XAI enhancing cyber defence against adversarial attacks in industrial applications

1stGeorgios Makridis
Department of Digital Systems
University of Piraeus
Athens, Greece
gmakridis@unipi.gr

2ndSpyros Theodoropoulos
Department of Digital Systems
University of Piraeus
Athens, Greece
sptheod@unipi.gr

3rdDimitrios Dardanis
Department of Digital Systems
University of Piraeus
Athens, Greece
ddardanis@unipi.gr

4thIoannis Makridis
Department of Digital Systems
University of Piraeus
Athens, Greece
imakridi@unipi.gr

5thMaria Margarita Separdani
Department of Maritime Studies
University of Piraeus
Athens, Greece
ritasepa8@gmail.com

6thGeorgios Fatouros
Department of Digital Systems
University of Piraeus
Athens, Greece
gfatouros@unipi.gr

7thDimosthenis Kyriazis
Department of Digital Systems
University of Piraeus
Athens, Greece
dimos@unipi.gr

7thPanagiotis Koulouris
Department of Digital Systems
University of Piraeus
Athens, Greece
pa.koulouris@gmail.com

***Abstract*—In recent years there is a surge of interest in the interpretability and explainability of AI systems, which is largely motivated by the need for ensuring the transparency and accountability of Artificial Intelligence (AI) operations, as well as by the need to minimize the cost and consequences of poor decisions. Another challenge that needs to be mentioned is the Cyber security attacks against AI infrastructures in manufacturing environments. This study examines eXplainable AI (XAI)-enhanced approaches against adversarial attacks for optimizing Cyber defense methods in manufacturing image classification tasks. The examined XAI methods were applied to an image classification task providing some insightful results regarding the utility of Local Interpretable Model-agnostic Explanations (LIME), Saliency maps, and the Gradient-weighted Class Activation Mapping (Grad-Cam) as methods to fortify a dataset against gradient evasion attacks. To this end, we "attacked" the XAI-enhanced Images and used them as input to the classifier to measure their robustness of it. Given the analyzed dataset, our research indicates that LIME-masked images are more robust to adversarial attacks. We additionally propose an Encoder-Decoder schema that timely predicts (decodes) the masked images, setting the proposed approach sufficient for a real-life problem.**

***Index Terms*—XAI, Deep Learning, Image Classification, Adversarial Attack, LIME, Grad-Cam, Saliency map**

I. Introduction

The current day and age, also known as the Digital or Information Age, is characterized by complex computing systems which generate enormous amounts of data on a daily basis. The digital transformation of industrial environments lead to the fourth industrial revolution -Industry4.0 [1], with Artificial Intelligence (AI) being the key facilitator of the Industry4.0 era by enabling innovative tools and processes including predictive quality management (Quality4.0) [2]. Despite that fact, the Industrial sector's state-of-the-art AI development is still far from utilizing the most sophisticated machine and Deep Learning (DL) competencies [1]. Lots of challenges have arisen regarding the implementation of AI approaches in real-life tasks. The Defense against poisoning attacks along with the Explainability and Interpretability of AI algorithms can be considered essential parts for 'triggering' the wide use of such methods in production environments. A challenge that may hamper the adoption of AI methods in the manufacturing sector is the Cyber security attacks against AI infrastructures in manufacturing environments. Adversarial attacks against AI systems in manufacturing can compromise the data used for training AI systems or even disclose the rules of AI operations. These attacks can therefore lead to Intellectual Property (IP) theft, while at the same time compromising the proper operation of AI systems which could eliminate their benefits and introduce risks in the production processes. Beyond Cyber security attacks, data unreliability can be caused by other factors such as ultra-high temperatures, data transfer errors, interference, and more. Unreliable data represent one of the main challenges for the graceful operation of AI systems and as such could lead to biased AI applications [3].

To address the Cyber security aspect, one could pursue strategies for strengthening the pretrained classifier against poisoning attacks as well as detecting adversarial samples that could harm the model. Methods such as Defensive Distillation [4], Gradient Obfuscation [5], Feature Squeezing [6] and more, are used to derive robust models including poisoning training.

The significance of this component against poisoning attacks was emphasized in [7], that states that even with a strong "defense", a poisoning of a 3% of the total training dataset can lead to a drop of up to 11% in accuracy.

This research proposes a novel method for making a model

more robust against adversarial attacks. We evaluated 3 different types of XAI techniques (LIME, Saliency Maps, and) enhancing DL approaches for image classification in an industrial use case against various levels of graded evasion attacks. The proposed approach of this research consists of an 'encoder-decoder' architecture for decoding an image in a novel masked/perturbed image that depicts the most predictive part of the image that is less vulnerable to adversarial variations.

