Explanation-Driven Adversarial Attacks against Multimedia Edge Applications

Theodora Anastasiou
UBITECH LTD
Limassol, Cyprus
tanastasiou@ubitech.eu

Ioannis Pastellas
UBITECH LTD
Limassol, Cyprus
ipastellas@ubitech.eu

Sophia Karagiorgou
UBITECH LTD
Limassol, Cyprus
skaragiorgou@ubitech.eu

Mariza Konidi
Dept. of Electronic Engineering
Hellenic Mediterranean University
ddk184@edu.hmu.gr

***Abstract*—Generative Artificial Intelligence (GenAI) and automated attacks on AI models raise concerns about the credibility and reliability of AI systems. Specifically, for civilian applications, there is an increasing need for AI systems to be robust, transparent, and interpretable to earn the trust of decision-makers. To effectively address the challenge of learning with enhanced explanations in constrained edge applications under conditions of massive adversarial attacks, we benchmark and present an Explanation-Driven Adversarial approach. By harvesting multimedia data collected from drones at the edge and augmenting it with diverse and massive adversarial examples, we transparently train explainers and robustify Neural Networks (NNs) to improve the AI model's fidelity. The new and robust AI model performs exceptionally well against novel and unseen attack types and concept drifts. Finally, through benchmarking across diverse adversarial attacks, we extend research in sensitive application domains and promote the adoption of more responsible and informed AI integration.**

***Index Terms*—adversarial explanations, adversarial attack benchmarking, robust and trustworthy explainability, AI model interpretability with confidence scores**

I. Introduction

In civilian applications, citizen protection involves a wide range of actors, including individuals, Law Enforcement Agencies (LEAs), first responders, and governmental agencies. Recent advancements in Artificial Intelligence (AI) have driven innovations across various traditional AI applications, including edge AI and computer vision. To enhance the safety of people and the environment, numerous efforts are being made to take advantage of the progress of these technologies to increase situational awareness in civilian applications. Key functions of these applications include tracking and surveillance, object detection, reconnaissance, intelligence analysis, and training. However, despite the potential of AI in civilian applications, several challenges must be addressed. For instance: (i) civilian AI systems must be transparent and interpretable to prevail on decision-makers and enable real-time risk analysis of the landscape; this is challenging because many AI techniques function as black boxes, preserving the internals of the decision paths that produce the results and lacking in transparency and interpretability; (ii) civilian AI systems must be robust and reliable; this presents difficulties, as research has shown that AI methods may be vulnerable to imperceptible input data manipulations and concept drifts; (iii) distributed and multimodal big data originate from various, albeit low-cost, Internet of Things (IoT) and sensor-equipped devices (e.g., drones equipped with cameras, accelerometers, telemetry sensors, etc.). Since many AI techniques rely on Machine Learning (ML), which requires a vast amount of training data, a challenge arises from the frequent lack of accurate and sufficient data in civilian and environmental applications. Furthermore, these data streams necessitate different processing modalities (e.g., online vs. offline; federated vs. centralized), as well as alignment, cleaning, and fusion, before they feed AI systems. In this work, we focus on the implementation of Neural Networks (NNs) in resource-constrained edge environments, processing online multimedia streams to address adversarial explanations' challenges and ensure reliable real-time conclusions for decision-makers and citizens.

II. Literature Review

This study builds upon our earlier work [1], which is one of several studies that explore novel applications of Adversarial Explainability (AE). Adversarial and explainable AI (XAI) combine the benefits of explanatory analysis and adversarial robustness at the same time. Given the broad spectrum of research in this area, using a variety of AI algorithms is crucial to mitigate risks such as adversarial attacks on AI models, low data quality, or malicious users. In this work, we focus on improving the robustness of NNs in resource-constrained multimedia applications on the edge. To achieve this, we orchestrate a strategic deployment of a multitude of adversarial attacks on AI models, accompanied by dynamic data augmentation infused with diverse concept drifts. This approach is designed to rigorously assess and amplify the models' robustness. Our findings culminate in the delivery of deeply insightful explanations, reinforced by confidence scores that resonate with clarity and trust in AI models.

