Toward Trustworthy and Responsible Autonomous Drones in Future Smart Cities

TOWARD TRUSTWORTHY AND RESPONSIBLE DEPLOYMENT OF AUTONOMOUS DRONES IN FUTURE SMART CITIES

Abdul-Rasheed Ottun
Institute of Computer Science
University of Tartu
Tartu, Estonia.
ottun@ut.ee

Zhigang Yin
Institute of Computer Science
University of Tartu
Tartu, Estonia.
zhigang.yin@ut.ee

Mohan Liyanage
Institute of Computer Science
University of Tartu
Tartu, Estonia.
mohan.liyanage@ut.ee

Michell Boerger
Fraunhofer Institute for Open Communication Systems
Berlin, Germany.
michell.boerger@fokus.fraunhofer.de

Mehrdad Asadi
Institute of Computer Science
University of Tartu
Tartu, Estonia.
mehrdad.asadi@ut.ee

Pan Hui
Department of Chemistry
University of Helsinki
Helsinki, Finland
pan.hui@helsinki.fi

Sasu Tarkoma
Department of Computer Science
University of Helsinki
Helsinki, Finland
sasu.tarkoma@cs.helsinki.fi

Nikolay Tcholtchev
Fraunhofer Institute for Open Communication Systems
Berlin, Germany.
nikolay.tcholtchev@fokus.fraunhofer.de

Petteri Nurmi
Department of Computer Science
University of Helsinki
Helsinki, Finland
petteri.nurmi@cs.helsinki.fi

Huber Flores
Institute of Computer Science
University of Tartu
Tartu, Estonia.
huber.flores@ut.ee

November 1, 2022

ABSTRACT

Autonomous drones are reaching a level of maturity when they can be deployed in cities to support tasks ranging from medicine or food delivery to environmental monitoring. These operations rely on powerful AI models integrated into the drones. Ensuring these models are robust is essential for operating in cities as any errors in the decisions of the autonomous drones can cause damage to the citizens or the urban infrastructure. We contribute a research vision for trustworthy city-scale deployments of autonomous drones. We highlight current key requirements and challenges that have to be fulfilled for achieving city-scale autonomous drone deployments. In addition, we also analyze the complexity of using XAI methods to monitor drone behavior. We demonstrate this by inducing changes in AI model behavior using data poisoning attacks. Our results demonstrate that XAI methods are sensitive enough to detect the possibility of a data attack, but a combination of multiple XAI methods is better to improve the robustness of the estimation. Our results also suggest

that currently, the reaction time to counter an attack in city-scale deployment is large due to the complexity of the XAI analysis.

***K*eywords** Autonomous vehicles · Cooperative Robots · Network Swarm

1 Introduction

The integration of AI into drones is critical for enabling autonomous operations that require minimal or no human intervention. Indeed, AI is essential for navigation, trajectory estimation, collision avoidance, and localization [1] to name but some examples. As these techniques have matured, drone applications that automate our daily life activities have become a reality, e.g., delivery of food or medicine and applications that harness drones for environmental purposes, e.g., air quality monitoring [2], litter detection and separation [3]; and water pollution monitoring [4]. The emergence of these application domains has started to pave the way toward city-scale deployments of autonomous drones, yet there still are challenges that need to be overcome before these deployments can become a reality.

City-scale deployments of autonomous drones, such as ground drones, cars or other vehicles, aerial drones, or even aquatic drones, are only possible if the operations of the drones are *trustworthy*, i.e., they can be guaranteed to operate safely without causing harm to the citizens, the urban infrastructure, or the environment. Ensuring trustworthiness, however, is a complex challenge as it requires testing a wide range of environments and dependencies between system components [5]. Traditionally trustworthiness has been analyzed using formal verification methods, but these are only suitable for systems that expose their internal logic. Modern autonomous drones largely rely on deep learning models which are black-box in nature, making it difficult to verify the decisions that are made [6]. The emergence of federated learning and other paradigms that perform continuous updates on the model becomes even more complicated, as continuous updates can significantly change the behaviour of the AI models.

