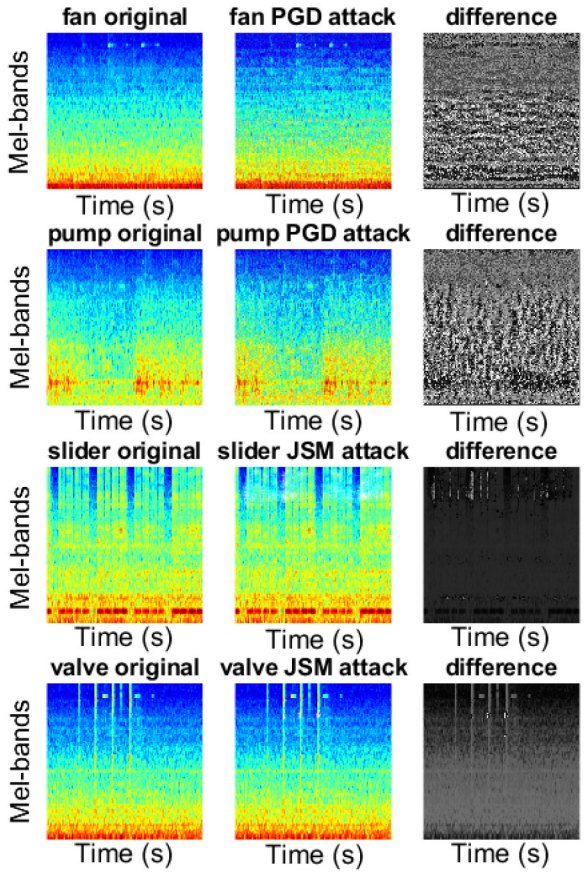

Fig. 8. Examples of how the best-performing attacks (PGD and JSM) perturb the spectral content of sounds indicative of machine faults for creating adversarial samples. The log-Mel spectrogram difference is depicted as a third subfigure.

of perturbations can purposefully mislead a system designed to monitor the health of industrial machines.

VI. Conclusion

This work exposed the vulnerabilities of modern audio-based industrial machine health monitoring systems and systematically investigated the efficacy of several adversarial attacks on such monitoring tools. After a carefully designed experimental process considering diverse aspects of the specific problem, it highlighted that such systems are indeed vulnerable in both types of errors, i.e., missing an existing fault and detecting a fault when there is not (*AAS1* and *AAS2*). In addition, we extensively analyze four attack types applied onto four machines and identified the best-performing attacks for each machine by considering various criteria at the same time (s_r, p, and α). It was shown that a potentially successful attack can depend on state of the art adversarial algorithms, while injecting only minor perturbations to the input audio. To the best of our knowledge, this is the first time that adversarial attacks have been systematically investigated in the specific field of research. This study can assist the design of efficient defence mechanisms able to compensate for the perturbations injected by the adversarial attacks.

In the future, we are going to focus on the following aspects.

1) Assess the performance of adversarial attacks in other types of audio signals containing speech and nonspeech content, as they are currently understudied,
2) Explore a wide range of defense strategies, e.g., based on psychoacoustic principles [31], audio steganography [32], exploiting temporal dependencies [33], etc. A promising direction could include the compression, e.g., MP3, of manipulated audio recordings based on psychoacoustic models which might remove imperceptible artifacts that might confuse the model. An additional direction considers the application of transformations, such as quantization, local smoothing, downsampling, etc., which may potentially limit or even eliminate the perturbations of adversarial attacks.
3) Investigate the transferability of adversarial attacks across DL models with diverse architectures, parameters, and training data sets.

Overall, this work highlighted the vulnerabilities existing in modern DL-based solutions and wishes to trigger more researchers into investigating the potentially severe consequences of adversarial attacks in diverse application domains.

Acknowledgment

This work was carried out within the project entitled "Advanced methods for sound and music computing" funded

by the University of Milan. We would like to thank the creators of the Adversarial Robustness toolbox and MIMII Data set.

REFERENCES

[1] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–9.

[2] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. ICLR Workshop*, 2017, pp. 1–14.

[3] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives [review article]," *IEEE Comput. Intell. Mag.*, vol. 14, no. 2, pp. 68–81, May 2019.

[4] Y. Song, T. Liu, T. Wei, X. Wang, Z. Tao, and M. Chen, "FDA^3: Federated defense against adversarial attacks for cloud-based IIoT applications," *IEEE Trans. Ind. Informat.*, vol. 17, no. 11, pp. 7830–7838, Nov. 2021.

[5] P. Tyrväinen, M. Silvennoinen, K. Talvitie-Lamberg, A. Ala-Kitula, and R. Kuoremäki, "Identifying opportunities for AI applications in healthcare—Renewing the national healthcare and social services," in *Proc. IEEE SeGAH*, 2018, pp. 1–7.

[6] S. Ntalampiras, "Fault identification in distributed sensor networks based on universal probabilistic modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 1939–1949, Sep. 2015.

[7] M. X. Cheng and W. B. Wu, "Data analytics for fault Localization in complex networks," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 701–708, Oct. 2016.