Gao et al. [2] investigate how NNs resist adversarial attacks in classification tasks. They demonstrate the adversarial fragility of neural network-based classifiers through matrix-theoretic derivations and specific classification cases. This fragility stems from the fact that neural networks often rely on compressed features to complete classification tasks. Therefore, in order to alter the limited feature subsets that neural networks use, perturbations must be included via adversarial attacks. According to their theoretical findings, when the input dimensionality increases, the adversarial robustness of the neural network may decrease.

Mei et al. [3] propose a new benchmark for evaluating model robustness against various noises, including natural distortions and adversarial attacks, to investigate how deep learning models handle challenges in remote sensing image classification. Their study experiments with several deep learning models and develops publicly available datasets with different levels of noise that can be used in future research. Their findings provide critical insights into building more robust and reliable deep learning systems for remote sensing applications.

An intrusion detection system integrating Generative Adversarial Networks (GANs), Multi-Scale Convolutional Neural Networks (MSCNNs), and Bidirectional Long Short-Term Memory (BiLSTM) networks is presented by Benchama et al. [4]. For interpretability, they employ Local Interpretable Model-Agnostic Explanations (LIME). Their system detects intrusions by synthesizing network traffic data that includes both typical and malicious patterns. The elucidation of the model's decisions is facilitated through the integration of LIME.

Card et al. [5] investigate the adversarial weaknesses of an NN-based malware classification system in various online and dynamic analysis settings. They use the state-of-the-art technique, SHapley Additive exPlanations (SHAP), to inform adversarial attackers about features critical to classification decisions. A Feed Forward Neural Network (FFNN) is trained for malware categorization. Their results demonstrate a notable occurrence of evasion in specific attack scenarios, highlighting the apparent susceptibility of malware classifiers to adversarial manipulation.

Wickstrøm et al. [6] describe how selected hyperparameters in XAI evaluations can be manipulated. Since there is no fixed set of "correct" explanations, such choices can significantly impact evaluation outcomes. They identify two (2) types of manipulations: one that makes a method seem better, and another that changes comparisons between methods. Their experiments show that even minor changes in settings can profoundly affect evaluation results. To mitigate this issue, they develop a ranking methodology that reduces manipulation risks that would make the ranking in XAI fairer and more reliable.

Baniecki and Biecek [7] investigate how adversaries can manipulate explanation methods in AI systems, potentially leading to misleading interpretations. While they provide a unified framework and taxonomy to classify adversarial attacks, they discuss very few defense strategies to increase the robustness of explanation methods.

Wang et al. [8] explored different methods that use gradients to explain how NNs make decisions. They classified these methods into four (4) groups and explained how they have been enhanced over time. Grad-CAM technique is discussed, which highlights the most important areas of an image that influence a model's decision by using gradient information from the last convolutional layer. This also looks at ways to evaluate these explanation methods, including how well humans can understand them and how accurate they are, while the authors also discuss the key challenges in making AI models more explainable using gradient-based techniques.

Large-scale practical AI implementation requires the capacity to not only achieve high performance but also to understand and defend its predictions, especially in complex multimedia edge applications. Our work demonstrates that explanation-driven adversarial attacks offer a powerful framework for these scenarios, merging the strengths of XAI and adversarial robustness to foster trust between AI systems and human operators.

By providing clear and interpretable adversarial explanations, our approach bridges the gap between the intricate inner workings of sophisticated algorithms and the practical knowledge of end users. This transparency in decision-making is critical for the safe and widespread adoption of AI across fields such as environmental monitoring, situational awareness, and other civilian applications requiring secure assurances.

Breaking new ground beyond the limitations of past studies that focus solely on either explainability or adversarial robustness, our innovative method weaves together a spectrum of attack strategies. This integration generates meaningful cause-and-effect insights, secured by objective confidence scores that are as solid as they are informative. Our approach accelerates the process of uncovering the fundamental causes within object segmentation and detection tasks while significantly enhancing overall system reliability and trustworthiness, setting a new basis for trust in AI.