The remainder of the paper is structured as follows: Section II presents a brief background of the utilized methods and approaches, while Section III depicts a short literature review focusing on XAI in adversarial attack scenarios. Section IV includes the proposed approach, describes the datasets used, and how these are leveraged in the proposed architecture. Section V dives deeper into the results of the conducted research and the implemented methods/techniques, with their performance being depicted in both plots and matrices formats. Section VI makes a short conclusion emphasizing the recommendations for future research.

II. Background

In order to make the motivation behind our research more clear we introduce the basic concepts of Image Classification, the XAI methods, and adversarial attacks.

A. eXplainable AI (XAI)

The notion of explaining and expressing a Machine Learning (ML) model is called interpretability or explainability [8]. This need for interpretability mainly exists in Deep Neural Networks (DNN) which are defined by large levels of complexity, thus appearing to be "black boxes" [9]. There has been a growing field of research addressing the problem of ML "black boxes" as not having clarity or insight also known as eXplainable Artificial Intelligence (XAI) [10].

Industrial applications combine various AI solutions. To this end, the XAI domain plays a very sensitive but important role in industrial applications, as it serves as a bridge between complex DL models and non-IT experts. To that end, the outcomes of the XAI method must be precise and easy to understand by domain experts, in order to increase the notion of "trust" in a real-time industrial environment. Although during the last couple of years several XAI methodologies, strategies, and frameworks have been presented, for the purposes of this research which focuses on industrial applications we will classify XAI methods according to their simplicity, the extent of interpretability, and percentage of dependencies of the analyzed ML/AI model Fig. 1.

Furthermore, complexity-related methods can be further distinguished to i) intrinsically explainable (Ante-Hoc) models, which are also referred to as transparent or glass box approaches and ii) forecasting black-box (Post-hoc) models that require an understanding of the prediction's reasoning steps concerning the explainability source. Apart from that we can categorize each method based on the scope: i) global explainability methods examine the algorithm as a whole, including the training data used, and appropriate uses of the algorithms while ii) Local explainability refers to the ability of the system to tell a user why a particular decision was made.

Last but not least, in terms of the various explainability approaches one needs to explore and identify the differences between model-specific and model-agnostic solutions. The main distinction between the two is whether the XAI approach is dependent on the underlying ML model or whether it could be applied to others as well.

A more comprehensive overview of various XAI methods and scenarios, as well as evaluation metrics applied in the manufacturing domain, was conducted in [11].

B. Adversarial Attack

A lot of research conducted within the last years highlighted [12] the vulnerability of DNN models from adversarial attacks. Especially in the domain of image classification, malicious inputs were used to purposefully create synthetic pictures that, while almost identical to the actual photos, can trick the classifier into producing inaccurate prediction outputs as noted by [13].

Solutions which aim to minimize the impact of adversarial examples affecting ML/AI output can be classified in the following groups: (a) Gradient Masking; (b) Robust Optimization; (c) Adversary Detection.

- Gradient Masking: as [14] and [15] mentioned Gradient Masking approaches aim to obfuscate the gradient information of the ML/AI algorithm in order to mislead potential malevolent actions.
- Robust Optimization: [16], [17] highlight methods and techniques whose purpose is to introduce robust classifiers that will not be highly affected in the presence of adversarial examples.
- Adversary Detection: based on [18] the main purpose of Adversary Detection methods is to serve as a buffer between the input space and the ML/AI model. To that end the main focus of such approaches is to correctly identify malicious cases and not allow them to reach the classifier.

III. XAI in Adversarial Attack Scenarios

In this research, a lot of focus has been placed on creating successful and self-resilient XAI algorithms against poisoning attacks [19]. Many researchers have attempted to fortify the XAI model from poisoning attacks (e.g., [20] exposed the vulnerabilities of state-of-the-art Saliency-map-based systems by manipulating the system's adversarial model). Similarity Difference And Uniqueness method (SIDU) [21] was designed as an XAI algorithm, able to provide visual explanations, with an ability to identify local and overall regions of objects that mostly affect the prediction outcome of the model. Furthermore, [21] SIDU tends to be even more robust in the existence of adversarial attacks in comparison with "black-box" models as its performance can be better explained and understood by domain experts [21]. The performance of SIDU is compared to in [22]. When compared to the fixation maps, the results show that outperforms SIDU; however, when the algorithms are compared to noisy inputs, the results switch, indicating that SIDU is more solid to adversarial attacks. As [23] Dombrowski et al. highlighted, explanation maps are vulnerable to adversarial attacks and manipulations that aim to corrupt the input data. By simplifying the explanation process, [23] Dombrowski et al. were also capable of strengthening the system's robustness of such attacks.