Explainable AI (XAI) methods have recently emerged as a mechanism that can help to understand the behaviour of AI models and the factors affecting them. For example, XAI methods can be used to evaluate the quality of the data used for training [7] or to assess the suitability of the internal model structure [8]. While these methods provide a foundation for analyzing the trustworthiness of the AI models integrated onto drones, they alone are insufficient as it is difficult to distinguish between multiple factors that have a similar effect on the model. For example, a decrease in model performance may result from a malfunction in the drone, a targeted attack, or a situational data bias. XAI methods also require heavy instrumentation of the model, and they tend to rely on offline analysis of the model, making these methods unsuitable for ongoing deployments where threats need to be detected immediately to ensure the safe operation of the drones.

We contribute a research vision for *trustworthy city-scale deployments of autonomous drones*. The vision, illustrated in Figure 1, assumes autonomous drones are widely used for societal and other functions and that they are instrumented with methods that enable diagnosing the AI and detecting threats. We first reflect on the current research landscape to identify gaps and open challenges to establish a roadmap that serves as a catalyst for research. We then experimentally demonstrate the importance of trustworthiness by conducting experiments on the effects of data poisoning on autonomous functionality. Finally, we conduct a small-scale field test in ground drone based litter monitoring to analyze the performance of different XAI methods in identifying data poisoning attacks. The results demonstrate that adverse effects on the model can be identified but that the root cause may not be easy to undercover. We conclude the article by highlighting further challenges and future directions for research.

2 Key Challenges and Requirements

Realising the vision of large-scale deployments of trustworthy autonomous drones is currently infeasible as there are technical and technological challenges that need to be addressed. We next highlight some important challenges and reflect on state-of-the-art to identify research gaps.

Data bias and drift detection Drones that are deployed should continue learning over time to improve their operation , e.g., by taking advantage of federated learning [7]. Optimally this process should integrate data from drones operating in different parts of the environment as this helps improving the generality and robustness of the AI models. Unfortunately, the model is vulnerable to biases in the data and it can be a target of attacks that affect the drone's operations. As a result, there is a need for methods that can quantify the resilience of the model updates and detect abnormal or erroneous updates before they affect the drone's functionality. A key challenge is to separate 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 separate between the different causes. Another challenge is to