III. Dataset Overview

A. Fire Images

A varied collection of fire images, depicting different fire incidents, is used to increase situational awareness of vital infrastructures and environmental protection. To provide a wide range of fire types, sizes, and contextual scenarios, these images were obtained from several datasets. They capture different environmental conditions and fire cases from different angles and resolutions. We used and fused publicly available images from the Fire Detection Dataset, which features fire images captured by a drone camera, along with public fire images from the NASA Space Apps Challenge in 2018. The fire images are characterized by the following attributes:

- **Diversity**: Images show both indoor and outdoor fires, ranging from small flames to large conflagrations.
- **Complexity**: Scenes have different amounts of occlusions, lighting conditions, and smoke levels.
- **Annotations**: Bounding boxes are added to each image to mark fire zones, providing ground truth for AI model training and assessment.
- **Big Data**: Dataset is extensive in both volume and velocity, as training involves massive amounts of data, measured in gigabytes, collected by a drone's integrated camera and processed in real-time.

The total volume of harvested images is 10GB.

B. Adversarial Attack Images

Images that have been purposefully altered to deceive an AI system, specifically its object detection module, are known as adversarial attack images. The purpose of these adversarial attacks is to evaluate the AI model's resilience under various conditions. At their core, adversarial techniques have been extensively employed to manipulate and disrupt the same images, as elaborated in detail below.

C. Dataset Statistics and Attacks

In the YOLO [9] experiments, the dataset consists of the aforementioned fire images and their adversarially attacked counterparts. The classification process involved five (5) distinct categories, namely fire (benign) and four (4) different types of attack, with each category differentiated based on the attack criteria. The dataset is comprised of a balanced distribution of both normal and attacked images to guarantee an unbiased assessment. In Grad-CAM [10], using a CNN-ResNet50-Classifier, we analyze both the normal fire dataset and the adversarially attacked images from the different attack types.

Table I provides a summary of the dataset statistics.

TABLE I: Dataset Statistics

Category	Training Set	Validation Set
Normal Fire Images	3527	150
Attacked Images by DeepFool	3527	150
Attacked Images by Fast Gradient	3527	150
Attacked Images by NewtonFool	3527	150
Attacked Images by PGD	3527	150
Attacked Images by Square	3527	150
Total Number	21162	900

IV. Logical Modules and Adversarial Explanations Architecture

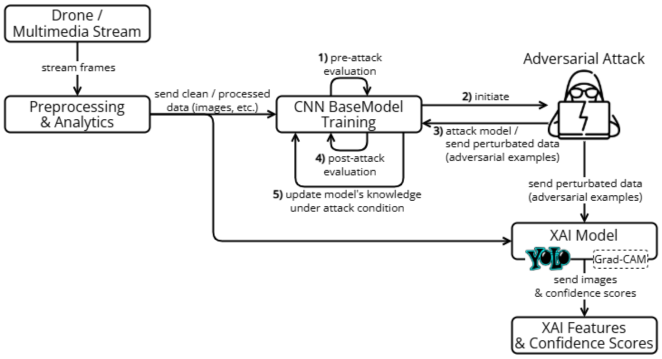
Drone / Multimedia Stream
stream frames
Preprocessing & Analytics
send clean / processed data (images, etc.)
1) pre-attack evaluation
CNN BaseModel Training
2) initiate
Adversarial Attack
3) attack model / send perturbated data (adversarial examples)
4) post-attack evaluation
5) update model's knowledge under attack condition
send perturbated data (adversarial examples)
XAI Model
YOLO
Grad-CAM
send images & confidence scores
XAI Features & Confidence Scores

Fig. 1: Adversarial Explanations Architecture

Figure 1, describes the overall architecture of the adversarial explanations, and a detailed description of each module will be discussed in the next sections.

A. Data Collection from Edge Devices

The first logical module describes the collection of data after being captured by an embedded camera on a drone. The data is then persisted for local processing and later streamed through a multimedia streaming server for cloud-based training.

B. Streaming, Preprocessing and Analytics Module

To address wildfires and environmental monitoring, the preprocessing module ensures that the attributes of the collected multimedia video streams are split into images and are appropriate for a situational awareness scenario. In this step, to guarantee abundant feature extraction, the images are resized, reshaped, or filtered. The preprocessing module utilizes Keras and skimage. More specifically, following the visual data exploration, we apply data normalization, scaling, and normalizing within a given numeric range, i.e. [0-1]. Keras is used to convert the images into arrays. Subsequently, we prepare and divide the images for binary or multiclass categorization. Binary classification focuses solely on the differentiation between attacked images, characterized as malicious, and non-attacked images, characterized as normal. In contrast, multiclass classification predicts the category of the image and verifies whether the AI system has been subjected to an attack. If an attack is detected, it further identifies the specific nature/category of the attack.

