Machine Beats Machine: Machine Learning Models to Defend Against Adversarial Attacks.

Jože M. Rožanec*
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Dimitrios Papamartzivanos
Ubitech Ltd
Chalandri, Athens, Greece
dpapamartz@ubitech.eu

Entso Veliou
Department of Informatics and Computer Engineering, University of West Attica
Athens, Greece
eveliou@uniwa.gr

Theodora Anastasiou
Ubitech Ltd
Chalandri, Athens, Greece
tanastasiou@ubitech.eu

Jelle Keizer
Philips Consumer Lifestyle BV
Drachten, The Neatherlands
jelle.keizer@philips.com

Blaž Fortuna
Qlector d.o.o.
Ljubljana, Slovenia
blaz.fortuna@qlector.com

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

We propose using a two-layered deployment of machine learning models to prevent adversarial attacks. The first layer determines whether the data was tampered, while the second layer solves a domain-specific problem. We explore three sets of features and three dataset variations to train machine learning models. Our results show clustering algorithms achieved promising results. In particular, we consider the best results were obtained by applying the DBSCAN algorithm to the structured structural similarity index measure computed between the images and a white reference image.

CCS CONCEPTS

• **Information systems → Data mining**; • **Computing methodologies → Computer vision problems**; • **Applied computing**;

KEYWORDS

Cybersecurity, Adversarial Attacks, Machine Learning, Automated Visual Inspection

ACM Reference Format:
Jože M. Rožanec, Dimitrios Papamartzivanos, Entso Veliou, Theodora Anastasiou, Jelle Keizer, Blaž Fortuna, and Dunja Mladenić. 2021. Machine Beats Machine: Machine Learning Models to Defend Against Adversarial Attacks.. In *Ljubljana '22: Slovenian KDD Conference on Data Mining and Data Warehouses, October, 2022, Ljubljana, Slovenia.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Artificial Intelligence (AI) solutions have penetrated the Industry 4.0 domain by revolutionizing the rigid production lines enabling innovative functionalities like mass customization, predictive maintenance, zero defect manufacturing, and digital twins. However, AI-fuelled manufacturing floors involve many interactions between the AI systems and other legacy Information and Communications Technology (ICT) systems, generating a new territory for malevolent actors to conquer. Hence, the threat landscape of Industry 4.0 is expanded unpredictably if we also consider the emergence of adversary tactics and techniques against AI systems and the constantly increasing number of reports of Machine Learning (ML) systems abuses based on real-world observations. In this context, Adversarial Machine Learning (AML) has become a significant concern in adopting AI technologies for critical applications, and it has already been identified as a barrier in multiple application domains. AML is a class of data manipulation techniques that cause changes in the behavior of AI algorithms while usually going unnoticed by humans. Suspicious objects misclassification in airport control systems [7], abuse of autonomous vehicles navigation systems [11], tricking of healthcare image analysis systems for classifying a benign tumor as malignant [15], abnormal robotic navigation control [23] are only a few examples of AI models' compromise that advocate the need for the investigation and development of robust defense solutions.

Recently, the evident challenges posed by AML have attracted the attention of the research community, the industry 4.0, and the manufacturing domains [20], as possible security issues on AI systems can pose a threat to systems reliability, productivity, and safety [2]. In this reality, defenders should not be just passive spectators, as there is a pressing need for robustifying AI systems to hold against the perils of adversarial attacks. New methods are needed to safeguard AI systems and sanitize the ML data pipelines from the potential injection of adversarial data samples due to poisoning and evasion attacks.

SiKDD '22, October, 2022, Ljubljana, Slovenia

We developed a machine learning model to address the abovementioned challenges, detecting whether the incoming images are adversarially altered. This enables a two-layered deployment of machine learning models that can be used to prevent adversarial attacks (see Fig. 1): (a) the first layer with models determining whether the data was tampered, and (b) a second layer that operates with regular machine learning models developed to solve particular domain-specific problems. We demonstrate our approach in a real-world use case from *Philips Consumer Lifestyle BV*. This paper explores a diverse set of features and machine learning models to detect whether the images have been tampered for malicious purposes.

