technologies

Article

Disrupting Audio Event Detection Deep Neural Networks with White Noise

Rodrigo dos Santos *, **Ashwitha Kassetty and Shirin Nilizadeh**

Computer Science and Engineering Department, Main Campus, The University of Texas at Arlington,
701 S Nedderman Dr., Arlington, TX 76019, USA; ashwitha2000@gmail.com (A.K.);
shirin.nilizadeh@uta.edu (S.N.)
* Correspondence: rodrigoaugusto.silvadossantos@uta.edu

Abstract: Audio event detection (AED) systems can leverage the power of specialized algorithms for detecting the presence of a specific sound of interest within audio captured from the environment. More recent approaches rely on deep learning algorithms, such as convolutional neural networks and convolutional recurrent neural networks. Given these conditions, it is important to assess how vulnerable these systems can be to attacks. As such, we develop AED-suited convolutional neural networks and convolutional recurrent neural networks, and attack them next with white noise disturbances, conceived to be simple and straightforward to be implemented and employed, even by non-tech savvy attackers. We develop this work under a safety-oriented scenario (AED systems for safety-related sounds, such as gunshots), and we show that an attacker can use such disturbances to avoid detection by up to 100 percent success. Prior work has shown that attackers can mislead image classification tasks; however, this work focuses on attacks against AED systems by tampering with their audio rather than image components. This work brings awareness to the designers and manufacturers of AED systems, as these solutions are vulnerable, yet may be trusted by individuals and families.

Keywords: AED; neural networks; deep learning; spectrograms

check for updates

Citation: dos Santos, R.; Kassetty, A.; Nilizadeh, S. Disrupting Audio Event Detection Deep Neural Networks with White Noise. *Technologies* **2021**, *9*, 64. https://doi.org/10.3390/technologies9030064

Academic Editor: Pedro Antonio Gutiérrez

Received: 25 August 2021
Accepted: 1 September 2021
Published: 6 September 2021

1. Introduction

IoT-based cyber–physical systems (CPSs) have been developed to advance personalized health care, emergency response, traffic flow management, and electric power generation and delivery [1–7]. IoT-based CPSs include smart networked systems with embedded sensors, processors and actuators that sense and interact with the physical world. A core element of IoT-based CPSs is *decision making*, which analyzes the data obtained from sensors. As more sensors became pervasive, generating huge volumes of data, deep learning algorithms, in particular, more recent neural networks, are also becoming pervasive, playing a vital role in the good performance of decision-making in real-life CPS applications.

The growing use of deep learning approaches in such critical purpose applications has raised concerns about their robustness against malicious attacks, where the attacker tries to fool the deep learning model. Some studies have already shown that deep neural network classifiers are susceptible to attacks aimed at causing misclassifications [8–10]. Most attacks against these classifiers are focused on image classification tasks [11–14]. An example would be changing some particular pixels [15,16] or creating new images that will misclassify to a target [17,18], maliciously chosen by the attacker. A few recent works studied attacks against speech and speaker recognition applications [19–21]. These attacks involve generating malicious audio commands that are recognized and interpreted by the audio models, used in voice processing systems, but are either inaudible or sound like mere noise to the human ear.

Speech recognition is just one example of the possible audio processing capabilities made available by CPS systems. Another one would be that of audio event detection (AED), where the systems obtain the audio inputs from some acoustic sensors and then process them to detect and classify some sonic events, such as gunshots. This is different from speech recognition in the sense that AED algorithms, unlike SR ones, do not look for specific phonetic sequences to identify a specific sound event [22]. Additionally important to mention is that, unlike in SR, where the basic units of sounds are similar (e.g., phonemes) [23,24], in AED, the different sound events (e.g., dog bark, gunshot) all present very distinct patterns, thus making the AED task more challenging.

AED systems have been employed for safety purposes, through the detection of suspicious sounds such as gunshots, footsteps and others. Gunshot sound detection has been extensively researched, and represents a good starting point for this work. AED systems for gunshot detection can be employed anywhere, from home to business, and even public spaces, where they would constantly monitor the environment for suspicious events. These systems currently make extensive use of state-of-the-art deep learning classifiers, such as convolutional neural networks (CNN) [25] and convolutional recurrent neural networks (CRNN) [26], as their primary detection and classification algorithms.

In this paper, we examine these two state-of-the-art gunshot detection algorithms against simple, accessible and easy-to-reproduce disturbances made of noise, to be used as a means of disrupting the classifiers. Several studies have shown that unwanted noise can have a detrimental effect on classifier performance [27,28]. We focus on this niche, and as such, study how to attack deep learning AED systems, focusing on employing simple, accessible and easy-to-reproduce disturbances made of white noise, to be used as a means of disrupting the classifiers. A major reason for choosing white noise is the concrete possibility of employing it as part of practical, on-the-field attacks. Another reason is the simplicity of the attack, thus making it largely available for a large roster of attackers.

