AI Robustness against Attacks in City-Scale Autonomous Drone Deployments

Abdul-Rasheed Ottun[1], Adeyinka Akintola[1], Mohan Liyanage[1], Michell Boerger[2], Pan Hui[3,4], Sasu Tarkoma[3], Nikolay Tcholtchev[2], Petteri Nurmi[3], and Huber Flores[1]
[1]University of Tartu, Estonia; [2]Fraunhofer FOKUS, Berlin, Germany;
[3]University of Helsinki, Finland [4]HKUST, Hong Kong
ottun@ut.ee

***Abstract*—The use of autonomous drone technology in crowded urban environments necessitates AI models that are able to operate robustly to ensure the safety of humans and the surrounding infrastructure. Adversarial attacks, particularly through data poisoning, can pose significant threats to the robustness of AI models. This paper contributes by assessing threats to AI robustness and the deployment of autonomous drones in cities. We analyze the impact of poisoning attacks on autonomous drones and demonstrate how explainable artificial intelligence (XAI) techniques can be employed to detect them. The results show that, while XAI is beneficial, it may not be all-sufficient, as covering the full spectrum of potential data manipulations is cumbersome. We then delve into the risks, opportunities, and research challenges, ultimately paving the way for city-scale deployments of autonomous drones.**

***Index Terms*—Autonomous Drones; XAI; Human-in-the-loop**

I. Introduction

Autonomous drone technology has undergone significant advancements, encompassing autonomous ground vehicles (AGV) such as delivery robots, service robots, and unmanned aerial drones (UAVs) [1], [2]. These technological strides have enabled drones to be effectively utilized for delivering essential goods, including food and medicine, with further applications anticipated in environmental monitoring, urban surveillance, and related fields, are anticipated [3]. These diverse applications are made possible by sophisticated AI models that provide the capabilities for autonomous operations, such as navigation and localization support [4].

As autonomous drones operate in crowded urban settings, ensuring the robustness, resilience, and consistency of their AI models is crucial for safe and reliable operation. Adversarial attacks pose a significant threat to AI robustness, as they aim to induce unintended consequences in the AI models, potentially leading to damage to the city or harm to individuals (Fig. 1), for instance, by colliding with urban infrastructure. Of particular concern are data manipulation attacks, which can be executed by manipulating the environment without any direct access to the autonomous drone. For example, an attacker wearing an adversarial generative patch can deceive the drone, causing it to misinterpret its location and change course. To ensure safe and trustworthy operations of autonomous drones, it is essential to comprehensively understand how these attacks affect autonomous drones and develop effective strategies to mitigate their effects for safe and trustworthy operations.

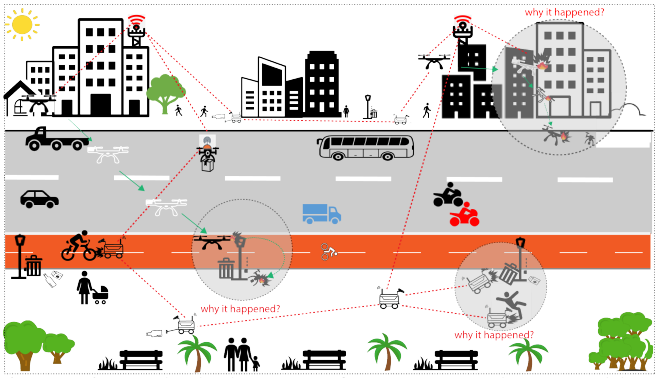


Fig. 1: City-scale deployment of autonomous drones and how these can malfunction or misbehave in urban settings.

This paper contributes by assessing the threats posed by adversarial attacks on AI robustness and the widespread deployment of autonomous drones in urban environments. The primary objective is to appraise the readiness of these deployments before they become a reality. A benign use case was employed to demonstrate the impact of adversarial data manipulation attacks on AI robustness. Following this, the potential of explainable AI (XAI) methods to enhance AI robustness and identify attacks is examined. XAI methods are techniques used to make the decisions and processes of AI models understandable to humans. Our findings indicate that XAI methods demonstrate high accuracy in identifying basic attacks by examining how the attack significantly alters the data features used for learning. However, identifying the root cause of the issue may be challenging. Additionally, XAI methods require prior knowledge of potential attacks, making it difficult to cover the full spectrum of potential data manipulations. The paper concludes by reflecting on the results and elaborating on the risks, opportunities, and research challenges that need to be addressed to enhance AI robustness. Potential technologies that could improve the robustness and resilience of AI models are highlighted. This work lays the groundwork for the practical implementation of city-scale deployments of autonomous drones, envisioning a future where they become a common sight in our everyday environments.