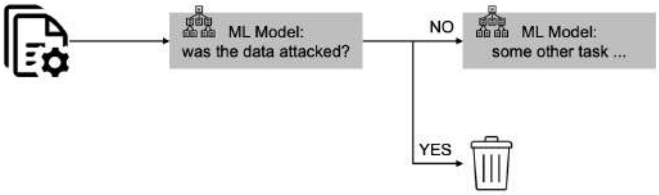

Figure 1: Two-layered deployment of machine learning models can be used to prevent adversarial attacks.

This paper is organized as follows. Section 2 outlines the current state of the art and related works, Section 3 describes the use case, and Section 4 provides a detailed description of the methodology and experiments. Finally, Section 5 outlines the results obtained, while Section 6 concludes and describes future work.

2 RELATED WORK

AML attacks are considered a severe threat to AI systems, and, that is, the research community seeks new robust defensive methods. Image classifiers, those analyzed in this work, are the focal point of the vast majority of the AML literature, as those have been proved prone to noise perturbations. According to the literature, prominent solutions focus on denoising the image classifiers, training the target model with adversarial examples, known as adversarial training, or applying standalone defense algorithms.

Yan et. al. [21] proposed a new adversarial attack called Observation-based Zero-mean Attack, and they evaluated the robustness of various deep image denoisers. They followed an adversarial training strategy and effectively removed various synthetic and adversarial noises from data. In [17], pre-processing data defenses for image denoising are evaluated, highlighting the advantages of such approaches that do not require the retraining of the classifiers, which is a computationally intense task in computer vision.

However, the robustness of adversarial training via data augmentation and distillation is advocated by the majority of the works in the domain. Specifically, Bortsova et al. [3] have focused on adversarial black-box settings, assuming that the attacker does not have full access to the target model as a more realistic scenario. They tuned their testbed to ensure minimal visual perceptibility of the attacks. The applied adversarial training dramatically decreased the performance of the designed attack. Hashemi and Mozaffari [8] trained CNNs with perturbed samples manipulated by various transformations and contaminated by different noises to foster robustness using adversarial training.

On top of the above, several standalone solutions have been proposed. CARAMEL system in [13] offered a set of detection techniques to combat security risks in automotive systems with embedded camera sensors. Hybrid approaches and more general alternatives intrinsically improve the robustness of AI models. A defensive Distillation mechanism against evasion attacks is proposed in [16] being able to reduce the effectiveness of adversarial sample creation from 95% to less than 0.5% on a studied DNN. Subset Scanning was presented in [19] to give the ability to DNNs to recognize out-of-distribution samples.

3 USE CASE

The Philips factory in Drachten, the Netherlands, is an advanced factory for mass manufacturing consumer goods (e.g., shavers, OneBlade, baby bottles, and soothers). Current production lines are often tailored for the mass production of one product or product series in the most efficient way. However, the manufacturing landscape is changing due to global shortages, manufacturing assets and components are becoming scarcer [1], and a shift in market demand requires the production of smaller batches more often. To adhere to these changes, production flexibility, re-use of assets, and a reduction of reconfiguration times are becoming more critical for the cost-efficient production of consumer goods. One of the topics currently investigated within Philips is quickly setting up automated quality inspections to make reconfiguring quality control systems faster and easier. Next to working on the technical challenges of doing this, safety and cyber-security topics are explored, aiming to implement AI-enabled automated quality systems with state-of-the-art defenses, the latter of which is the focus point discussed in this paper.

The dataset used contains images of the decorative part of a Philips shaver. This product is mass-produced and important for the visual appearance of the shavers. Next to that, the part is very close to or in direct contact with the user's skin, where any deviations in its quality could impact shaver performance or even shaver safety. The dataset contains 1.194 images classified into two classes: (a) attacked with the Projected Gradient Descent attack [5], and (b) not attacked.