C. Training and Evaluation of the Classifier

The clean processed data is then used to train and assess a CNN Model, which serves as the baseline classifier for initializing a particular attack method, as shown in step 1 & step 2 of Figure 1 .

To determine the CNN Model's robustness, as measured by accuracy, two (2) specific conditions must be met: (i) after applying a selected evasion attack (e.g., Deepfool, FGM, Newtonfool, PGD, and Square Attack) from the ART library; and (ii) after adversarial training, to determine the percentage at which the CNN Model's accuracy may recover from the attack, as indicated in Table IV. The classifier from the previously trained CNN Model is used to launch an informed (i.e. labeled) attack, generating adversarial imagery examples. This labeled data is then utilized as examples for testing and evaluation using XAI features and confidence scores. To rigorously evaluate our model's performance accuracy, we follow a standard common methodology during the AI model training process by dividing the dataset into separate training (70%) and validation (30%) sets. Additionally, we use a variety of Keras layers, such as Sequential, Dense, Flatten, Conv2D, MaxPooling2D, Activation, and Dropout, to construct the CNN Model. Depending on the training task and the supporting learning objective, the activation functions for binary and multiclass classification differ. Finally, the CNN Model's accuracy is assessed using the validation set and classification reports in an average of K-fold splits.

Attack Category	norm	eps	eps_step	decay	max_iter	targeted	num_random	batch_size	nb_grads	eta	p_init	nb_restarts
DeepFool	-	0.000001	-	-	10	-	-	1	10	-	-	-
Fast Gradient Method (FGM)	np.inf	1	0.1	-	-	FALSE	5	-	-	-	-	-
NewtonFool	-	-	-	-	10	-	-	1	-	0.01	-	-
Projected Gradient Descent (PGD)	np.inf	0.3	0.1	None	10	FALSE	0	5	-	-	-	-
Square Attack	np.inf	0.3	-	-	100	-	-	128	-	-	0.8	1

TABLE II: Adversarial Attacks Configuration Parameters

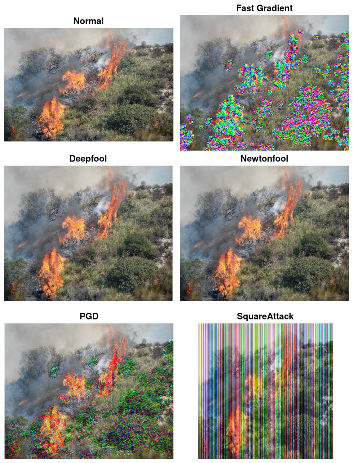

Fig. 2: Attacked Image

D. Adversarial Imagery Examples Generation

In Figure 2, we illustrate how additional adversarial examples emerge consistently when the original image undergoes meticulous preprocessing and vigorous attacks. It becomes evident that while DeepFool and NewtonFool may initially appear to leave the data untouched, the reality is different regarding their impact on the model's performance.

1) Evasion Attacks: Adversarial Evasion (AE) attacks were first created for imagery data, where the main goal is to hide any alterations from human sight [11].

Adversarial attacks are categorized into white-box, where attackers have full model access for accurate gradient computation, and black-box, which rely solely on input-output queries. In our experiments, we employed three (3) different types of adversarial attacks, detailed below :

- Optimization Attacks [12]: These attacks systematically distort an image and deceive a model while minimizing noticeable changes. They use mathematical techniques, such as gradient descent, to detect the least identifiable way of fooling the model.
- Gradient-Based Attacks [13]: Gradient-Based Attackers exploit different levels of model access to manipulate its outputs using gradient-based methods to generate adversarial inputs. By taking advantage of model vulnerabilities during training and inference, they manage to create misclassifications or steal sensitive information.
- Query-Based Attacks [14]: These attacks interact with the target model to generate adversarial images with subtle perturbations, misleading the model in the process. They optimize these perturbations to make them harder to detect. Decision-based attacks are more common than score-based attacks because they are more realistic.