II. The Impact of Attacks on Autonomous Drones

To illustrate the potential vulnerabilities of autonomous drones, we begin by demonstrating how abnormal model

behaviour and potential disruption of AI performance can be caused by external data poisoning attacks.

Threat Model: A generic threat model is considered, where the AI model in the autonomous drone is targeted to fail by the attacker. The attack can result in specific misbehaviour, such as accidents caused by the failure of navigation support to recognize pedestrians or cars. Alternatively, it could be an attack that causes the AI to malfunction, for instance, a sponge attack that drains the autonomous drone's resources or a ransomware attack that prevents normal operations. The motivation for the attack could be to harm the citizens or the city, financial gain, or notoriety.

Application Scenario: The use case involves litter recognition with autonomous drones, serving as a representative example of AI-driven operations. In this scenario, thermal images are analyzed in real-time to identify various litter objects and determine their materials. Specifically, the drone analyzes the dissipation of sunlight-induced thermal radiation that is captured by a thermal camera integrated onto the autonomous drone [5]. Attacks against the model can disrupt the operations of the autonomous drones or drain their resources. More serious attacks could target navigation, obstacle detection, or other functions that could directly result in harm to citizens or damage to the environment. While our use case presents a benign example to illustrate the risks of attacks without risking the citizens or the environment, our findings are applicable to any AI applications that rely on computer vision.

Experimental Setup: Experiments were conducted using three common litter objects with different materials: (A) Plastic bottle, (B) Face mask and (C) Cardboard cup. Video footage of disposed litter was recorded by the autonomous drone, which was then pre-processed and analyzed to identify litter [5]. Data injection attacks were employed to manipulate the input data using blurring and steganography techniques [6], [7]. These attacks, while easy to implement, can have significant impacts, including the installation of backdoor triggers [7] to drain the autonomous drone's resources [8] or create unexpected behaviours. Notably, as we focus on data manipulation, the attacks do not require direct access to the vision system of the drone, as they can manipulate objects in the environment or use additional devices, such as lasers, to alter the data captured and analyzed by the sensors [9].

Results: The thermal dissipation times for the litter objects were measured and are as follows: plastic bottle 62.5s, cardboard cup 72.5s, and face mask 82s. The relative differences align with those reported in [5] for the same materials. However, the absolute values differ due to the varying intensity of the thermal source, the size of the material, and the total exposure time. To analyze poisoning, two levels of poisoning are considered: 10% (low) and 40% (high). Values higher than 40% result in poisoning taking over the model. For blurring, the dissipation times after poisoning are 51.5s (plastic bottle), 51.1s (cardboard cup), and 38.4s (face mask) for 10% poisoning, and 49.3s, 22.9s, and 40.5s when 40% is poisoned, respectively. The relative differences in the thermal dissipation values thus change significantly, breaking the AI model used for detecting litter materials. The resource drain on the autonomous drone also increased notably, highlighting the potential real-world implications of such attacks. In contrast to blurring, steganography attacks did not influence thermal dissipation times, indicating a varying response of the model to different attack types.

III. XAI as Model Diagnostics

Explainable AI (XAI) methods offer a potential solution for overcoming attacks by offering diagnostics that can identify when an attack occurs. In the following, the potential of different XAI methods to detect targeted poisoning attacks is analyzed, and their benefits and disadvantages are evaluated. The quantifiable values provided by XAI methods in benign cases are examined, and their performance against poisoned data is analyzed. Additionally, the impact of different processing techniques on the behavior of XAI methods when applied to the full image and the processed image with the background removed is investigated to understand how different processing techniques affect the behaviour of XAI methods.