In our threat model (seen in Figure 1), we assume that the attacker, while attempting to cause harm, actively adds white noise perturbations to the sound being fed to the AED system; in other words, white noise is overlaid to to the gunshot sounds being used as the AED input. The ultimate goal of the attacker is to prevent the AED system from detecting the gunshot sounds. We implement a CNN and a CRNN, and we use gunshot sound datasets from [29–31]. We first test the classifiers with undisturbed gunshot samples in order to examine their performance under baseline conditions, and then digitally inject white noise perturbations, interleaving them with the gunshot sounds. Experiments covering real on-the-field experiments will be covered in future publications. Our consolidated results show that AED classifiers are susceptible against these noisy attacks, as the performance of both the CNN and CRNN are strongly affected, being degraded by nearly 100% when tested against the perturbations.

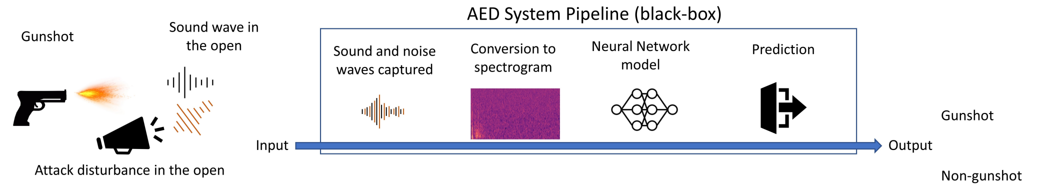

Gunshot
Sound wave in the open
Attack disturbance in the open
AED System Pipeline (black-box)
Sound and noise waves captured
Conversion to spectrogram
Neural Network model
Prediction
Input
Output
Gunshot
Non-gunshot

Figure 1. A hypothetical scenario where an attacker tries to evade detection by injecting noise sounds, in the open, to the audio input captured by AED systems.

In particular, our contributions reside on attacking deep learning classifiers with a simple and very easy to reproduce disturbance, which is relevant in present day when there is not only a proliferation of real deep learning devices that rely on deep learning classifiers for suspicious sound detection, but also of sophisticated audio equipment that could be used for practical attack purposes.

2. Materials and Methods

In this section, we present our overall research methodology and setup as well as the details behind the two supervised neural network algorithms we implemented, namely CNN and CRNN. These two neural network algorithms are state-of-the-art classification algorithms that were recently used for audio detection and classification in IoT-based CPSs, including gunshots and other safety-driven applications. We implement these algorithms as binary classifiers, where we feed them audio samples as an input, and they provide an output label of *gunshot (1)* or *non-gunshot (0)*.

2.1. CNN Models in Audio Detection Systems

Convolutional neural networks are considered to be the best among learning algorithms in understanding image contents [32]. CNNs were inspired by the organization of the animal visual cortex [33], providing increasingly better performance as they become deeper, while also becoming harder to train [34]. We implement a CNN model based on the work of Zhou et al. [25], as it was successfully used for the purpose of urban sound recognition, gunshots included. Our model is tailored after much experimentation, and it is composed of the following:

1. Convolutional layers: three convolutional blocks, each one with two convolutional 2D layers. These layers have 32, 64, 64, 64, 128 and 128 filters (total of 480) of size 3 by 3. Same padding is also applied to the first convolutional layer of each block.
2. Pooling layers: three 2 by 2 max pooling layers, each coming right after the second convolutional layer of each convolutional block.
3. Dense layers: two dense (also known as fully connected) layers come after the last convolutional block.
4. Activation functions: these functions compute the weighted sum of inputs and biases, and as such, are responsible for the firing or no firing of neurons [35]. For the presented CNN, ReLU activation is applied after each convolutional layer as well as after the first fully connected layer, while Softmax activation is applied only once, after the second fully connected layer. In other words, ReLU is applied to all inner layers, while Softmax is applied to the most outer layer.
5. Regularization: applied in the end of each convolutional block as well as after the first fully connected layer, with 25, 50, 50 and 50% respectively. Regularization, also known as dropout, per [36], addresses the overfitting problem, among other common neural network issues.

   The CNN uses sparse categorical cross entropy as a loss function and RMSprop as an optimizer. A visual representation of its architecture can be seen in Figure 2.