4 METHODOLOGY

We framed adversarial attack detection as a classification problem. We experimented with three kinds of features: (a) image embeddings (obtained from the Average Pooling Layer of a pre-trained ResNet-18 model ([9])), (b) histograms reflecting grayscale pixel frequencies (with pixel values extending between zero and 255), and (c) structural similarity index measure (SSIM) computed against a white image. While the embeddings provide information about the image as a whole, we considered the histograms and SSIM metric could be useful given the apparent difference between the original and perturbed images. Furthermore, we computed the features across three different datasets (see Fig. 2 for sample images): (a) original set of images, (b) images cropped considering an image slice extending from top to bottom (coordinates (160, 0, 200, 369) - we name this dataset set "Cropped (v1)"), and (c) images cropped

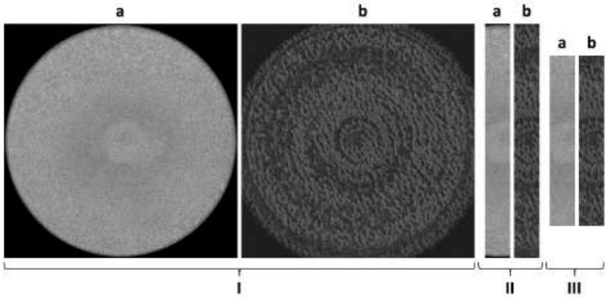

Figure 2: Three sets of images: (a) indicates the original image, while (b) indicates the images attacked with the Projected Gradient Descent attack. The subsets I, II, and III indicate (I) the whole image, (II) cropped image (v1 (considering coordinates (160, 0, 200, 369))), and cropped image (v2 - (considering coordinates (160, 50, 200, 319))).

considering a slice of the central part of the image (coordinates (160, 50, 200, 319) - - we name this dataset set "Cropped (v2)"). By comparing the original image dataset against those obtained by slicing the central part, we sought to understand if the models' predictive power increased by looking at a specific area of the image rather than the whole.

We first trained three machine learning models: Catboost [18] with Focal Loss [14] (trained over 150 iterations, and considering a tree depth of ten, while evaluating the performance during training with the logloss metric), Logistic Regression (the dataset was scaled between zero and one, considering the train set, and transformed to ensure zero mean and unit variance), and KMeans (the dataset was transformed to ensure zero mean and unit variance, and the model initiated with random initialization and seeking to generate two clusters). We evaluated our experiments with a ten-fold stratified cross-validation ([12, 22]), using one fold for testing and the rest of the folds to train the model. Furthermore, to avoid overfitting, we performed a feature selection using the mutual information to evaluate the most relevant ones and select the *top K* features, with $K = \sqrt{N}$, considering N to be equal to the number of data instances in the train set [10]. Finally, we measured our models' performance with a custom metric (DP_{AUCROC}) that summarizes the discriminative power as computed from the area under the receiver operating characteristic curve (AUC ROC, see [4]) (see Eq. 1). The metric ranges from zero (no discriminative power) to one (perfect discriminative power) and it preserves the AUC ROC desirable properties of being threshold independent and invariant to *a priori* class probabilities.

$$DP_{AUC_ROC} = 2 \cdot |(0.5 - AUCROC)| \tag{1}$$

Based on the good results obtained in the clustering setting, we decided to conduct additional experiments, running the DBSCAN algorithm [6] over all existing data. The advantage of such an algorithm is that it can estimate the clusters with no prior information regarding the number of expected clusters. Therefore, if working well, it would be useful to generalize the approach toward detecting new cyberattacks where no labeled data exists yet. We consider such a characteristic to be fundamental to production environments. For the models resulting from the three abovementioned datasets, we measured the estimated number of clusters, estimated number of noise points, homogeneity (whether the clusters contain only samples belonging to a single class), completeness (whether all the data points members of a given class are elements of the same cluster), V-measure (harmonic mean between homogeneity and completeness), adjusted Rand index (similarity between clusterings obtained by the proposed and random models), and the Silhouette Coefficient (estimates the separation distance between the resulting clusters). We ran the DBSCAN algorithm measuring the distance between clusters with the Euclidean distance, considering the maximum distance between two samples for one to be considered as in the neighborhood of the other to be 0,3. Furthermore, we considered that at least ten samples in a neighborhood were required for a point to be considered as a core point.