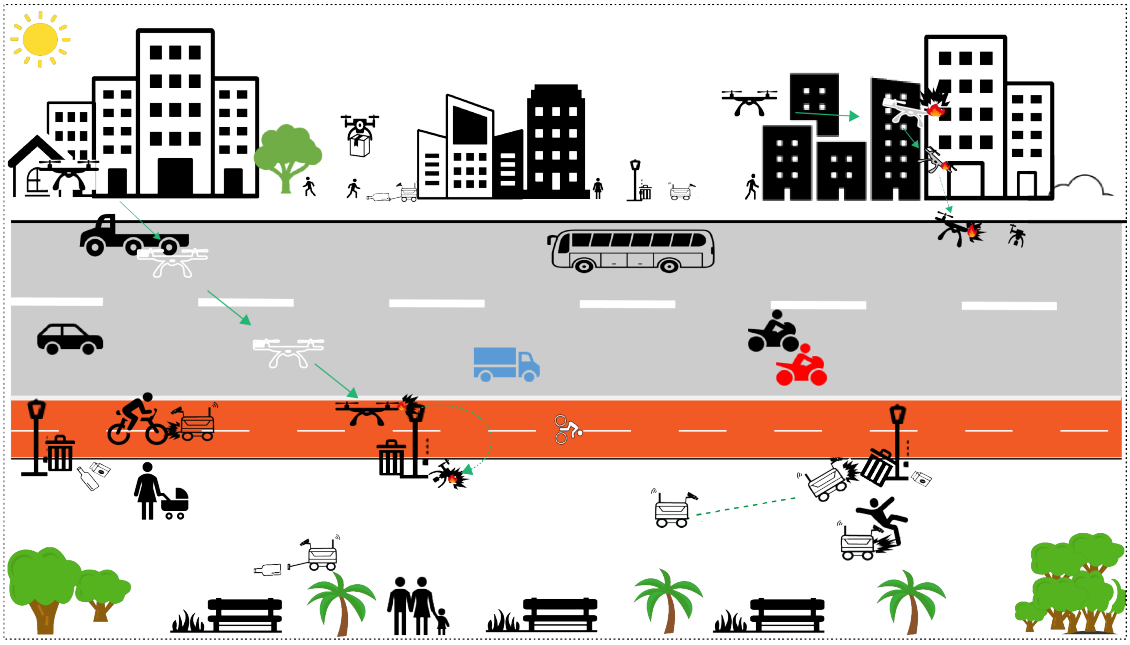


Figure 1: City-scale deployment of autonomous drones and how drones can malfunction in urban settings.

ensure the methods can operate at different temporal scale, i.e., can identify problems even when biased or erroneous data is aggregated with valid data and when the erroneous data arrives gradually.

Continuous model verification Besides detecting issues in the data, trustworthy operations require analyzing and verifying the decisions the drones make [9]. For large-scale deployments (e.g., a city), the analysis needs to happen continuously and on-site as otherwise the effort needed for verification limits the scale of the deployments. Indeed, model diagnosis requires accessing the internal structure of the model, which typically requires instrumenting the source code of the drone, halting operations, and accessing the internal logic of the drone. This process usually requires taking the drone to a lab as accessing the internal logic requires bypassing internal security features. While XAI methods offer a partial solution for continuous verification, they similarly need access to data and the model structure. Hence, they cannot be adopted as a general solution. A partial solution is to integrate the XAI methods directly as part of the security features (e.g., as part of trusted execution environments), but also this poses its own challenges as the security features often limit available resources.

Model interpretability and resources The performance of AI models is intrinsically linked to the resources and components integrated on the drones [5]. Over time, these components need maintenance, or may be upgraded to improve the operations of the drone. These operations can affect the model and result in unexpected behavior. For example, integrating a higher resolution camera on the drone affects the dimensionality of the input data and may contain more detail than previously. This can require replacing the model or at least re-training it. In terms of interpretability, this requires linking model diagnostics with physical components of the drone and being able to analyze and interpret the effects individual physical components have on the model's decisions. Note that these changes do not necessarily affect the input data. For example, drones can operate using different payloads which affects the weight and resource consumption of the drone. This requires integrating physical configuration directly into the model diagnosis. This is essential also for detecting unsafe operation, e.g., detecting unsafe payloads. Current methods are insufficient as they are unable to link model behaviour with physical characteristics of the operating environment. For example, while changes in input data can be detected with current model diagnosis methods, they are unable to detect whether these changes are a result of the physical configuration or external interference.

Network-group diagnosis Effective large-scale operations of autonomous drones are likely to require cooperation between the drones as this is essential for reducing resource drain and ensuring optimal performance. Effective coordination results in dependencies between the AI models deployed on the different drones and understanding potential errors or threats requires analysing the combined logic of all drones working in tandem, e.g., swarm intelligence [10]. Current XAI and other model diagnosis techniques are tailored to analysing individual models and hence they can only be used if the drones have a global model that integrates the decision logic of all drones together. Note that this task is more complex than analysing the performance of individual drones as attacks or errors can affect only some