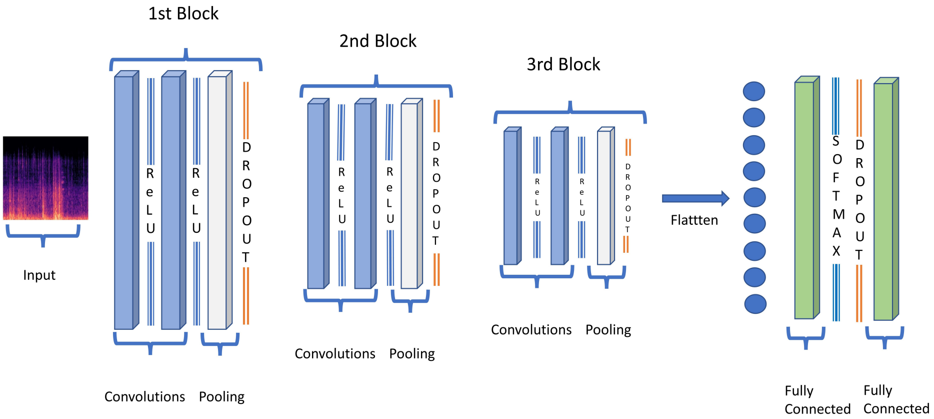


Figure 2. CNN architecture.

2.2. CRNN Models in Audio Detection Systems

Convolutional recurrent neural networks (CRNN) address one shortfall of regular CNNs—the lack of memory about long-term dependencies over sequences [37,38]. Xinyu Fu [37] proposed to implement a CRNN, where common sigmoid activation function is replaced by a long short-term memory (LSTM) advanced activation. It was shown that CRNNs work better with longer or lengthier inputs [38] because LSTM activation ensures that outputs of the previous point in time connect to the next point in time.

For the purpose of this study, we implement a CRNN model inspired by the work of Lim et al. [26], which was also successfully used for gunshot recognition with some good performance. We tailor the model after much experimentation, which is composed by the following:

1. Convolutional layers: one convolutional block, with one convolutional layer. This block is made by 128 filters of size 32, ReLU activation and batch normalization, pooling layer of size of 40 and a dropout layer of 0.3.
2. LSTM layer: one backwards LSTM layer with 128 units, followed by tanh activation and a new dropout of 0.3.
3. Dense layers: two stacked dense layers, the first with 128 units and the second with two, each one followed by batch normalization and the first one followed by a ReLU activation and the last one by a Softmax activation.

The CRNN used sparse categorical cross entropy as loss function and Adam as optimizer.

2.3. Datasets

DCASE 2017 Challenge [30] provides datasets for several different acoustic classification tasks. We acquired the *detection of rare sound events* dataset. The reason for the selection of this specific dataset as our main source of data points stem from the fact that it is publicly available, making it easily acquirable, while also containing a relatively high number of good quality sounds. This dataset includes one training and one testing set, each one containing 1500 sound samples. From 3000 samples in total, 1000 are gunshot sounds, which are the ones used in our experiments. To increase the number of data points available to our research, we also obtained gunshot samples from [29,31], bringing the total number of samples to 2000. For our experiments, 1500 of these sounds are dedicated for training and 500 for testing. Finally, we obtained the samples for the negative classes from several other sources, such as MIMII [39], environmental Sounds [40], FreeSound [41], Zapsplat [42] and Fesliyan Studios [43]. The negative classes are made of *pump, children playing, fan, valve* and *music*, in other words, sounds that did not carry gunshot sounds.

It is important to mention that we cleaned and preprocessed the audio by performing *frequency normalization* (normalizing the frequencies of all samples at 22,000 Hertz), *audio channel normalization* (set all samples to monaural) and *audio length normalization* (made all samples to have 3 seconds in length by cropping lengthier ones and discarding shorter ones).

2.4. Spectrograms

We adopt the same approach adopted by other authors, such as [25–27] with regards to the input to be fed to our models, relying on spectrograms. Audio files can be represented by *spectrograms*. Spectrograms display in a graph (usually 2D) the spectrum of frequency changes over time for a sound signal, by chopping it up and then stacking the slices close to each other [44]. Unlike speech, sound events often have a shorter duration but with more distinctive time-frequency representations, which was shown to be a good feature for sound classification [25,26,45]. Spectrograms are images, and they fit well as input to both CNN and CRNN models.

The authors in [25,26] used spectrograms as input to their neural network models. Similarly, we also convert the audio samples into spectrograms and then pass them to the CNN and CRNN classification algorithms. Note that the disturbances represented by the proposed attacks are introduced directly to the audio files, prior to their conversion to spectrograms. As such, in our threat model, the attacker does not have direct access to the spectrogram generation algorithm (black-box IoT/CPS system). This is because in our attack model, we assume that the attacker does not have any knowledge about the system and simply tries to alter the sounds generated by the gun, before capture by the hypothetical AED system. Samples of spectrograms generated as part of this study can be seen in Figure 3.