5 RESULTS AND ANALYSIS

Model		**Catboost**	**KMeans**	**Logistic regression**
Embeddings	**Original image**	0.0167	**1.0000**	*0.0228*
	Cropped (v1)	*0.0014*	**1.0000**	0.0003
	Cropped (v2)	0.0181	**1.0000**	*0.0213*
SSIM	**Original image**	0.0152	**1.0000**	*0.0184*
	Cropped (v1)	*0.0008*	**1.0000**	0.0004
	Cropped (v2)	0.0179	**1.0000**	*0.0195*
Histograms	**Original image**	0.0016	**1.0000**	*0.0030*
	Cropped (v1)	0.0003	**1.0000**	*0.0011*
	Cropped (v2)	0.0018	**1.0000**	*0.0031*

Table 1: Results obtained across classification experiments. We measure models' performance with Eq. 1. Best results are bolded, *second-best are italicized.*

We present the results obtained in our classification experiments in Table 1. We found the KMeans models achieved perfect discrimination in all cases, while the second-best model was the Logistic regression, which had second-best results in all but two cases. Nevertheless, the Logistic regression and the Catboost models achieved a low discriminative power, almost unable to distinguish between tampered and non-tampered images. Regarding the features, we found that the best average performance was obtained when training the models on the *Cropped (v2)* dataset, followed by those trained on the whole images.

When running the DBSCAN algorithm (see results in Table 2), we found the best results were obtained considering the SSIM measure. Furthermore, using the SSIM issued excellent results in all cases. The best ones were obtained considering the *Cropped (v1) dataset*, while the second-best was achieved with the *Cropped (v2) dataset*. Using the SSIM only, the DBSCAN algorithm was able to correctly group the instances into two groups and misclassified at most a single instance. However, the performance achieved either with embeddings or histograms was not satisfactory. When considering histogram features, the DBSCAN algorithm was not able to discriminate between instances, creating a single cluster. On the other hand, when considering embeddings, DBSCAN created three clusters that issued a bad performance, considering most of the points to be noisy. We, therefore, conclude that the only promising

	Embeddings			SSIM			Histograms		
	Original image	**Cropped (v1)**	**Cropped (v2)**	**Original image**	**Cropped (v1)**	**Cropped (v2)**	**Original image**	**Cropped (v1)**	**Cropped (v2)**
Number of clusters	3	1	1	**2**	**2**	**2**	1	1	1
Number of noise points	1010	794	887	1	**0**	1	621	603	606
Homogeneity	0.1770	0.4550	0.3170	**1.0000**	**1.0000**	**1.0000**	0.8550	0.9290	0.9150
Completeness	0.2090	0.4940	0.3860	*0.9910*	**1.0000**	*0.9910*	0.8560	0.9290	0.9150
V-measure	0.1920	0.4740	0.3480	*0.9960*	**1.0000**	*0.9960*	0.8550	0.9290	0.9150
Adjusted Rand index	0.0710	0.4350	0.2540	*0.9980*	**1.0000**	*0.9980*	0.9020	0.9600	0.9500
Silhouette coefficient	0.0750	0.4310	0.2660	0.8980	**0.9590**	*0.9070*	0.8330	0.8970	0.8800

Table 2: Results obtained across clustering experiments. Best ones are bolded, *second-best are italicized.*

results were those obtained considering the SSIM. Nevertheless, we consider further research is required to understand whether this kind of feature can be useful across a wide range of attacks and in the real-world. SSIM provides metadata describing the images. Given high-quality attacks aim to reduce the visual footprint on the images, it remains an open question to which extent can the SSIM capture weak footprints and therefore enable similar discriminative capabilities on machine learning models.

6 CONCLUSION

In this work, we explored multiple sets of features and machine learning models to determine whether an image has been tampered with for the purpose of an adversarial attack. While the difference between attacked and non-attacked images is evident to the human eye, it is not to the machine learning algorithms. We found that the Catboost and Logistic regression models could almost not discriminate between both cases. On the other hand, the clustering algorithms (KMeans and DBSCAN) had a stronger performance. While the KMeans models did so perfectly, regardless of the features, the DBSCAN model only performed well using the SSIM. We consider the strength of such a model the fact that no *a priori* information regarding the classes is required, therefore saving the annotation effort and providing greater flexibility towards future adversarial attacks. Our future research will focus on testing a wider range of cyberattacks while ensuring the attack will not be discernable to the human eye.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 program project STAR under grant agreement number H2020-956573.