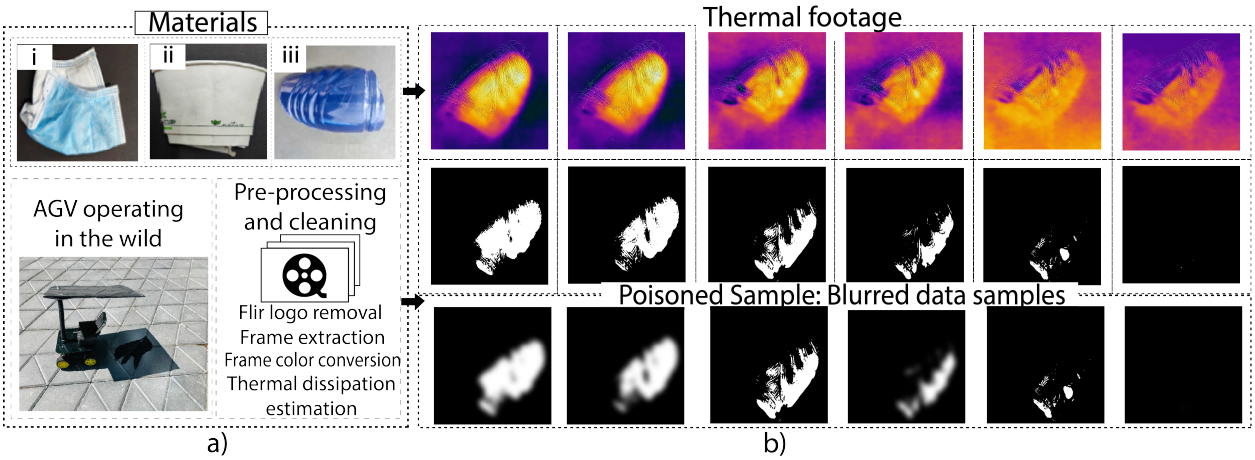


Figure 2: Autonomous ground drone for litter identification using sunlight and thermal imaging, a) Prototype sensing litter in the wild and litter materials used in the experiment; b) Analysis of thermal samples and estimation of thermal dissipation time.

of the drones, yet have an influence on all of the drones by compromising the coordination of the drones through the network [11]. Understanding the effects on coordination requires improved diagnosis mechanisms to analyze network formation groups, individual parts of the network (slices), as well as models of how targeted errors can affect the performance of drone collaboration.

3 Impact of Attacks on Autonomous Drones

Model diagnostics is essential not only for offering a mechanism to analyze and understand the behaviour of AI models but also for mitigating risks of external attacks that do not require access to the model itself. We next highlight the need for model diagnostics by drawing on an example from computer vision based object detection to demonstrate how external data poisoning attacks can result in abnormal model behaviour and potentially even break the AI performance.

Threat Model We consider a generic threat model where the attacker attempts to case the AI model used in the autonomous drone to fail. This can be a targeted attack that results in a specific misbehaviour, e.g., causing accidents by making the navigation support to fail to recognize pedestrians or cars, or an attack that simply causes the AI to malfunction, e.g., a sponge attack that drains the drone's resources or a ransomware attack that prevents normal operations. The motivation for the attack can be causing damage or harm to the city or the citizens, financial incentive, or desire for fame.

Application Scenario We consider litter recognition on autonomous drones as a representative example of drone applications that rely on AI. The application operates on thermal images which are analyzed in real-time to identify different litter objects and to determine their material. Specifically, the drone analyses the dissipation of sunlight-induced thermal radiation that is captured by a thermal camera integrated onto the drone [3]. Attacks against the model can break the operations of the drones or drain their resources. More serious attacks naturally would target navigation, obstacle detection, or other function that could directly result in harm to citizens or damage to the environment. Our use case presents a benign case that allows illustrating the risks of attacks without risking the citizens or the environment, and our findings are applicable to other AI applications that rely on computer vision.

Experimental Setup Figure 2 shows our testbed and illustrates the use of thermal dissipation to analyze the thermal dissipation fingerprints of materials. We consider three common litter objects with different materials in our experiment: (A) Plastic bottle, (B) Face mask and (C) Cardboard cup. The drone records video footage of disposed litter which is pre-processed and analyzed to identify litter [3]. To attack the model, we poison the data using blurring [7].