Experiment Setup: The TrashNet litter classification dataset, comprising 2527 litter images [10], was utilized for the experiment. This dataset was chosen due to its large collection of real-world images, enabling the analysis of different environments and contexts for litter classification. A convolutional neural network (CNN) model was trained because it had demonstrated strong performance with this data [10]. Images were resampled to 300×300 to have consistent input dimensionality. Data augmentation techniques, including horizontal and vertical flipping and rescaling were applied to the training set, which consisted of 2276 images trained with a batch size of 32 for each epoch iteration. The remaining images were used for testing. A collection of 10 poisoned and non-poisoned images was separately considered to illustrate the performance of XAI methods. The experiment was conducted on the Google Colab platform using the latest version of the Keras library (2.8.0) with TensorFlow (v2.8.2).

XAI methods: Three model-agnostic XAI methods were considered for the analysis: LIME [11], SHAP [12], and Occlusion sensitivity. These methods do not rely on CNN gradients; instead, they use perturbations to interpret model behaviour and derive versatile (global and local) insights that can be compared across different models. LIME segments images into superpixels, SHAP attributes importance to features, and Occlusion sensitivity [13] uses sensitivity heat maps to observe prediction impact. The selected XAI methods were applied separately to images with the background removed (i.e., only the litter object) and to the original input image. The object extraction process, depicted in Figure 2, involved applying a dynamic patch (determined using object detection) on the image to isolate it. From the final output of the XAI methods, a pixel percentage metric was calculated to captures the importance of a pixel.

Samples and Poisoning: Six litter categories were considered: glass, paper, cardboard, trash, metal and plastic. For poisoning the data, two attacks were considered: blurring and steganography. Blurring can cause autonomous drones to misidentify targets in urban areas, such as crossing signals

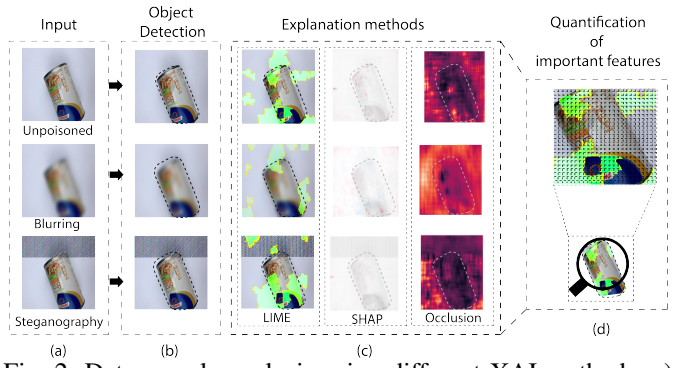


Fig. 2: Data sample analysis using different XAI methods, a) Data samples (poisoned and unpoisoned), b) Object detection, c) XAI methods output over samples (LIME, SHAP and Occlusion sensitivity) and d) Object extraction

and pedestrian side walks. Steganography introduces extra information in the binary information of the images, which can become resources-intensive for the autonomous drone as more processing power is required to extract relevant information, similar to a sponge attack. The level of poisoning was systematically assessed by poisoning the data in 10% increments from 10% to 40%.

IV. Results

Model Performance under Poisoning: The performance of the CNN in classifying litter is 0.7 when no data is poisoned, but this performance gradually decreases as the data is poisoned. After the blurring attack, the model accuracy was decreased to 0.61 (10% poisoned); 0.53 (20% poisoned); 0.53 (30% poisoned) and 0.60 (40% poisoned). Similarly, the steganography attack decreased the model accuracy to 0.52 (10% poisoned); 0.52 (20% poisoned); 0.62 (30% poisoned) and 0.67 (40% poisoned). In both cases, a clear drop in accuracy was observed. Unlike the earlier experiment, the performance drop was higher for data poisoned with steganography than with blurring. This difference in results is simply due to differences in the sensors (RGB vs thermal camera) and the processing pipeline, highlighting how the effectiveness of the attack is influenced by the task and the specifics of the AI being used. The performance drop resulting from poisoning depends on how much the attack affects the patterns in the data. In general, as larger amounts of the data become poisoned, the inference process starts to be dominated by the poisoned patterns, while smaller amounts result in distortions that can confuse the model. This pattern is observed with both attacks, with the sole exception being blurring at a 10% rate, as a small level of blurring does not distort the patterns of the litter object sufficiently to impact the AI model.