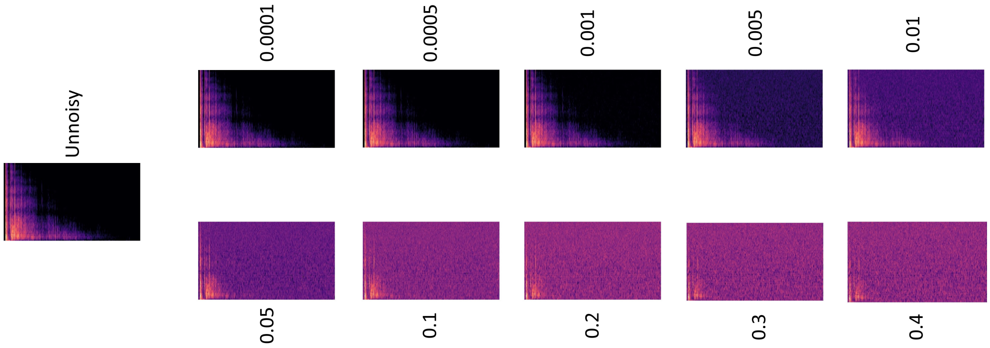


Figure 3. Spectrogram samples under unnoisy and noisy conditions.

2.5. *Attacks with Noise*

Every practical application that deals with audio signals also deals with the issue of noise. As stated by Prasadh et al. [46], "natural audio signals are never available in pure, noiseless form". As such, even under ideal conditions, natural noise may be present in the audio in use. In this work, we craft noisy disturbances made of white noise, which, according to [47], happens when "each audible frequency is equally loud," meaning no sound feature, "shape or form can be distinguished". We choose this noise variant, due to its ubiquity in day-to-day life, and especially its simplicity with regard to how to reproduce it. It is important to highlight once again that in our threat model, the attack occurs during audio capturing, prior to spectrogram generation; as such, our proposed attack is audio-based, not image-based. This type of noise is widely adopted by different research across different domains [48–50], thereby facilitating future comparisons.

2.6. *Experiments*

Our experiments involve the use of our two neural network classifiers set as two different representation of an AED system that detects gunshot sounds. We employ digital gunshot samples, first in unnoisy conditions, and then we infuse the same samples with progressively higher levels noise. We crafted our own training and testing sets, and employ them to train and test our CNN models. The experiments were binary (output could be either *gunshot* and *non-gunshot*. Both the training and test sets always had the two participating classes in a balanced fashion. In other words, we always made sure to have the same number of samples per class in each experiment.

A summary of our experiments follows next.

1. Unnoisy experiments: Both AED classifiers exposed to digital gunshot sounds, without any disturbance.
2. White noise experiments: Both AED classifiers exposed to digital gunshot sounds. The disturbances, made of white noise, are dynamic in nature, and after much experimentation, ten different noisy thresholds, namely *0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3 and 0.4*. The process for generating the white noise infused samples can be seen in Algorithm 1.

Algorithm 1: White Noise Generation Algorithm

```
Result: Perturbed audio sample
initialization;
for number of audio files in the test set do
    sample = load audio file as an array;
    noise = adjustment factor * max element of the array;
    perturbed sample = sample + noise * normal distribution;
    save perturbed sample;
end
```

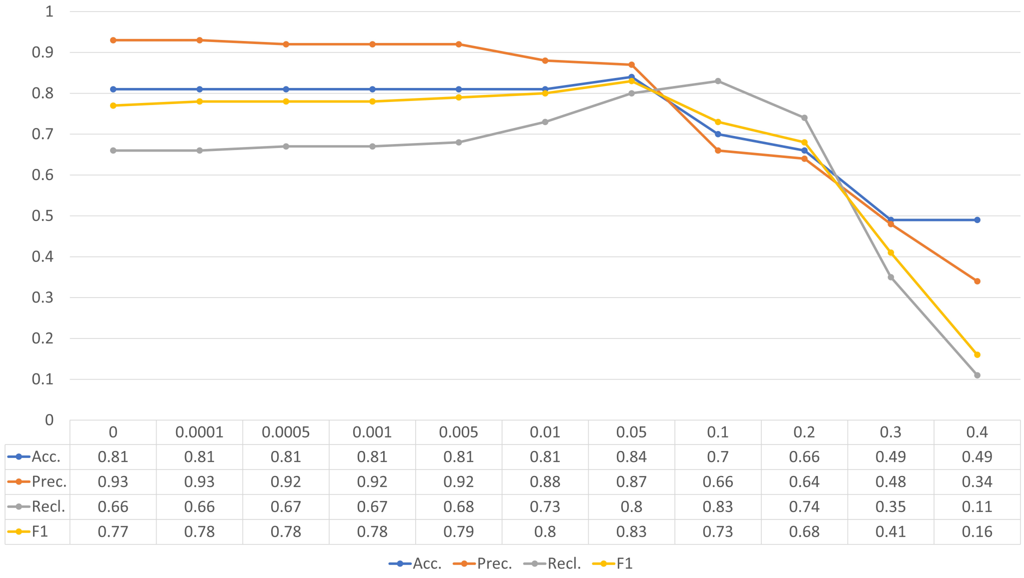
3. Results

In this section, we present the results obtained from the several experiments that show the robustness of CNN and CRNN classifiers against the white noise attacks. We start by providing the baseline classification results achieved by both models, CNN and CRNN, using the undisturbed test samples, which can be found in Table 1. Both models were trained for 100 epochs; however, the last learning occurred at 52 and 98 epochs (CNN) and at 56 and 29 epochs (CRNN), respectively. Both algorithms are trained and tested on the same datasets. The full experiment results can be seen in Table 1.