TABLE III: Comparison of Attack Methods

Method	**Type**	**Access**	**Feature**	**Limitations**
DeepFool	Optimization	White-box	Iterative, boundary-focused	Limited for deep, highly non-linear networks.
FGM	Gradient-based	White-box	Fast, single-step attack	Less effective than iterative methods.
NewtonFool	Optimization	White-box	Newton's method optimization	Expensive due to second-order calculations.
PGD	Gradient-based	White-box	Projected steps for robustness	High cost with many iterations.
Square Attack	Query-based	Black-box	Random square perturbations	Requires many queries; larger perturbations.

In this work, we experimented with several types of adversarial attacks. Table III summarizes the main differences between DeepFool [15], FGM [16], NewtonFool [17], PGD [18], and Square Attack [19], which we used to demonstrate and benchmark the proposed robustification approach.

These attacks involve various trade-offs between speed, effectiveness, and computational cost, making them suitable for different scenarios based on the objectives and model defenses. For this study, we employed specific configuration parameters to launch the attacks, as shown in Table II. In addition, to encourage the research community to further investigate these issues, we have made publicly available a dataset [20] that is created by initiating these particular adversarial attacks.

E. Explain and Justify Decision-making through XAI Features and Confidence Scores

Figure 1 illustrates in detail the steps involved in explaining how adversarial attacks are initiated. First, we perform a pre-attack evaluation on the baseline model, which is then used to initiate our attack and generate the adversarial examples. A post-attack evaluation follows to assess the performance of the model on the perturbated data. The resulting adversarial images (perturbated data) are then used to augment the dataset that includes both benign and attacked images and update the model's knowledge under attack conditions. After generating the perturbated data, the images are sent to the XAI Model Module.

As part of the XAI process, a CNN classifier is constructed using this augmented dataset, which consists of two (2) data classes: "fire" (representing normal instances of fire alongside the attacked fire images) and "non-fire" (representing background images without fire). The CNN classifier is designed to be robust against adversarial attacks and employs Grad-CAM to generate a significance map of salient pixels, providing explanations even when attacked data instances are present.

Meanwhile, YOLOv8 has been trained on five (5) classes of data: "fire" (representing normal instances) and "fgm", "deepfool", "newtonfool", "pgd", "square attack" (representing different types of adversarial occurrences). Explainability is provided through YOLO's bounding boxes and confidence scores, while also serving as the attack type detector since it can also detect and classify the specific adversarial attack applied.

V. Experimental Results on Diverse Attacks with Explanations and Confidence Scores

As shown in Table IV, the NewtonFool ranks as the least effective among the tested attacks. The Base CNN Model initially has an accuracy of 85%, which drops to 42% after the attack. Although adversarial training does not fully restore robustness, NewtonFool consistently demonstrates lower efficacy, as the adversarial training does not restore the accuracy of the model (accuracy after adversarial training - 44%). Conversely, while DeepFool shows promising effectiveness with a decrease to 10% accuracy, the anticipated improvements from the robustification process are not fully realized as the accuracy shows a light increase to 65% accuracy. Both the Fast Gradient Method (FGM) and PGD significantly impact CNN model performance, with a decreased to 10% of the accuracy of the base model 90% and 87% respectively, while the robustification mechanism achieves strong results in these cases as it manages to regain the accuracy of the models from 10% to 88% for FGM and to 86% for PGD.