Analysis of XAI Methods: The effectiveness of XAI methods was analyzed considering 10 randomly chosen poisoned samples from each litter category to report the accuracy of estimating the correct class for each sample. Table I summarizes the results for the different XAI methods. The effect of poisoning depends on the litter category and the extent of poisoning. Paper and cardboard objects with regular shapes are the easiest for the XAI methods, whereas classes containing irregular shapes (metal, plastic, trash) exhibit the highest variation in results. Similar to the results for the CNN model, a higher level of poisoning can result in a smaller drop, or in some cases even an increase, in performance. This pattern is more common for steganography, as the poisoned data starts to dominate the inference process once a higher fraction of the data is poisoned. While XAI methods can only help recognize poisoning without directly enhancing the performance of the classifiers, they can indirectly offer insights that can help to improve the classifiers. For example, samples that are identified as poisoned can be used to develop data augmentation techniques, which can be incorporated into the model training process to improve the robustness of the classification models. To illustrate this point, blurring is already a commonly used data augmentation technique for improving the training of AI models. From our experiment, it was visually observed that the attack tends to impact the background more than the foreground. Thus, processing techniques that separate the object from the background are likely to improve performance.

Diagnosing Objects with XAI: Lastly, the effect of data poisoning over the important features of the object when it is isolated from the background is examined. The coefficient of variation of the poisoned pixels, which depicts the ratio of the standard deviation to the mean, was considered. The higher the value of the coefficient, the higher the dispersion, and thus the better the method is at identifying poisoned data. The results for the 10 test samples of each class are shown in Figure 3. For the blurring attack, the average values of the XAI methods are 0.35 (LIME), 0.17 (SHAP) and 0.3 (Occlusion). For data poisoned with steganography, the corresponding values are 0.22 (LIME), 0.10 (SHAP) and 0.26 (Occlusion). One-way ANOVA between the three XAI methods indicates statistical significance, ($F(2,1794)=118.4$, p-value < 0.001), indicating that there are differences in the applicability of the different XAI methods. The higher average values of LIME and Occlusion indicate that they are better at identifying the poisoned data. SHAP performs well for metal objects which are the most irregular, but struggles with other categories. A one-way ANOVA test was also used to verify that the difference in variation across classes is significant across all XAI methods, poisoning attacks, and levels of poisoning ($F(5,1791)= 14.76$, p-value < 0.001). Across all XAI methods, the coefficients of variation are larger for steganography than for blurring, indicating that XAI methods can also provide clues about the nature of the error. The effect between attack type and data poison level was also investigated. A two-way ANOVA test between attack type and data poisoning level indicates a significant effect ($F(1,4)=3.396$, p-value < 0.01), i.e., the coefficients of variation depend not only on the attack type but also the extent of poisoning. Taken together, these results show that XAI methods help to identify the important features of the image, even after data is poisoned. However, their effectiveness is affected by the object, the type of attack, and the extent of poisoning generated by the attack. In any case, even when the objects can be separated and analyzed, an elaborate processing pipeline is required for processing,