Table 1. CNN / CRNN classification performance on noise-free test datasets followed by increasing levels of white noise.

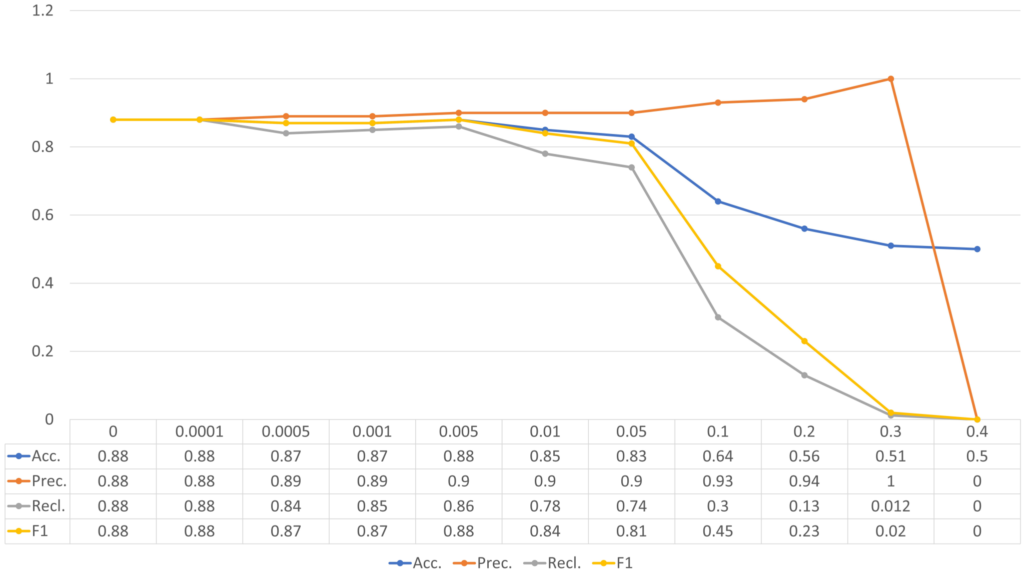
	CNN				CRNN			
Condition	**Acc.**	**Prec.**	**Rcl.**	**F1**	**Acc.**	**Prec.**	**Rcl.**	**F1**
Unnoisy	0.88	0.88	0.88	0.88	0.81	0.93	0.66	0.77
0.0001	0.88	0.88	0.88	0.88	0.81	0.93	0.66	0.78
0.0005	0.87	0.89	0.84	0.87	0.81	0.92	0.67	0.78
0.001	0.87	0.89	0.85	0.87	0.81	0.92	0.67	0.78
0.005	0.88	0.90	0.86	0.88	0.81	0.92	0.68	0.79
0.01	0.85	0.90	0.78	0.84	0.81	0.88	0.73	0.80
0.05	0.83	0.90	0.74	0.81	0.84	0.87	0.80	0.83
0.1	0.64	0.93	0.30	0.45	0.70	0.66	0.83	0.73
0.2	0.56	0.94	0.13	0.23	0.66	0.64	0.74	0.68
0.3	0.51	1	0.012	0.02	0.49	0.48	0.35	0.41
0.4	0.5	0	0	0	0.49	0.34	0.11	0.16

When executing the baseline unnoisy experiments, both models perform reasonably well, with accuracies above 80%. This sets the tone for the experiments that come next, where we proceed to attack these same classifiers with white noise. When this happens, both models present drops in classification performance as soon as such noise is introduced to the test sets. The drops are small but cumulative, and a sharper drop is noticed when the 0.1 threshold is reached, only to become unacceptably worse from there on, to the point of rendering both models essentially useless. One can also realize that the CRNN is proved to be slightly more robust than the CNN, and we credit this to its memory advantage over the CNN [37,38]. The graphical visualization of these results can be seen in Figures 4 and 5.



	0	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4
Acc.	0.81	0.81	0.81	0.81	0.81	0.81	0.84	0.7	0.66	0.49	0.49
Prec.	0.93	0.93	0.92	0.92	0.92	0.88	0.87	0.66	0.64	0.48	0.34
Recl.	0.66	0.66	0.67	0.67	0.68	0.73	0.8	0.83	0.74	0.35	0.11
F1	0.77	0.78	0.78	0.78	0.79	0.8	0.83	0.73	0.68	0.41	0.16

Figure 4. CNN experiments detailed results.