Another innovative detection approach designed for the inner structures/layers of DNN classifiers, [24] is the distinction between normal and adversarial inputs by using Shapley Additive Explanations (SHAP) values. The majority of the approaches treated the identification of adversarial data as an anomaly detection task, in both the input space as well as the internal activation layer and model architecture, notwithstanding the fact that much research has also been conducted with an approach to increase detector performance by manipulating the inputs [6] or changing the training process [25]. According to [26], the presence of adversarial cases is an inherent aspect of the underlying dataset. To that end, one needs to carefully identify those features that tend to be robust and those that are not. Furthermore, [26] Ilyas et al. also highlight that the existence of adversarial examples in the input space can be transferred/affect more than a single classification model. Consequently, in the presence of features that are non-robust, the predictive values

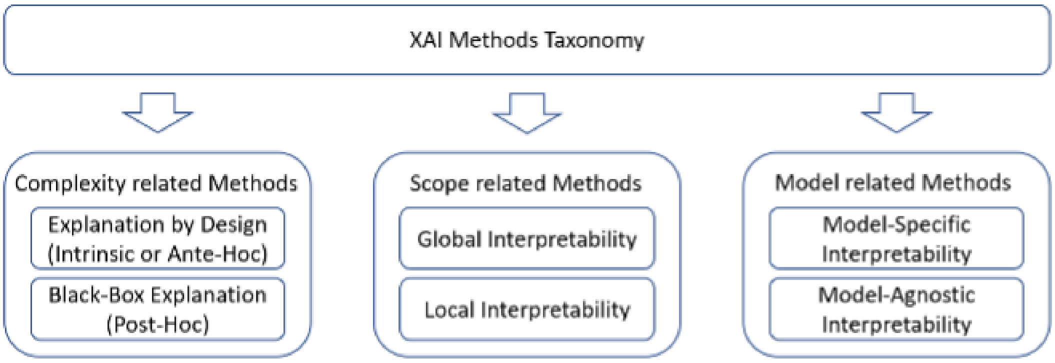

XAI Methods Taxonomy
Complexity related Methods
Explanation by Design (Intrinsic or Ante-Hoc)
Black-Box Explanation (Post-Hoc)
Scope related Methods
Global Interpretability
Local Interpretability
Model related Methods
Model-Specific Interpretability
Model-Agnostic Interpretability

Fig. 1: Taxonomy of XAI Methods

cannot be trusted without further evaluation, since even minor changes in the input space could have a direct impact on the output. As a result, adversarial evasion attacks focus on modifying the non-robust characteristics of the input space whilst substantially maintaining the values of robust features [26]. This happens since implementing an effective alteration to robust characteristics necessitates major input changes.

The novelty of this paper is the application and evaluation of three different XAI methods that enhance the robustness against adversarial attacks regarding a real use case and dataset from the manufacturing sector. Since the underlying dataset is unbalanced and consists of defect detection data, small differences between the input images need to be detected by the proposed model. Considering that the intent of adversarial examples is to (slightly) modify the input space in order to affect the classification output, the correct operation and analysis is a very challenging task.

Moreover, this study aims to propose a framework that leverages XAI techniques to enhance DL approaches for image classification in industrial use cases. The framework consists of an 'encoder-decoder' architecture for decoding an image in a novel masked/perturbed image that depicts the most predictive part that is less vulnerable to adversarial variations.

IV. Method

A. Proposed Approach (XAI against adversarial attacks)

Most Machine Learning approaches were designed to address domain-specific problems in which the same statistical distribution was used to generate both the training and test sets. Adversaries may provide data that contradicts that statistical assumption when such models are used in reality. This data could be adjusted to exploit specific flaws and manipulate the results. In order to provide a more robust Image Classification against data poisoning attacks, we leveraged XAI methods (i.e., LIME and Saliency maps) that may enhance an already trained DNN model for image classification.

B. Dataset

The dataset analyzed in this research was collected and provided by Philips. A large portion of (potential) nonconformances within the Philips factory are related to the visual appearance of parts and products. Due to the complexity and costs of the (partial) automation, most of the visual quality inspections within the Philips factory in Drachten are still performed manually. This dataset has been collected to encourage the exploration of solutions for (partial) automation of visual quality systems based on AI that can be trained using small and incomplete datasets. The key role in developing flexible automated visual quality inspections has the explainability of the AI models and the defense against adversarial attacks.