	LIME					SHAP					Occlusion Sensitivity				
Poisoning Level	**0%**	**10%**	**20%**	**30%**	**40%**	**0%**	**10%**	**20%**	**30%**	**40%**	**0%**	**10%**	**20%**	**30%**	**40%**
Poisoning type	**Blurring**														
Cardboard	1	0.9	0.9	0.8	0.9	1	0.8	0.8	0.8	0.9	1	0.8	0.9	0.8	0.9
Glass	1	0.7	0.7	0.8	0.6	0.9	0.8	0.8	0.8	0.6	1	0.8	0.8	0.8	0.6
Metal	0.7	0.8	0.7	0.8	0.4	0.6	0.8	0.7	0.8	0.4	0.7	0.7	0.7	0.8	0.4
Paper	0.9	0.9	0.7	0.7	1	0.9	0.9	0.9	0.7	0.9	0.9	0.9	0.9	0.7	1
Plastic	0.8	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7
Trash	0.9	0.6	0.6	0.8	0.7	0.9	0.6	0.6	0.8	0.6	0.9	0.6	0.6	0.8	0.7
Average	**0.9**	**0.8**	**0.7**	**0.8**	**0.7**	**0.9**	**0.8**	**0.7**	**0.8**	**0.7**	**0.9**	**0.8**	**0.8**	**0.8**	**0.7**
Poisoning type	**Steganography**														
Cardboard	1	1	1	0.8	0.8	1	1	1	0.8	0.8	1	1	1	0.8	0.8
Glass	1	0.7	0.6	1	1	1	0.6	0.6	1	0.9	1	0.7	0.6	1	0.9
Metal	0.7	0.6	0.6	0.7	0.7	0.7	0.6	0.6	0.7	0.8	0.7	0.6	0.6	0.7	0.8
Paper	0.9	1	1	1	1	0.9	1	1	0.9	0.9	0.9	1	1	0.9	0.9
Plastic	0.8	0.7	0.7	0.7	0.8	0.8	0.7	0.7	0.7	0.8	0.8	0.7	0.7	0.7	0.8
Trash	0.9	0.8	0.6	0.9	1	0.9	0.6	0.6	0.9	0.9	0.9	0.7	0.6	0.9	0.9
Average	**0.9**	**0.8**	**0.8**	**0.9**	**0.9**	**0.9**	**0.8**	**0.8**	**0.8**	**0.9**	**0.9**	**0.8**	**0.8**	**0.8**	**0.9**

TABLE I: Individual performance of XAI methods on selected poisoned and unpoisoned samples.

which drains the resources of the autonomous drones faster and limits their operations.

V. Reducing Failures and Improving Resilience

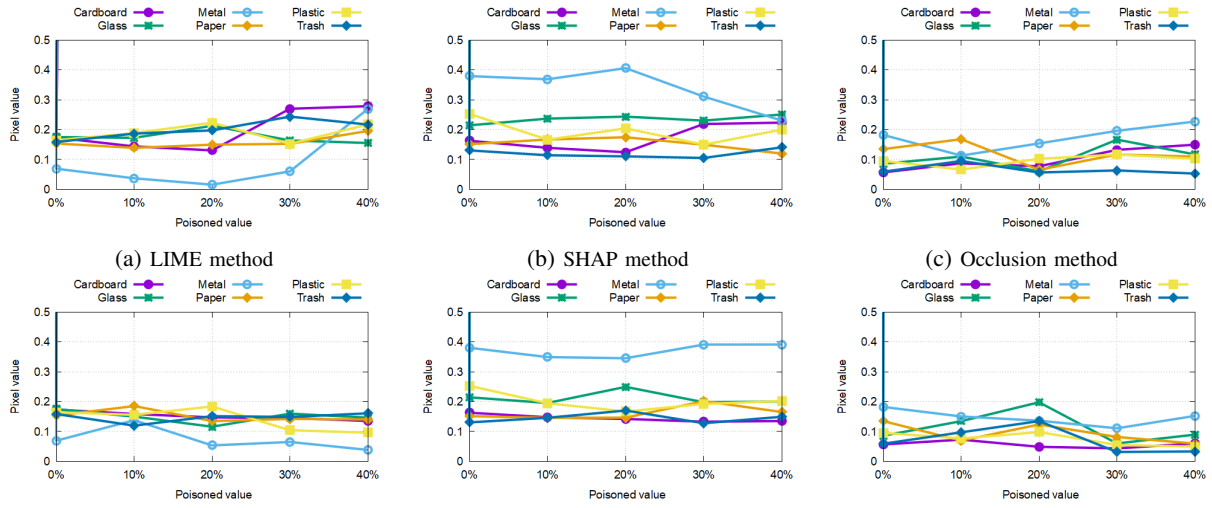
Practical drone deployments can be enabled through existing technologies, but the lack of resilience and accountability in AI models prevents these deployments from being robust. Latest regulatory requirements mandate that the AI systems must exhibit robustness against various challenges and adversarial conditions. Next, important challenges and state-of-the-art approaches that can improve AI robustness are highlighted.