[8] R. Marino, C. Wisultschew, A. Otero, J. M. Lanza-Gutierrez, J. Portilla, and E. D. L. Torre, "A machine-learning-based distributed system for fault diagnosis with scalable detection quality in industrial IoT," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4339–4352, Mar. 2021.

[9] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, "A taxonomy and terminology of adversarial machine learning," Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Internet-Draft NISTIR 8269, Oct. 2019.

[10] M.-I. Nicolae *et al.*, "Adversarial robustness toolbox v1.0.0," 2018, *arXiv:1807.01069*.

[11] V. Subramanian, A. Pankajakshan, E. Benetos, N. Xu, S. McDonald, and M. Sandler, "A study on the transferability of adversarial attacks in sound event classification," in *Proc. ICASSP*, 2020, pp. 301–305.

[12] H. Kwon, Y. Kim, H. Yoon, and D. Choi, "Selective audio adversarial example in evasion attack on speech recognition system," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 526–538, 2020.

[13] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE SPW*, 2018, pp. 1–7.

[14] J. Henry, M. Ergezer, and M. Orescanin, "Perceptually constrained fast adversarial audio attacks," in *Proc. ICMLA*, 2021, pp. 819–824.

[15] B. L. Junchenl, B. C. Bingqinc, and Z. Z. Zhuoran, "Real world audio adversary against wake-word detection systems," in *Proc. NeurIPS*, Vancouver, BC, Canada, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/ebbdfea212e3a756a1fded7b35578525-Abstract.html

[16] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "A robust approach for securing audio classification against adversarial attacks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2147–2159, 2020.

[17] H. Rafezi and F. Hassani, "Tricone bit health monitoring using wavelet packet decomposed vibration signal," in *Proc. CoDIT*, 2018, pp. 1012–1016.

[18] Q. Fu, B. Jing, P. He, S. Si, and Y. Wang, "Fault feature selection and diagnosis of rolling bearings based on EEMD and optimized Elman_AdaBoost algorithm," *IEEE Sensors J.*, vol. 18, no. 12, pp. 5024–5034, Jun. 2018.

[19] B. Li, M.-Y. Chow, Y. Tipsuwan, and J. C. Hung, "Neural-network-based motor rolling bearing fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 47, no. 5, pp. 1060–1069, Oct. 2000.

[20] Q. Yang, C. Hu, and N. Zheng, "Data-driven diagnosis of nonlinearly mixed mechanical faults in wind turbine gearbox," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 466–467, Feb. 2018.

[21] J. Xiong, Q. Liang, J. Wan, Q. Zhang, X. Chen, and R. Ma, "The order statistics correlation coefficient and PPMCC fuse non-dimension in fault diagnosis of rotating petrochemical unit," *IEEE Sensors J.*, vol. 18, no. 11, pp. 4704–4714, Jun. 2018.

[22] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Trans. Ind. Informat.*, vol. 11, no. 3, pp. 812–820, Jun. 2015.

[23] R. Tanabe *et al.*, "MIMII due: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," in *Proc. IEEE WASPAA*, 2021, pp. 21–25.

[24] S. Ntalampiras, "Moving vehicle classification using wireless acoustic sensor networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 129–138, Apr. 2018.

[25] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019.

[26] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, Vancouver, BC, Canada, Apr./May 2018, pp. 1–23. [Online]. Available: OpenReview.net

[28] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE EuroSP*, 2016, pp. 372–387.

[29] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy (SP)*, 2017, pp. 39–57.

[30] M. Casey, "General sound classification and similarity in MPEG-7," *Organised Sound*, vol. 6, no. 2, pp. 153–164, 2002.

[31] N. Das, M. Shanbhogue, S.-T. Chen, L. Chen, M. E. Kounavis, and D. H. Chau, "ADAGIO: Interactive experimentation with adversarial attack and defense for audio," in *Machine Learning and Knowledge Discovery in Databases*. Dublin, Ireland: Springer Int., 2019, pp. 677–681.

[32] J. Wu, B. Chen, W. Luo, and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2282–2294, 2020.

[33] Z. Yang, B. Li, P. Chen, and D. Song, "Characterizing audio adversarial examples using temporal dependency," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

Stavros Ntalampiras received the engineering and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Patras, Greece, in 2006 and 2010, respectively.

He is an Associate Professor with the Department of Computer Science, University of Milan, Milan, Italy. He has carried out research and/or didactic activities with Politecnico di Milano, Milan, the Joint Research Center of the European Commission, the National Research Council of Italy, Rome, Italy, and Bocconi University, Milan. His research interests include content-based signal processing, machine learning, audio pattern recognition, medical acoustics, bioacoustics, and cyber–physical systems.

Dr. Ntalampiras is currently an Associate Editor of IEEE ACCESS, *PLOS One*, *IET Signal Processing*, and *CAAI Transactions on Intelligence Technology*, as well as a member of the IEEE Computational Intelligent Society Task Force on Computational Audio Processing.