	0	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4
Acc.	0.88	0.88	0.87	0.87	0.88	0.85	0.83	0.64	0.56	0.51	0.5
Prec.	0.88	0.88	0.89	0.89	0.9	0.9	0.9	0.93	0.94	1	0
Recl.	0.88	0.88	0.84	0.85	0.86	0.78	0.74	0.3	0.13	0.012	0
F1	0.88	0.88	0.87	0.87	0.88	0.84	0.81	0.45	0.23	0.02	0

Figure 5. CRNN experiments detailed results.

4. Discussion

We tested CNN and CRNN algorithms for AED, and while their detection performance was reasonable under ideal circumstances, a sharp drop in it was seen, even when little white noise was injected into the test audio samples. This is important because white noise is simple to reproduce and to be employed by non-technically savvy individuals. It also can be hard to be filtered out without affecting the sound capture capability needed for an AED system, especially when higher noisy thresholds are used and when its amplitude is tailored to closely follow that of the sound of interest.

AED practical solutions are already a reality, currently being in the process of becoming ubiquitous for a large audience. These are real physical devices that employ deep learning models for the detection of suspicious events for security purposes, being largely available for purchase and deployment to homes around the world. Some examples of these devices are ones manufactured by major companies, such as the *Echo Dot* and *Echo Show* by Amazon [51], and *Nest Mini* and *Nest Hub* by Google [52,53]. Despite still being limited in terms of detection capabilities, as most of these devices can detect only a few variety of audio events, attempts to create general purpose devices, capable of detecting a wide spectrum of audio events, are known to be in the making, e.g., See-Sound [54]. Practical attacks against these devices are, as such, just a matter of time.

As a matter of fact, white-noise reproduction capable gear based on speakers and other specialized equipment have been widely available to the broad audience for a long time now [55]. These can become physical devices that generate audio disturbances on the field. More sophisticated gear with increased capabilities are also a reality and are intensively researched and used by military and law enforcement agencies around the world [56–59]. As such, attack solutions that could rely on all sorts of gear, from tiny and discrete devices to large and complex ones are available today.

One cannot ignore the apparent heavy planning needed in order to implement such attacks from a practical standpoint; however, one cannot also ignore the motivation of attackers who intend to do harm. For example, the attack that occurred in the Mandalay Hotel at Las Vegas [60] showcases such motivation, as the attacker spent months smuggling guns and ammunition into his hotel room, and even went to the extent of setting, possibly, other sensors in the corridor leading to his room, so he would be better prepared to deal with law enforcement officials when they responded to the emergency situation he was about to set. Therefore, it is not a stretch to envision a scenario where an attacker (e.g., burglar) could plan for days, weeks, or even months in advance on how to deploy attacks against an audio-based AED system.

By doing so, such an attacker would either delay or avoid detection by an AED system, and as such, gain time to perform their malicious intents. For instance, a burglar could use an "on-the-field" variant of the white noise attack to disrupt a home-based AED system, and thus be able to invade an empty residency without being detected and/or triggering AED-associated alarms and notifications (since a potential glass breakage would not be detected by the under-attack AED system). After gaining entrance, the burglar could potentially perform their activities, such as robbery, without being disturbed. As AED systems gain popularity and scale, it is not difficult to envision a scenario where an AED system may be protecting a public area, and terrorists, aware of such monitoring, employ white noise disturbances to disrupt such a system. This, in turn, would make it hard for authorities to respond properly through the negation of the system's ability to detect a sound of interest (e.g., gunshots being fired) and subsequently to relay the location of the sound.

5. Conclusions

Through extensive amount of experiments, we evaluated the robustness of AED systems against noisy attacks, which are generated by adding different levels of white noise to test samples fed to neural network classifiers. Our consolidated results show that AED systems are susceptible against white noise attacks, as the performance of both the CNN and CRNN classifiers were degraded by nearly 100% when tested against the perturbations. These strongly suggest that actually deployable solutions that rely on these AED classifiers cannot be trusted, especially if they are used for safety purposes. More research is currently underway, seeking to test the white disturbances as well as other types of audio disturbances on real, existing AED capable devices. We also are currently actively researching practical countermeasures for these attacks and will report the appropriate results in upcoming publications.