Immersive evaluation and remediation: Current AI model evaluation and testing processes are not extensive enough to assess performance in all possible situations in which AI models can fail. Evaluating AI models extensively within a short period without delay is a key challenge, especially for autonomous drones operating in urban scenarios. A promising synergy of technologies that can aid in this manner is digital twins and generative AI. Digital twin technology can provide the means to build digital representations of autonomous drones running AI models, whereas generative AI can provide immersive experience to adjust the situational context of the autonomous drone in a large spectrum of different settings. For instance, the urban context of the autonomous drone can be dynamically changed to evaluate navigation in different terrains, e.g., rural vs urban areas. Immersive evaluation can also be used to apply different counter-measurements in case autonomous drones are compromised or hampered. For instance, label sanitization methods can be applied to digital autonomous drones to select the most suitable for their physical counterparts.

Data bias and drift detection: Autonomous drones deployed in natural environments engage in ongoing learning processes, enhancing AI resilience via distributed model training like federated learning. However, data collected by these drones may include privacy-sensitive information (e.g., face, speech, or car registration plates). Ensuring privacy necessitates employing techniques like data obfuscation and privacy-preserving methods [6]. As AI models are vulnerable to data poisoning attacks that can disrupt their operations, there is a need for techniques that can quantify model resilience to erroneous updates before they impact the functionality of autonomous drones. A key challenge is the separation between non-intentional malfunctions (e.g., camera failure), intrinsic data biases, and targeted attacks. Existing methods largely target one type of issue (e.g., drift or poisoning) without being able to distinguish between the different causes. Another challenge is to ensure that the methods can operate at different temporal scales, i.e., they can identify problems even when biased or erroneous data is aggregated with valid data and when erroneous data arrives gradually.

Continuous model verification: Trustworthy autonomous drone operations also require easy-to-use solutions for continuous, on-site analysis and verification of decisions, especially for large-scale deployments like in cities [14]. Otherwise, the effort needed for verification can limit the scalability of deployments. Diagnosing AI models often involves accessing the internal structure of the model, which typically requires halting drone operations, modifying source code, and bypassing security features - often necessitating lab work. Explainable AI (XAI) methods provide valuable insights into AI behaviour, but their effectiveness is limited by the need for access to data and the model structure, making them less viable as a comprehensive solution. Integrating XAI into security features, such as trusted execution environments (TEEs), could help. However, TEEs currently have significant computational limitations for such practical applications. Additionally, other formal verification methods for AI models encounter challenges, particularly when dealing with non-linear activation functions [15], [16].

Model interpretability and physical drone components: The performance of AI models is intrinsically linked to the resources and components of the autonomous drone [17]. Over time, these components need maintenance, or may be upgraded to improve the operations of the drone. These changes can affect the model and result in unexpected behaviour. For example, integrating a higher-resolution camera affects the dimensionality of the input data and may capture more detailed images. This can require replacing the model or at least re-



(a) LIME method (b) SHAP method (c) Occlusion method

(d) LIME method (e) SHAP method (f) Occlusion method

Fig. 3: Object analysis with each XAI method as data is poisoned with, (a-c) Blurring and (d-f) Steganography

training it. In terms of interpretability, this requires linking model diagnostics with the physical components of the autonomous drone and being able to analyze and interpret the effects that individual physical components have on the model's decisions. It is important to note that these changes do not necessarily affect the input data. For example, autonomous drones can operate using different payloads which affects the weight and their resource consumption. This necessitates integrating physical configuration directly into the model diagnosis. This integration is essential for detecting unsafe operations, such as identifying unsafe payloads. Current methods are insufficient as they are unable to link model behaviour with the physical characteristics of the operating environment.

Cooperative operations and failures: Effective cooperation between autonomous drones is important for balancing workload between them. This coordination results in dependencies between the AI models deployed on the different autonomous drones, and understanding potential errors or threats requires analysing the combined logic of all autonomous drones working in tandem, e.g., swarm intelligence [18]. Current XAI and other model diagnostics techniques are tailored to analysing individual models, and hence, they can only be used if the autonomous drones have a global model that integrates the decision logic of all autonomous drones working together. It is important to note that this task is more complex than analysing the performance of individual autonomous drones, as attacks or errors can affect only some of the autonomous drones yet have an influence on all of them by compromising the coordination of the drones through the network [19]. Understanding the effects of target errors on autonomous drones' coordination network and collaboration requires the use of improved diagnostic mechanisms for analysing network formation groups, individual parts of the network (slices) and the model.