REFERENCES

[1] [n.d.]. European Economic Forecast. Autumn 2021. https://economy-finance.ec.europa.eu/publications/european-economic-forecast-autumn-2021_en. Accessed: 2022-08-05.

[2] Adrien Bécue, Isabel Praça, and João Gama. 2021. Artificial intelligence, cyberthreats and Industry 4.0: Challenges and opportunities. *Artificial Intelligence Review* 54, 5 (2021), 3849–3886.

[3] Gerda Bortsova, Cristina González-Gonzalo, Suzanne C. Wetstein, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bart Liefers, Bram van Ginneken, Josien PW Pluim, and Mitko Veta. 2021. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis* 73 (2021), 102141. Publisher: Elsevier.

[4] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145 – 1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[5] Yingpeng Deng and Lina J Karam. 2020. Universal adversarial attack via enhanced projected gradient descent. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1241–1245.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *kdd*, Vol. 96. 226–231.

[7] Dan-Ioan Gota, Adela Puscasiu, Alexandra Fanca, Honoriu Valean, and Liviu Miclea. 2020. Threat objects detection in airport using machine learning. In *2020 21th International Carpathian Control Conference (ICCC)*. IEEE, 1–6.

[8] Atiyeh Hashemi and Saeed Mozaffari. 2021. CNN adversarial attack mitigation using perturbed samples training. *Multim. Tools Appl.* 80 (2021), 22077–22095.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, and Edward R Dougherty. 2005. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* 21, 8 (2005), 1509–1515.

[11] A. Kloukiniotis, A. Papandreou, A. Lalos, P. Kapsalas, D.-V. Nguyen, and K. Moustakas. 2022. Countering adversarial attacks on autonomous vehicles using denoising techniques: A Review. *IEEE Open Journal of Intelligent Transportation Systems* (2022). Publisher: IEEE.

[12] Max Kuhn, Kjell Johnson, et al. 2013. *Applied predictive modeling*. Vol. 26. Springer.

[13] Christos Kyrkou, Andreas Papachristodoulou, Andreas Kloukiniotis, Andreas Papandreou, Aris Lalos, Konstantinos Moustakas, and Theocharis Theocharides. 2020. Towards artificial-intelligence-based cybersecurity for robustifying automated driving systems against camera sensor attacks. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 476–481.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[15] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition* 110 (2021), 107332.

[16] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2015. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *CoRR* abs/1511.04508 (2015). arXiv:1511.04508 http://arxiv.org/abs/1511.04508

[17] Marek Pawlicki and Ryszard S. Choraś. 2021. Preprocessing Pipelines including Block-Matching Convolutional Neural Network for Image Denoising to Robustify Deep Reidentification against Evasion Attacks. *Entropy* 23, 10 (2021), 1304. Publisher: MDPI.

[18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).

[19] Skyler Speakman, Srihari Sridharan, Sekou Remy, Komminist Weldemariam, and Edward McFowland. 2018. Subset scanning over neural network activations. *arXiv preprint arXiv:1810.08676* (2018).

[20] Entso Veliou, Dimitrios Papamartzivanos, Sofia Anna Menesidou, Panagiotis Gouvas, and Thanassis Giannetsos. 2021. *Artificial Intelligence and Secure Manufacturing: Filling Gaps in Making Industrial Environments Safer.* Now Publishers. 30–51 pages. https://doi.org/10.1561/9781680838770.ch2

[21] Hanshu Yan, Jingfeng Zhang, Jiashi Feng, Masashi Sugiyama, and Vincent YF Tan. 2022. Towards Adversarially Robust Deep Image Denoising. *arXiv preprint arXiv:2201.04397* (2022).

[22] Xinchuan Zeng and Tony R Martinez. 2000. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 1 (2000), 1–12.

[23] Fangyi Zhang, Jürgen Leitner, Michael Milford, Ben Upcroft, and Peter Corke. 2015. Towards vision-based deep reinforcement learning for robotic motion control. *arXiv preprint arXiv:1511.03791* (2015).