Results We first calculate the thermal dissipation time of each material without attacks: plastic bottle 62.5s, cardboard cup 72.5s, and face mask 82s. The relative differences match those reported in [3] for the same materials. The absolute values differ due to the different intensity of thermal source, size of the material and the total exposure time. To analyze poisoning, we consider two levels of poisoning: 10% (low) and 40% (high). Higher values 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 that is used for

detecting litter materials. We also observed a clear increase in the processing time of the drone, thus resulting in higher resource drain. We also separately tested a stenography attack to poison the data. In this case the thermal dissipation times were not influenced, i.e., the model is resilient against this attack. This highlights how the model response may vary depending on the attack type.

4 XAI as Model Diagnostics

xplainable AI (XAI) methods provide a natural starting point for integrating model diagnostics on the drones and to overcome the effects of attacks. We next analyze different the potential of different XAI methods to detect targeted poisoning attacks, focusing on understanding the benefits and disadvantages of different methods. We first examine the quantifiable values provides by XAI methods in benign case, after which we analyze them against poisoned data. We separately analyze the full image and a processed image where the background is removed to better understand how different processing techniques affect the behaviour of XAI methods.

Experiment Setup We perform the experiment using the TrashNet litter classification dataset which consists of 2527 litter images [12]. We rely on this dataset as it contains a large amount of real-world images which makes it possible to analyze different environments and contexts for litter classification. As AI model we consider a convolutional deep learning model (CNN) as this has been shown to achieve good performance on the dataset [12]. Images are resampled to 300×300 to have consistent input dimensionality. We augment the training data using horizontal and vertical flipping and rescaling. We train the dataset using 2276 images with batch size of 32 for each epoch iteration. The remaining images are used for testing and we separately consider a collection of 10 poisoned and non-poisoned images for illustrating the performance of XAI methods. Our 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 We consider three model-agnostic XAI methods that can be applied for any type of AI model: LIME (Local the Interpretable Model-agnostic Explanation), SHAP (Shapley Addictive Explanations) and Occlusion sensitivity. As these methods are model-agnostic, they do not require any information about the CNN gradients to analyze model behaviour. These methods are also perturbation-based, which means that they manipulate the input (i.e., image pixels) to extract details that can be linked with the predictions. LIME creates an interpretable representation of the litter image by segmenting the image pixels, based on similarity, into superpixels [13]. SHAP explains the CNN's prediction by attributing importance values to the features that contributed to a prediction. This is accomplished using a binary vector of simplified inputs by perturbing the pixels (i.e. input space) of the litter image into superpixels (features) that represent the image [14]. The features that contribute to the probability of the prediction are highlighted in one color (red) and the features that decreases the probability of the predicted class in another color (blue). Finally, occlusion sensitivity explains predictions by using a sensitivity heat map to observe the impact a perturbation mask has on the neural network predictions. Specifically, an occlusion mask (i.e small gray square patch) is placed on the image and changes in the prediction probabilities are observed [15]. As stated, the XAI methods are applied separately for images where the background is removed (i.e., only the litter object) and for the original input image. The process for extracting the object is shown in Figure 3 and works by applying a dynamic patch (determined using object detection) on the image to isolate it. From the final output of the XAI methods, we calculate a pixel percentage metric that captures the importance of a pixel.

Samples and Poisoning We considered six litter categories: glass, paper, cardboard, trash, metal and plastic. For poisoning we consider two attacks, blurring and steganography, as describe in the previous section. Blurring can make autonomous drones to misidentify targets in urban areas, e.g., crossing signals and pedestrian sides. 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). We systematically assess how the level of poisoning affects the results by poisoning the data in 10% increments from 10% to 40%.

5 Result

Model Performance under Poisoning The performance to classify litter of our CNN is 0.7 when no data is poisoned, but this performance is reduced as data is gradually poisoned. After blurring attack, the model accuracy is reduced to 0.61 (10% poisoned); 0.53 (20% poisoned); 0.53 (30% poisoned) and 0.60 (40% poisoned). Similarly, after steganography, the model accuracy is reduced to 0.52 (10% poisoned); 0.52 (20% poisoned); 0.62 (30% poisoned) and 0.67 (40% poisoned). In both cases, we observe a clear drop in accuracy. Unlike our earlier experiment, the performance drop is 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 and highlights how the effectiveness of the attack