Human oversight and AI: The advanced human-like reasoning of AI models has led to concerns about trust and safety among human operators and developers. Consequently, human oversight has been established as a key requirement for ensuring that AI models are trustworthy. Human-in-the-loop is preferred over fully autonomous approaches, such as automation-in-the-loop agents because it allows for human expertise and judgment to be applied throughout the life cycle of AI models. This ensures that the AI operates within defined boundaries, delivers desirable outcomes, and behaves as expected after deployment [20]. Moreover, human oversight is critical in practical drone deployments, as failures must be remediated rapidly to avoid halting operations. Human expertise and past experiences can be leveraged to improve AI robustness. Instead of performing a time-consuming remediation analysis, human experts can intuitively select near-optimal solutions to remediate issues. A key challenge, however, is communicating the dissected logic of AI models to human experts. Although regulations like the EU and US AI Acts, as well as China regulations, require human-in-the-loop involvement in generative AI, best practices for achieving this are not yet well-defined.

VI. Discussion

Additional Application Areas: The presented work utilizes an AI model for litter classification to demonstrate the impact of data poisoning on AI decision-making. While these attacks can be applied to various algorithms and data types, their impact may differ. For instance, random spoofing may exploit specific model characteristics in time series data compared to image data.

Stakeholders: Model diagnostics are essential for companies and organisations deploying autonomous drones in urban environments, such as delivery carriers for smart cities. Similarly, municipalities and government authorities may mandate the integration of these diagnostics into drones before issuing permits.

Practical Limitations: In addition to XAI, several other approaches exist to dissect the learning and inference processes of AI models. Model-agnostic XAI methods were employed

to ensure comparability across different models. Methods such as Guided Backpropagation and Grad-CAM can provide more detailed analyses of AI models, but they are not model-agnostic, requiring adapating them to the used AI model architecture. Notably, for autonomous drones operating in urban areas, XAI methods are critical for ensuring accountable operations. The choice of the XAI method for analyzing AI must meet not only technical criteria but also regulatory and legal requirements set by governmental entities.

Other Vulnerabilities: Attacks on AI models are not the sole threats to autonomous drones; hardware and software components are also vulnerable, leading to operational failures and misbehaviour. Hence, it is crucial to develop methods that can operate on autonomous drones and distinguish between these vulnerabilities in real-time. For example, a sponge attack can drain the battery of the autonomous drone, while a jamming attack can disrupt its GPS localization. Similarly, software backdoors can expose autonomous drones to manipulation. Consequently, the robust deployment of autonomous drones requires security measures that extend beyond just the AI components.

VII. Summary and Conclusions

Robust AI models are crucial for ensuring trustworthy operations of autonomous drones, especially in complex and dynamic environments such as cities. These models must exhibit accuracy and resilience; otherwise, there is a risk of causing harm to citizens or damaging the environment. The current state of practical drone deployments and their trustworthiness was explored, with a focus on the importance of model diagnostics in light of data poisoning attacks that affect the robustness of AI models used in autonomous drones. Notably, these attacks do not require access to the device or AI model, as they are executed solely by manipulating the inputs used. Furthermore, the significance of explainable AI (XAI) methods in identifying data poisoning issues was demonstrated, acknowledging that XAI methods also have their limitations. Based on these findings, challenges and opportunities to enhance AI robustness were presented, alongside the identification of promising technologies and requirements to make autonomous drone operations safer.

Acknowledgment

This research has been financed by the European Union and Estonian Research Council via project TEM-TA101, the Academy of Finland project 339614 and is also part of SPATIAL project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101021808. Special thanks to Zhigang Yin and Mehrdad Asadi for the technical discussions around the topic.

References

[1] N. Mohamed *et al.*, "Unmanned aerial vehicles applications in future smart cities," *TFSC*, vol. 153, p. 119293, 2020.

[2] K. Kuru *et al.*, "A framework for the synergistic integration of fully autonomous ground vehicles with smart city," *IEEE Access*, vol. 9, pp. 923–948, 2020.