The dataset which contains categorized images of shaver shell prints is used to train the AI-based systems. A short description of the dataset is presented in Table I , with three samples of the dataset are shown in Fig. 2. The label (i.e., the dependent variable) can take 3 values which are good, double print, and, interrupted print. Specifically, the dataset is quite imbalanced as the samples are distributed in each category as follows, while the interrupted ones are very similar to the good ones:

- good: 76%
- double print: 6%
- interrupted print: 18%

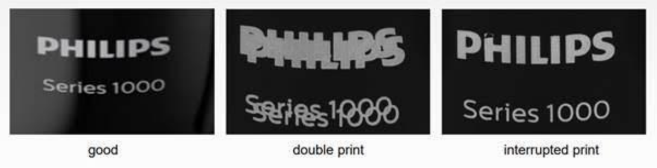


Fig. 2: One sample per class of the dataset

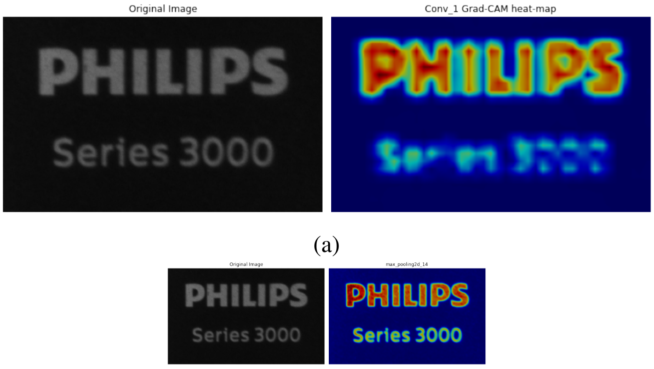
C. Pre-processing

The processing, analysis, and classification of image data is not a straightforward task and comes with a series of challenges. Data complexity, imprecision, and deficiency are some of the most well-known problems when dealing with Image classification tasks [27]. To that end, pre-processing mechanisms are of paramount importance to either increase the accuracy or decrease the complexity of a DL model. Although, various solutions to pre-process image data have been proposed in the literature [28], for the purposes of this research image normalization and grayscale conversion were preferred.

1) Although Grayscaling can be perceived as a simple task since it just removes the color from a given image consequently making it black and white, it could potentially affect the computational complexity of an ML model. Using greyscale, one could remove a series of pixels that are not required for the domain-specific task, thus reducing the sparsity and complexity of the given input space to an ML solution.

TABLE I: Dataset description

Name	**Shaver images for defect detection**
Inputs	- Small training data set containing categorized images of shaver shell prints (1 product) - Small training data for setting up automated quality inspection of a variety of products
Output	Candidate unlabelled images for which we request a human annotation.
Dataset description	Categorized images of quality inspection of the shaver shell prints. The dataset is imbalanced.
Dataset generation	Vision system implemented in pad-printing cell of shaver shells.
Categories	a) Good b) Double print c) Interrupted print
Total amount of data instances	3564
Type of Images	Color
Dimensions per Image	360 x 220, 96 DPI
Total size of Dataset	350 MB



(a)

(b)

Fig. 3: GradCam examples in 2 layers of the DNN

2) **Normalization or "re-scaling" is a well-known technique that projects pixels from a given image to a predefined range of values. For simplicity, one could consider the normalization range to span from 0 to 1 or even -1 to 1, although this is not a standard metric and can be customized. Therefore, all images that would be analyzed by an ML/AI solution will have the same influence on the model since the potential loss of pixel rate has been minimized. To that end, the learning rate of the ML/AI model can be standardized across the input space.**

D. XAI approaches

The Gradient-weighted Class Activation Mapping (Grad-Cam) model [29] was used as one of the State-of-the-ART methods for interpreting the top features (parts of the image) concerning the ''Label". To identify the abstracted notion of a given image, Grad-Cam utilizes gradients to generate localization maps and highlight essential parts of the image. Despite the high level of complexity, Grad-Cam is able to provide intuitive outputs, thus improving the model's accuracy and flexibility. An example of applying the Grad-Cam method is depicted in Fig. 3.

The Local Interpretable Model-agnostic Explanations LIME model is based on the work of [30] Ribeiro et al. and its focal point is to identify and understand the behavior of "black-box" classifiers. Using LIME, the steps that were followed to extract the ''important" parts of an image are the following and also depicted in Algorithm 1:

- **Define the number of superpixels based on the complexity of an image.**
- **Spawn similar samples to the input image and conceal the previously defined superpixels, thus generating "perturbations".**
- **Use the trained AI/ML model for image classification of the perturbed sample.**
- **Evaluation of sample importance. To measure the influence/importance of the given perturbed sample, one calculates the cosine distance of the sample in relation to the original image by using a kernel function. The smaller the distance between the two, the higher the influence of the perturbed sample.**
- **Fit a linear regression model to the most important perturbed samples that were identified in the previous step, to capture the fitted coefficients of the feature space. Features with the largest coefficients are the ones that are of most interest since they affect the "decision-making" process of a given "black-box" model the most.**

When dealing with problems that require elaborated Deep Learning solutions the level of complexity might rise in an unprecedented manner, thus making the decision-making process of CNN "obscure" even to experts. Saliency maps assist domain experts to achieve greater insights into the "inner" decision-making processes of an ML/AI solution, even at each convolutional layer when dealing with complex CNNs. The notion of Saliency maps was first introduced by [31] Simonyan and Zisserman as a technique to enhance human cognition when dealing with complex classification tasks in ML/AI systems. Saliency maps are designed to capture human attention in specific regions of the generated image, by identifying "special" features/pixels which are highlighted based on their underlying importance. To that end, human experts could have a better understanding of the classification output of a given ML/AI solution, thus making the human-computer interaction a straightforward task.