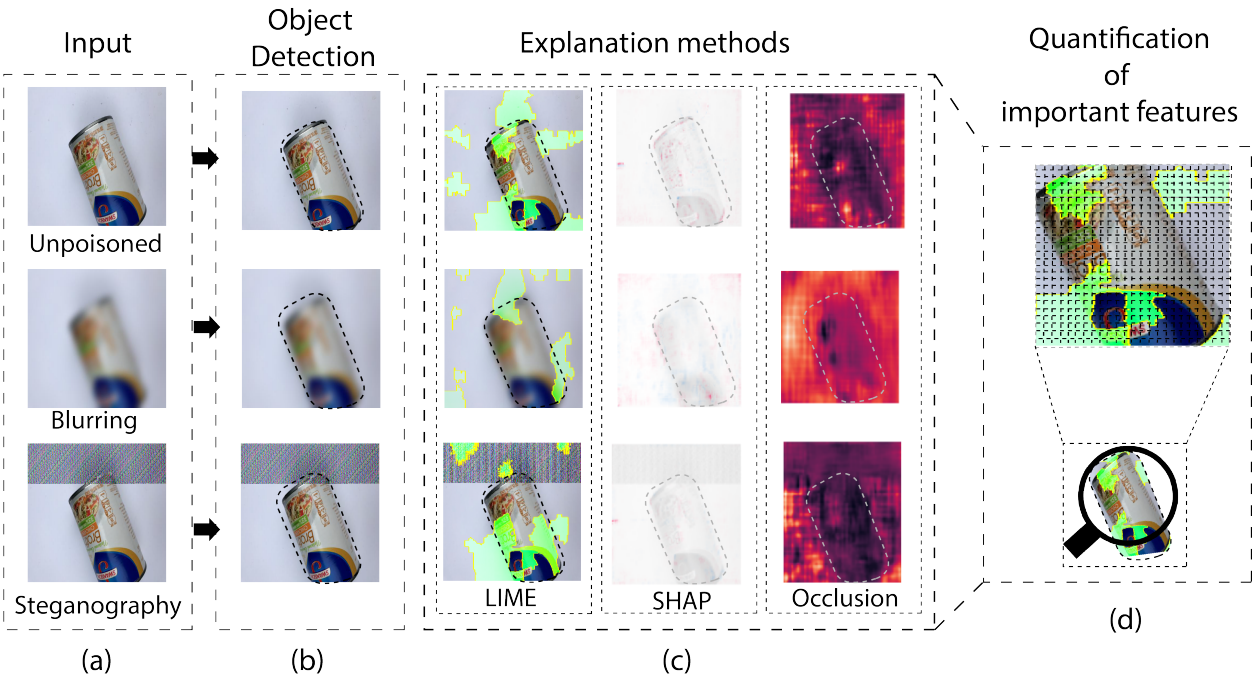
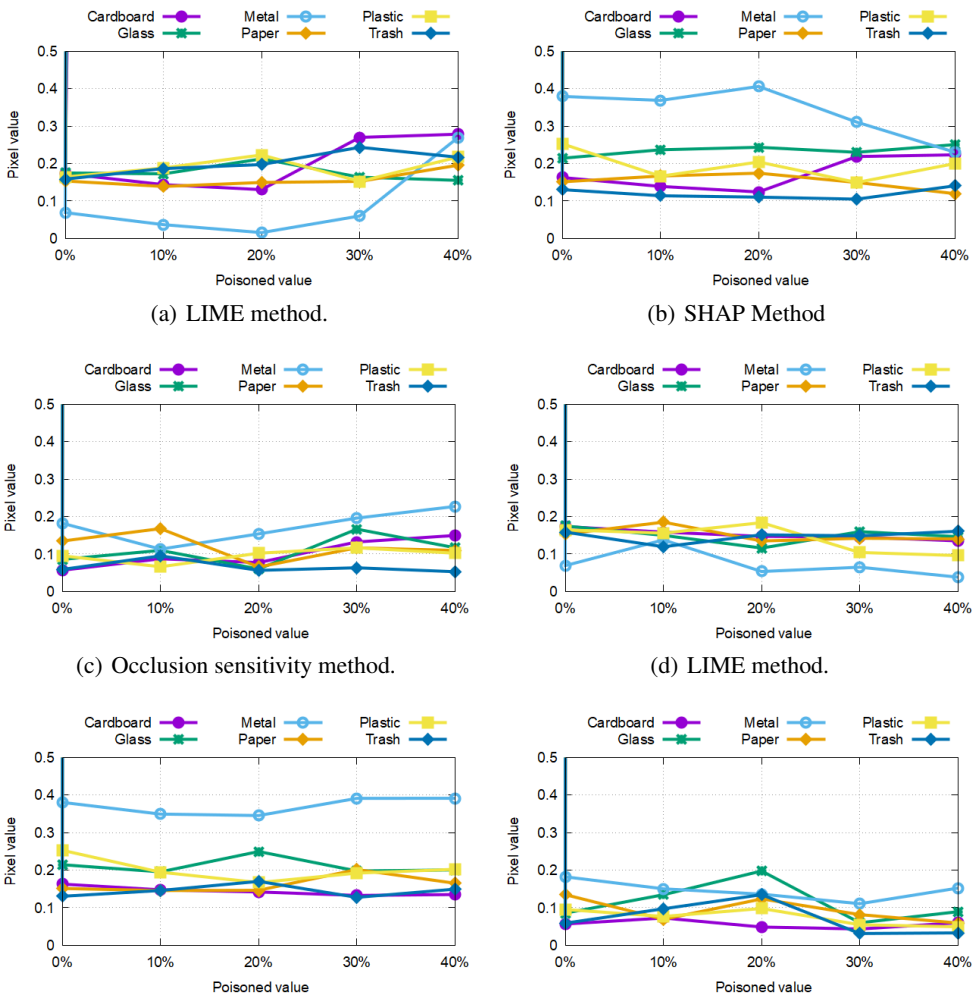