[3] E. Frachtenberg, "Practical drone delivery," *Computer*, vol. 52, no. 12, pp. 53–57, 2019.

[4] Y. Fu *et al.*, "A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance," *IEEE T-ITS*, 2021.

[5] Z. Yin *et al.*, "Toward city-scale litter monitoring using autonomous ground vehicles," *IEEE Pervasive Comput.*, 2022.

[6] A. Taïk *et al.*, "Data-quality based scheduling for federated edge learning," in *2021 IEEE LCN*. IEEE, 2021, pp. 17–23.

[7] S. Li *et al.*, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE T Depend and Secure*, vol. 18, no. 5, pp. 2088–2105, 2020.

[8] I. Shumailov *et al.*, "Sponge examples: Energy-latency attacks on neural networks," in *2021 IEEE EuroS&P*. IEEE, 2021, pp. 212–231.

[9] Z. Fu *et al.*, "Remote attacks on drones vision sensors: An empirical study," *IEEE T Depend Secure*, vol. 19, no. 5, pp. 3125–3135, 2021.

[10] R. A. Aral *et al.*, "Classification of trashnet dataset based on deep learning models," in *IEEE BigData*. IEEE, 2018, pp. 2058–2062.

[11] M. T. Ribeiro *et al.*, "" why should i trust you?" explaining the predictions of any classifier," in *ACM SIGKDD*, 2016, pp. 1135–1144.

[12] S. M. Lundberg *et al.*, "A unified approach to interpreting model predictions," *Adv Neural Inf Process Syst*, vol. 30, 2017.

[13] M. D. Zeiler *et al.*, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.

[14] D. Gunning *et al.*, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[15] J. M. Wing, "Trustworthy ai," *Communications of the ACM*, vol. 64, no. 10, pp. 64–71, 2021.

[16] T. Wu *et al.*, "Testing artificial intelligence system towards safety and robustness: State of the art." *IJCS*, vol. 47, no. 3, 2020.

[17] A. Wojciechowska *et al.*, "Designing drones: Factors and characteristics influencing the perception of flying robots," *Proceedings of IMWUT 2019*, vol. 3, no. 3, pp. 1–19.

[18] M. Anneken *et al.*, "Anomaly detection and xai concepts in swarm intelligence," 2021.

[19] M. K. Shehzad *et al.*, "Artificial intelligence for 6g networks: Technology advancement and standardization," *IEEE Veh Technol Mag*, 2022.

[20] D. Kaur *et al.*, "Trustworthy artificial intelligence: a review," *ACM CSUR*, vol. 55, no. 2, pp. 1–38, 2022.

Biographies

Abdul-Rasheed Ottun is a PhD-student at the University of Tartu. Research: AI, XAI, Trustworthy AI and sensing. Email:ottun@ut.ee.

Adeyinka Akintola is a PhD-student at the University of Tartu. Research: AGVs and smart-plants. Email:adeyinka@ut.ee.

Mohan Liyanage is a Lecturer at the University of Tartu. Research: IoT, edge computing and networking. Email:mohan.liyanage@ut.ee

Michell Boerger is a Research Scientist for the Fraunhofer FOKUS. Research: AI, XAI, and Cybersecurity. E-mail:michell.boerger@fokus.fraunhofer.de

Pan Hui is a Professor at HKUST and University of Helsinki. Research: Mobile computing, opportunistic networks, and AI. E-mail:pan.hui@helsinki.fi

Nikolay Tcholtchev is a researcher in Fraunhofer FOKUS. Research: Smart-Cities Cybersecurity, and AI. E-mail:nikolay.tcholtchev@fokus.fraunhofer.de

Sasu Tarkoma is a Professor at the University of Helsinki. Research: AI, data science, and sensing systems. E-mail:sasu.tarkoma@helsinki.fi

Petteri Nurmi is a Professor at the University of Helsinki. Research: Distributed systems, pervasive data science, and sensing systems. E-mail:petteri.nurmi@helsinki.fi

Huber Flores is an Associate Professor at the University of Tartu. Research: Distributed, mobile and pervasive computing systems. E-mail:huber.flores@ut.ee