V. Results

A. Preliminary Findings

Initially, we trained a CNN model for image classification. Following the standard evaluation scheme, 70-30 for train-test split the results of the evaluation on the test set of images are depicted in Table II where we can notice that this model achieves an accuracy of 97% in this 3-class classification task when normal data are used. This model will be the AI model that will be used to test its robustness against adversarial attacks.

To this end we used adversarial data created with the Gradient-based evasion attack [32] method and, evaluated this model Loss and accuracy on the test set. That means that the model was trained on normal data (train and validation set) and then we poisoned the test set using epsilon from 0 to 0.2 and measure the degradation of the accuracy. These are the results in Table III under the label ''Normal Images". As we can see, the accuracy falls heavily after the value of epsilon of 0.02.

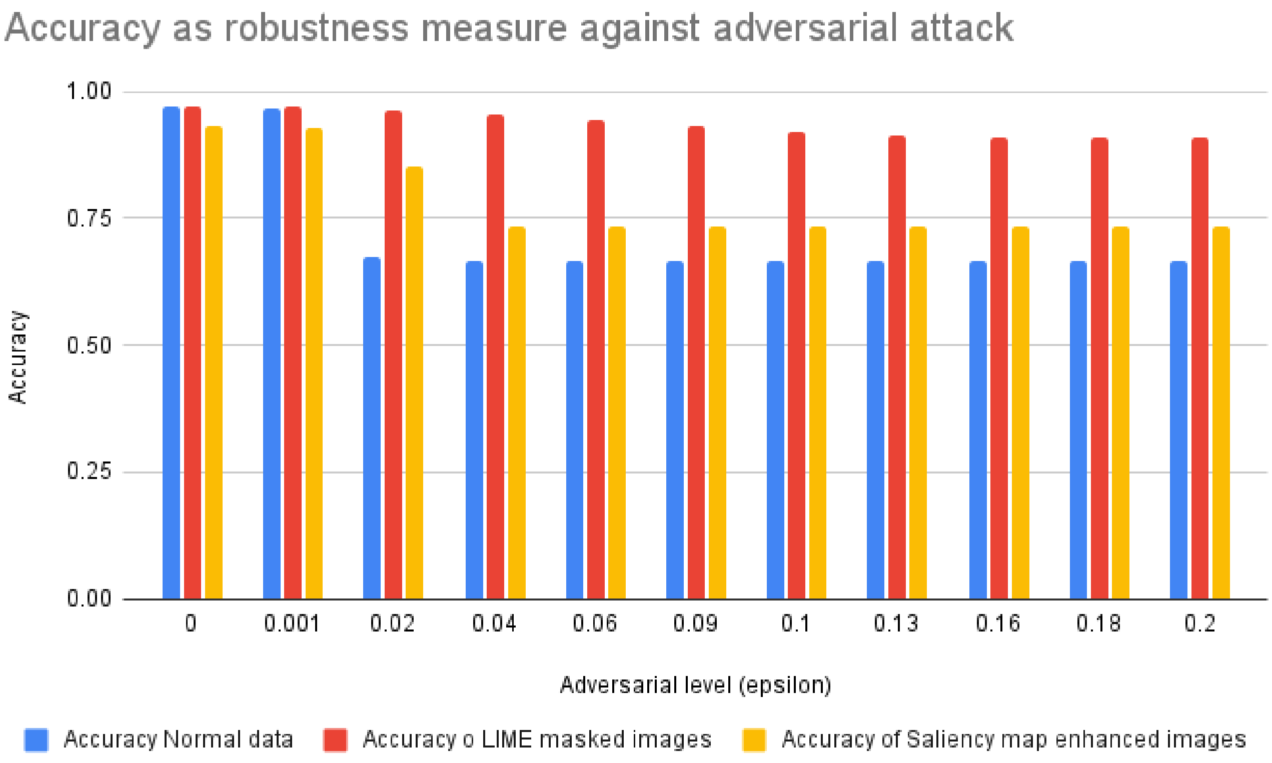


Fig. 4: Barplot showing the accuracy of the CNN image classifier against adversarial attacks of various levels (x-axis) on different types of input data