Attack Category	CNN Model	Attacked Model	Robust CNN
DeepFool	0.85	0.10	0.65
FGM	0.90	0.10	0.88
NewtonFool	0.82	0.42	0.44
PGD	0.87	0.10	0.86
Square Attack	0.87	0.14	0.63

TABLE IV: Accuracy of CNN Training, Attacked, and Robustified Models

In the following paragraphs, we present the YOLO results alongside the explanation outcomes with confidence scores. The results are summarized in three (3) tables. Table V establishes the baseline performance evaluation on well-known datasets and NN models. ResNet50 achieves an accuracy of 67% on CIFAR-100, while YOLOv8 (small and large) yield mAP50 scores of 0.446 and 0.478, respectively, on the COCO dataset. Table VI compares our proposed methods. Notably, the CNN-Res50-Robust-Classifier attains high performance with 98% accuracy and 97% precision, outperforming the YOLOv8 models—which register lower accuracy (58–62%), precision (72–79%), and mAP50 (0.25–0.30). Finally, Table VII details how YOLO's confidence scores vary under different adversarial attacks. While the model maintains relatively high performance under normal conditions (with a confidence score of 0.76), its performance drops significantly for certain attacks (e.g., DeepFool, with a score of 0.39), illustrating the YOLO framework's vulnerability to adversarial perturbations.

In addition, we can see the explanations of both YOLO and the Resnet-classifier with Grad-CAM. In Figure 3, YOLO successfully highlights the fire-affected area even under attack while also identifying the type of attack along with a confidence score. Figure 4 showcases the explanation robustness of the ResNet classifier, where the fire-related explanations remain valid even under adversarial attacked data instances. This robostify the models, as even under attack the model and the explanations still work well, and help us understand also how different attacks behave on the models.

These figures depict the varying impacts of different attacks on the images. The FGM attack stands out as it creates significant confusion for both models, resulting in difficulties in accurately identifying and articulating the presence of fire. In comparison, the DeepFool and NewtonFool attacks exhibit similar effects, leading to a less pronounced but still notable challenge in image interpretation. This emphasizes the distinct ways in which each attack affects the models' ability to process visual information.

TABLE V: Baselines YOLOv8 on COCO dataset & ResNet50 on CIFAR-100

	Accuracy	mAP50
ResNet50	0.67	-
YOLOv8 small	-	0.446
YOLOv8 large	-	0.478

TABLE VI: Results

	Accuracy	Precision	mAP50
CNN-Res50-Robust-Classifier	0.98	0.97	-
YOLOv8 small	0.58	0.72	0.25
YOLOv8 large	0.62	0.79	0.30

TABLE VII: Yolo Confidence score along different attacks

Class	Precision	Recall	Conf.Score
Normal Fire	0.89	0.90	0.76
Deepfool	0.43	0.10	0.39
Fast Gradient	1.0	0.55	0.52
Newtonfool	0.44	0.51	0.40
PGD	1.0	0.86	0.74
Square Attack	0.98	0.83	0.72

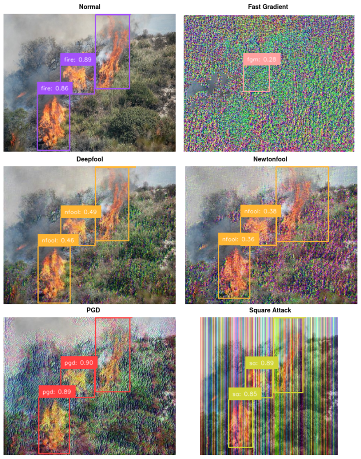

Fig. 3: YOLO explanations on examples

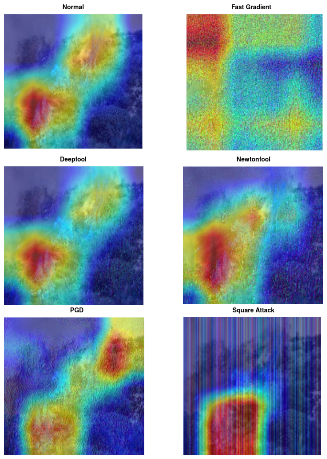

Fig. 4: Grad-CAM explanations on examples

VI. Conclusion

In conclusion, our approach addresses high-physical risk scenarios in critical infrastructures and complex environments. The robustified model, deployed on an edge processor aboard a drone, achieves high accuracy and precision in object detection and classification. Additionally, adversarial explanations from the YOLO large model offer valuable insights into potential attacks. By combining these explanations with Confidence Scores, our method provides a promising solution for applications demanding exceptional reliability in object detection and image classification. Furthermore, our research demonstrates that explanation-driven adversarial attacks are an effective approach to enhancing the security of multimedia edge applications. By integrating clear and interpretable adversarial insights with robust confidence scoring, our robustified model not only identifies existing vulnerabilities but also provides actionable awareness to improve model resilience.