Author Contributions: Conceptualization, R.d.S. and S.N.; methodology, R.d.S. and S.N.; software, R.d.S.; validation, R.d.S.; formal analysis, R.d.S.; investigation, R.d.S.; resources, S.N.; data curation, R.d.S. and A.K.; writing—original draft preparation, R.d.S.; writing—review and editing, S.N.; visualization, R.d.S. and S.N.; supervision, S.N.; project administration, R.d.S.; funding acquisition, S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AED	Audio Event Detection
SR	Speech Recognition
DL	Deep Learning
ML	Machine Learning

References

1. Wang, Z.; Song, H.; Watkins, D.; Ong, K.; Xue, P.; Yang, Q.; Shi, X. Cyber-physical systems for water sustainability: Challenges and opportunities. *IEEE Commun. Mag.* **2015**, *53*, 216–222. [CrossRef]
2. Zhang, Y.; Qiu, M.; Tsai, C.; Hassan, M.; Alamri, A. Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Syst. J.* **2015**, *11*, 88–95. [CrossRef]
3. Wang, L.; Onori, M. Current status and advancement of cyber-physical systems in manufacturing. *J. Manuf. Syst.* **2015**, *37*, 517–527. [CrossRef]
4. Zander, J.; Mosterman, P.; Padir, T.; Wan, Y.; Fu, S.X. Cyber-physical systems can make emergency response smart. *Procedia Eng.* **2015**, *107*, 312–318. [CrossRef]
5. Reddy, R.; Mamatha, C.; Reddy, R. A Review on Machine Learning Trends, Application and Challenges in Internet of Things. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 19–22 September 2018.
6. Chan, R. Internet of Things Business Models. *J. Serv. Sci. Manag.* **2015**, *8*, 552–568. [CrossRef]
7. Venkatesh, J.; Aksanli, B.; Chan, C.; Akyurek, A.; Rosing, T. Scalable-Application Design for the IoT. *IEEE Software* **2017**, *34*, 62–70. [CrossRef]
8. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017.
9. Brown, T.; Mane, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial Patch. *arXiv* **2017**, arXiv:1712.09665.
10. Goodfellow, I.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
11. Akhtar, N.; Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* **2018**, *6*, 14410–14430. [CrossRef]
12. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing robust adversarial examples. *arXiv* **2017**, arXiv:1707.07397.
13. Metzen, H.; Kumar, C.; Brox, T.; Fischer, V. Universal adversarial perturbations against semantic image segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
14. Su, J.; Vargas, D.; Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828–841. [CrossRef]
15. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2016**, arXiv:1607.02533.
16. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
17. Nguyen, A.; Yosinski, J.; Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
18. Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.; Zhai, J.; Wang, W.; Zhang, X. Trojaning attack on neural networks. In Proceedings of the Network and Distributed Systems Security (NDSS) Symposium, San Diego, CA, USA, 26 February–1 March 2017.
19. Carlini, N.; Mishra, P.; Vaidya, T.; Zhang, Y.; Sherr, M.; Shields, C.; Zhou, W. Hidden Voice Commands. In Proceedings of the USENIX Security, Austin, TX, USA, 10–12 August 2016.
20. Carlini, N.; Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In Proceedings of the IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24 May 2018.
21. Kwon, H.; Kim, Y.; Yoon, H.; Choi, D. Selective audio adversarial example in evasion attack on speech recognition system. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 526–538. [CrossRef]
22. Hamid, A.; Mohamed, A.; Jiang, H.; Deng, L. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1533–1545. [CrossRef]
23. Bilen, Ç.; Ferroni, G.; Tuveri, F.; Azcarreta, J. A Framework for the Robust Evaluation of Sound Event Detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020.
24. Cowling, M. Comparison of techniques for environmental sound recognition. *Pattern Recognit. Lett.* **2003**, *10*, 2895–2907. [CrossRef]
25. Zhou, H.; Song, Y.; Shu, H. Using deep convolutional neural network to classify urban sounds. In Proceedings of the IEEE Region 10 Conference (TENCON), Penang, Malaysia, 5–8 November 2017.
26. Lim, H.; Park, J.; Han, Y. Rare sound event detection using 1D convolutional recurrent neural networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, Munich, Germany, 16–17 November 2017.