Results of the image classification task with original dataset

TABLE II

	precision	recall	f1-score	support
double print	0.98	1	0.99	46
good	0.97	0.99	0.98	540
interrupted print	0.95	0.88	0.91	126
accuracy		0.97		712
macro avg	0.97	0.96	0.96	712
weighted avg	0.97	0.97	0.97	712

B. *Investigation with LIME*

Fig. 6 shows 3 examples of the process described regarding the explainability of the LIME model. We can see 3 examples of the initial dataset and how they are masked by the LIME based on the most ”important” super-pixels of the images for their category. These examples are labeled as perturbed. Then we can notice the images that are labeled as reconstructed. These are the results of an encoder-decoder DNN that is trained to generate the results of the LIME model. The inspiration behind the usage of this encoder-decoder is depicted in Fig. 5.

Afterwards, we performed a poisoning attack on the reconstructed images again with the same levels of poisoning and the results are presented in Table III under the label ”Pertrubed Images”. We notice that these images are almost as efficient as the initial ones when no adversarial data are present and more robust when in the presence of an attack. For the purposes of this particular study, the correct classification of poisoned samples is of paramount importance, and therefore this approach performs up to standard for the poisoned/malicious test set (as it has an accuracy of 90% when epsilon=0.2). Nevertheless, the original set of images drops to 66% when epsilon = 0.04, which lead us to conclude that the LIME mask approach is effective against Gradient-based evasion attack.

C. *Investigation with Grad-Cam*

Fig. 7 shows 3 examples of the process described regarding the explainability of the model. We can see 3 examples of the initial dataset and how they are masked. Afterward, we applied the Image Classifier model to the enhanced images. The results of the classification on the test set are presented in Table IV. We notice that the model does not perform well when the inputs are enhanced with highlights. So our initial hypothesis is rejected.

To that end, we applied these images as supplementary information. Specifically, we trained a Siamese type of DNN where the 2 similar networks were identical to the initial one. The results are not encouraging as the overall model did not achieve accuracy over 50%. That means that the image operates as noise to the original image.

D. *Investigation with Saliency Map*

Fig. 8 shows 3 examples of the process described regarding the explainability with the Saliency map visualization of important parts of an image explaining why the model classified each image. We see 3 examples of the initial dataset and how they are masked by the Saliency method. Specifically, one can notice the saliency map, the original image, and also the superimposition of them.

TABLE III: Results of difference types of enjoined images against various levels of adversarial attack

	Pertubed Images		Normal Images		Sialency Images	
Adversarial - epsilon	**Loss**	**Accuracy**	**Loss**	**Accuracy**	**Loss**	**Accuracy**
0	0.874	0.9677	0.1098	0.9705	0.448	0.9312
0.001	0.8741	0.9677	0.1283	0.9663	0.4532	0.9284
0.02	0.8838	0.9621	6.1446	0.6742	0.7465	0.8525
0.04	0.9125	0.9551	12.6796	0.6643	1.409	0.7331
0.06	0.9791	0.9424	18.6595	0.6643	2.0485	0.7331
0.09	1.0812	0.9298	24.2798	0.6643	2.64	0.7331
0.1	1.2043	0.9185	29.793	0.6643	3.2418	0.7331
0.13	1.3412	0.9143	35.163	0.6643	3.9503	0.7331
0.16	1.5038	0.9101	40.0465	0.6643	4.8538	0.7331
0.18	1.7087	0.9087	44.1805	0.6643	5.9688	0.7331
0.2	1.9866	0.9073	47.5323	0.6643	7.2715	0.7331