In future work, we aim to enhance the proposed Adversarial Explanations framework by incorporating advanced localization features, further supporting more informed decision-making processes. Through the use of adversarial attack types, a larger dataset, and performance measurements, we also plan to further expand our study to better benchmark our approach. We also intend to analyze how the results and attacks of explainability affect performance and computational costs.

Acknowledgment

This work has received funding from the Research and Innovation Foundation under Restart Research 2016-2020 Programme and the DUAL USE/0922/0024 agreement of CYGNUS project, and the European Union's Horizon Europe TALON project with GA No 101070181.

References

[1] I. P. Theodora Anastasiou and S. Karagiorgou, "Adversarial explanations for informed civilian and environmental protection," *2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 2672-2681*, 2024.

[2] J. Gao, R. Mudumbai, X. Wu, J. Yi, C. Xu, H. Xie, and W. Xu, "Towards unlocking the mystery of adversarial fragility of neural networks," *arXiv preprint arXiv:2406.16200*, 2024.

[3] X. W. Y. S. M. M. L.-P. C. Shaohui Mei, Jiawei Lian, "A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing: Surveying and benchmarking." 2024.

[4] A. Benchama and K. Zebbara, "Novel approach to intrusion detection: Introducing gan-mscnn-bilstm with lime predictions," *arXiv preprint arXiv:2406.05443*, 2024.

[5] Q. Card, K. Aryal, and M. Gupta, "Explainability-informed targeted malware misclassification," *arXiv preprint arXiv:2405.04010*, 2024.

[6] A. H. Kristoffer Wickstrøm, Marina Marie-Claire Höhne, "From flexibility to manipulation: The slippery slope of xai evaluation," *arXiv preprint arXiv:2412.05592*, 2024.

[7] P. B. Hubert Baniecki, "Adversarial attacks and defenses in explainable artificial intelligence: A survey." *arXiv preprint arXiv:2306.06123*, 2023.

[8] X. G. Z. S. Yongjie Wang, Tong Zhang, "Gradient-based feature attribution in explainable ai: A technical review." *arXiv preprint arXiv:2403.10415*, 2024.

[9] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-Time Flying Object Detection with YOLOv8," May 2024, arXiv:2305.09972 [cs]. [Online]. Available: http://arxiv.org/abs/2305.09972

[10] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, arXiv:1610.02391 [cs]. [Online]. Available: http://arxiv.org/abs/1610.02391

[11] I. G. A. Kurakin and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[12] N. C. D. Wagner, "Towards evaluating the robustness of neural networks," *arXiv preprint arXiv:1608.04644*, 2017.

[13] T. A. Kartik Gupta, "Improved gradient based adversarial attacks for quantized networks," *arXiv arXiv:2003.13511v2*, 2021.

[14] N. K. M. S. Naveed Akhtar, Ajmal Mian, "Advances in adversarial attacks and defenses in computer vision: A survey," *arXiv arXiv:2108.00401v2*, 2021.

[15] P. F. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, "Deepfool: a simple and accurate method to fool deep neural networks," *arXiv arXiv:1511.04599v3*, 2016.

[16] C. S. Ian J. Goodfellow, Jonathon Shlens, "Explaining and harnessing adversarial examples," *arXiv arXiv:1412.6572v3*, 2015.

[17] X. W. Uyeong Jang and S. Jha., "Objective metrics and gradient descent algorithms for adversarial examples in machine learning. in proceedings of the 33rd annual computer security applications conference (acsac '17)," *https://doi.org/10.1145/3134600.3134635*, 2017.

[18] L. S. D. T. A. V. Aleksander Madry, Aleksandar Makelov, "Towards deep learning models resistant to adversarial attacks," *arXiv arXiv:1706.06083v4*, 2019.

[19] N. F. M. H. Maksym Andriushchenko, Francesco Croce, "Square attack: a query-efficient black-box adversarial attack via random search," *arXiv arXiv:1912.00049v3*, 2019.

[20] T. Anastasiou and I. Pastellas, "Cygnus dataset addressing civilian application challenges," 2025.