27. Alraddadi, S.; Alqurashi, F.; Tsaramirsis, G.; Luhaybi, A.; Buhari, S. Aroma Release of Olfactory Displays Based on Audio-Visual Content. *Appl. Sci.* **2019**, *9*, 4866. [CrossRef]
28. Boyat, A.; Joshi, B. A review paper: Noise models in digital image processing. *arXiv* **2015**, arXiv:1505.03489.
29. The Free Firearm Library—Expanded Edition. Available online: airbornesound.com (accessed on 1 January 2021).
30. Detection of Rare Sound Events. Available online: cs.tut.fi/sgn/arg/dcase2017/challenge/task-rare-sound-event-detection (accessed on 1 January 2021).
31. A Dataset and Taxonomy for Urban Sound Research. Available online: justinsalamon.com (accessed on 1 January 2021).
32. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A. A survey of the recent architectures of deep convolutional neural networks. *arXiv* **2019**, arXiv:1901.06032
33. Yamashita, R.; Nishio, M.; Do, R.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef]
34. Thakkar, V.; Tewary, S.; Chakraborty, C. Batch Normalization in Convolutional Neural Networks—A comparative study with CIFAR-10 data. In Proceedings of the Fifth International Conference on Emerging Applications of Information Technology (EAIT), Kolkata, India, 12–13 January 2018.
35. Nwankpa, C.; Ijomah, W.; Gachagan, A.; Marshall, S. Activation Functions: Comparison of Trends in Practice and Research for Deep Learning. In Proceedings of the Machine Learning: Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
36. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1919–1958.
37. Fu, X.; Ch'ng, E.; Aickelin, U.; See, S. CRNN a joint neural network for redundancy detection. In Proceedings of the IEEE International Conference on Smart Computing (SMARTCOMP), Hong Kong, China, 29–31 May 2017.
38. Gao, S.; Lin, B.; Wang, C. Share price trend prediction using CRNN with LSTM structure. In Proceedings of the International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 6–8 December 2018.
39. Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection. Available online: zenodo.org/ (accessed on 1 January 2021).
40. ESC-50 Dataset for Environmental Sound Classification. Available online: github.com/karolpiczak/ESC-50 (accessed on 1 January 2021).
41. Freesound. Available online: freesound.org/help/faq/ (accessed on 1 January 2021).
42. Free Sound Effects & Royalty Free Music. Available online: www.zapsplat.com (accessed on 1 January 2021).
43. Fesliyan Studios Royalty Free Music. Available online: fesliyanstudios.com/contact (accessed on 1 January 2021).
44. What Is a Spectrogram. Available online: tomroelandts.com/articles/what-is-a-spectrogram (accessed on 1 January 2021).
45. Dennis, J.; Tran, H.; Li, H. Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions. *IEEE Signal Process. Lett.* **2011**, *18*, 130–133. [CrossRef]
46. Prasadh, S.; Natrajan, S.; Kalaivani, S. Efficiency analysis of noise reduction algorithms: Analysis of the best algorithm of noise reduction from a set of algorithms. In Proceedings of the International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 23–24 November 2017.
47. Edmonds, E. Abstraction and interaction: An art system for white noise. In Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation (CGIV), Sydney, NSW, Australia, 26–28 July 2006.
48. Dahlan, R. AdaBoost Noise Estimator for Subspace based Speech Enhancement. In Proceedings of the International Conference on Computer, Control, Informatics and its Applications (IC3INA), Tangerang, Indonesia, 1–2 November 2018.
49. Vasuki, P.; Bhavana, C.; Mohamed, S.; Lakshmi, E. Automatic noise identification in images using moments and neural network. In Proceedings of the International Conference on Machine Vision and Image Processing (MVIP), Coimbatore, India, 14–15 December 2012.
50. Montillet, J.; Tregoning, P.; McClusky, S.; Yu, K. Extracting White Noise Statistics in GPS Coordinate Time Series. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 563–567. [CrossRef]
51. How to Set up Alexa Guard on an Amazon Echo. Available online: cnbc.com/2019/05/14/how-to-set-up-alexa-guard-on-an-amazon-echo.html (accessed on 1 January 2021).
52. Nest Hub. Available online: store.google.com/us/product/google_nest_hubl (accessed on 1 January 2021).
53. Nest Mini. Available online: store.google.com/product/google_nest_mini (accessed on 1 January 2021).
54. See Sound. Available online: see-sound.com/devices/ (accessed on 1 January 2021).
55. The Best Sound Machines on Amazon, According to Hyperenthusiastic Reviewers. Available online: nymag.com/strategist/article/best-sound-machines-noise-machines.html (accessed on 1 January 2021).
56. So What Is This Secretive Chinese Sonic Weapon Exactly? Available online: popularmechanics.com/military/ (accessed on 1 January 2021).
57. U.S. Military Is Developing a Sound Weapon that Sounds Like a Retro Modem. Available online: digitaltrends.com/cool-tech/military-sound-weapon-old-modem/ (accessed on 1 January 2021).
58. Plug your Ears and Run': NYPD's Use of Sound Cannons Is Challenged in Federal Court. Available online: nbcnews.com/news/us-news/plug-your-ears-run-nypd-s-use-sound-cannons-challenged-n1008916 (accessed on 1 January 2021).

59. Using Sound to Attack: The Diverse World of Acoustic Devices. Available online: cnn.com/2017/08/10/health/acoustic-weapons-explainer/index.html (accessed on 1 January 2021).
60. A Man Stashed Guns in His Las Vegas Hotel Room. 3 Years Later, a Killer Did the Same. Available online: nytimes.com/2018/09/28/us/las-vegas-shooting-mgm-lawsuits.html (accessed on 1 January 2021).