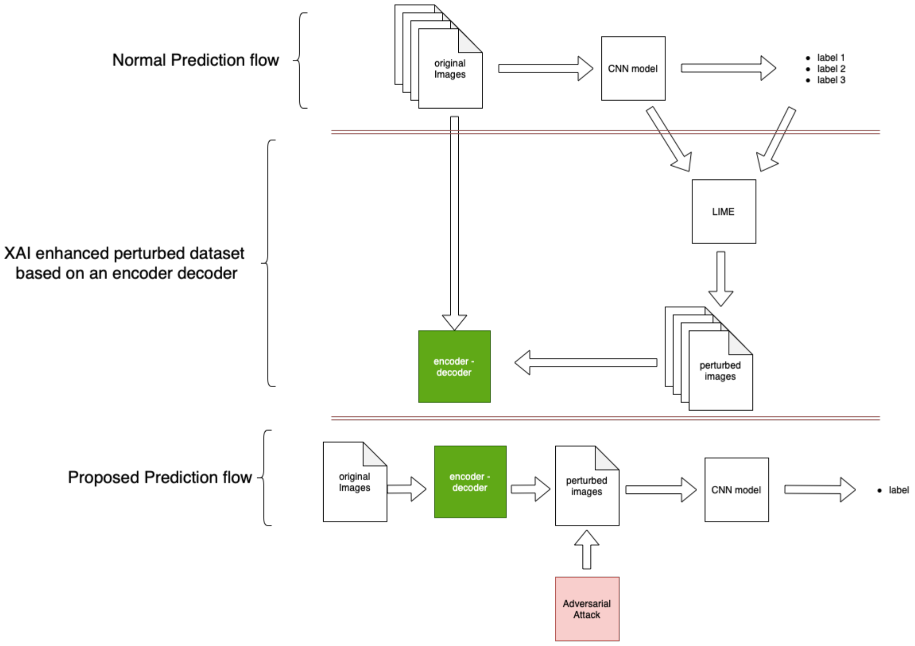


Fig. 5: How LIME can be used in the prediction phase to make a model more robust against adversarial attacks.

TABLE IV: Results of the image classification task with Grad-Cam masked dataset

	precision	**recall**	**f1-score**	**support**
double print	0.05	0.52	0.09	46
good	0	0	0	540
interrupted print	0.52	0.94	0.67	126
accuracy		0.2		712
macro avg	0.19	0.49	0.25	712
weighted avg	0.09	0.2	0.12	712

Afterwards, we applied the Image Classifier model to the Saliency-enhanced images. The results of the classification on the test set are shown in Table V.

Furthermore, we applied poisoning attacks to the saliency-enhanced image, with the same levels of poisoning. The results of this are in Table III under the label "Saliency Images". We can notice that these images are almost as efficient as the initial ones when no adversarial data are present and more robust when there is an attack. In this study, it is important to accurately classify the poisoned samples so this approach seems to perform better for small epsilon, while the performance drops significantly for epsilon > 0.04. It still performs better than the original images but worse than the LIME-masked images.

VI. Conclusion

This paper examines an XAI-enhanced DNN for addressing the problem of cyber defense against adversarial attacks for manufacturing image classification tasks. LIME, Salience maps, and Grad-Cam techniques are used to find the most informative parts of each image with respect to enhancing a

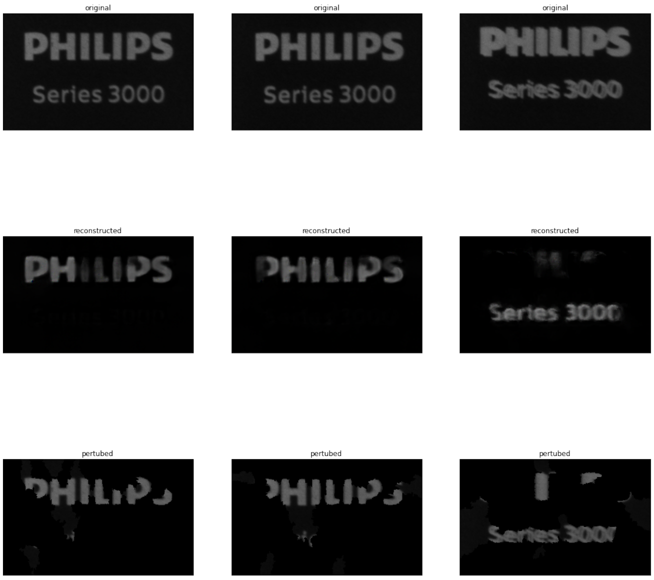


Fig. 6: Example of LIME important superpixel

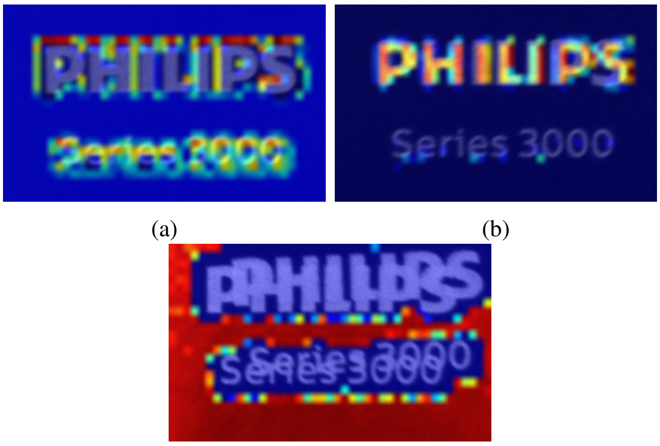

(a)
(b)
(c)

Fig. 7: Grad-Cam of 3 image examples, the first image is a good print, the second the interrupted one, and the third a double print one.

trained image classification DL model offering higher explainability and tolerance against adversarial data attacks. These experiments highlighted that masked/perturbed images (i.e., most important super-pixels) are the most tolerant XAI-based transformation/enhancement to a custom CNN image classifier against Gradient-based evasion attack. The outcomes of our experiments show that LIME in reconstruction is better than the other methods as occlusions make the classifier learn more robust features and not be affected by small perturbations. Our defenses rely on the insight that LIME-masked images can serve as a noise reduction process, helping reduce the magnitude of adversarial perturbations.

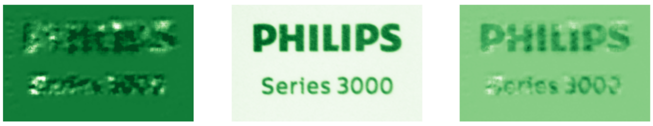


Fig. 8: Saliency Map explanation method of 1 image example, the first image is the salience map, the second the original image, and the third the combination of the original image with saliency map.

TABLE V: Results of the image classification task with saliency map enhanced dataset

	precision	recall	f1-score	support
double print	0.61	1.00	0.76	46
good	0.97	0.99	0.98	540
interrupted print	0.93	0.66	0.77	126
accuracy		0.93		712
macro avg	0.84	0.88	0.84	712
weighted avg	0.94	0.93	0.93	712

As future work we plan to develop a combination of XAI methods creating a super-XAI-enhanced image inspired by fusion models. The enhanced images that will be produced by this model will be tested against more adversarial and evasion types of attacks and will be evaluated accordingly. Then a global encoder-decoder model will be developed based on a pre-trained model such as the Resnet.

Acknowledgements: The research included in this paper has received funding from the Europeans Union's funded Project STAR under grant agreement no 956573.

REFERENCES

[1] G. Makridis, D. Kyriazis, and S. Plitsos, "Predictive maintenance leveraging machine learning for time-series forecasting in the maritime industry," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[2] Y. Bai, Z. Sun, J. Deng, L. Li, J. Long, and C. Li, "Manufacturing quality prediction using intelligent learning approaches: A comparative study," *Sustainability*, vol. 10, no. 1, p. 85, 2018.

[3] C. González-Gonzalo, E. F. Thee, C. C. Klaver, A. Y. Lee, R. O. Schlingemann, A. Tufail, F. Verbraak, and C. I. Sánchez, "Trustworthy ai: Closing the gap between development and integration of ai systems in ophthalmic practice," *Progress in Retinal and Eye Research*, p. 101034, 2021.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[5] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[6] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[7] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," *Advances in neural information processing systems*, vol. 30, 2017.

[8] J. Choo and S. Liu, "Visual analytics for explainable deep learning," *IEEE computer graphics and applications*, vol. 38, no. 4, pp. 84–92, 2018.

[9] T. Zahavy, N. Ben-Zrihem, and S. Mannor, "Graying the black box: Understanding dqns," in *International conference on machine learning*. PMLR, 2016, pp. 1899–1908.

[10] D. Gunning, "Explainable artificial intelligence (xai) darpa-baa-16-53," *Defense Advanced Research Projects Agency*, 2016.

[11] G. Sofianidis, J. M. Rožanec, D. Mladenić, and D. Kyriazis, "A review of explainable artificial intelligence in manufacturing," *arXiv preprint arXiv:2107.02295*, 2021.

[12] T. He and J. Glass, "Detecting egregious responses in neural sequence-to-sequence models," *arXiv preprint arXiv:1809.04113*, 2018.

[13] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, “Adversarial attacks and defenses in images, graphs and text: A review,” *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[14] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[15] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *International conference on machine learning*. PMLR, 2018, pp. 274–283.

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

[17] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” *arXiv preprint arXiv:1611.01236*, 2016.

[18] N. Carlini and D. Wagner, “Adversarial examples are not easily detected: Bypassing ten detection methods,” in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 3–14.

[19] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[20] J. Heo, S. Joo, and T. Moon, “Fooling neural network interpretations via adversarial model manipulation,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] S. M. Muddamsetty, N. J. Mohammad, and T. B. Moeslund, “Sidu: similarity difference and uniqueness method for explainable ai,” in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3269–3273.

[22] L. Fenoy and A. Ciontos, “Performance evaluation of explainable ai methods against adversarial noise.”

[23] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, “Explanations can be manipulated and geometry is to blame,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] G. Fidel, R. Bitton, and A. Shabtai, “When explainability meets adversarial learning: Detecting adversarial examples using shap signatures,” in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[25] K. Roth, Y. Kilcher, and T. Hofmann, “The odds are odd: A statistical test for detecting adversarial examples,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 5498–5507.

[26] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” *Advances in neural information processing systems*, vol. 32, 2019.

[27] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. Buydens, “Breaking with trends in preprocessing?” *TrAC Trends in Analytical Chemistry*, vol. 50, pp. 96–106, 2013.

[28] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, “” why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[31] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.

[32] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.