Figure 3: Pipeline to solely analyze objects using XAI methods, a) Data samples (poisoned and unpoisoned), b) Object detection, c) XAI methods output over samples (LIME, SHAP and Occlusion sensitivity) and d) Object extraction.

is influenced by the task and the specifics of the AI that is being used. The performance drop resulting from poisoning depends on how much the attack affects the patterns in the data. In general, once larger amounts of the data become poisoned, the inference process starts to be dominated by the poisoned patterns whereas in smaller amounts they result in distortions that can confuse the model. This pattern is observed with both attacks with the sole exception being blurring at 10% rate. The reason for this exception stems from small amount of blurring failing to distort the patterns of the litter object.

Analysis of XAI Methods We analyze the effectiveness of XAI methods by considering 10 randomly chosen poisoned samples from each litter category and report the accuracy of estimating the correct class for each sample. Table 1 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 shape are easiest for the XAI methods, whereas classes containing irregular shapes (metal, plastic, trash) showing highest variation in results. As with the results for the CNN model, in some cases a higher level of poisoning can result in smaller drop – or in some cases even in 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. The robustness of the model can also be improved by incorporating transformations that resemble the poisoned data as part of the training data. For example, blurring is a commonly used data augmentation technique for improving the training of AI models. From our experiments, we also visually observed that the attacks tend to affect more the background and thus processing techniques that separate the foreground object from background are likely to improve performance.

Diagnosing Objects with XAI Lastly, we examine the effect of data poisoning over the important features of the object when it is isolated from the background. As the metric we consider the coefficient of variation of the poisoned pixels, which depicts the ratio of the standard deviation to the mean. The higher the value of the coefficient, the higher the dispersion and thus the better the method is at identifying poisoned data. Figure **??** shows the results for the 10 test samples of each class. 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 generally are better at identifying poisoned data. SHAP performs well for metal objects which are the most irregular, but struggles with other categories. We also used one-way ANOVA test 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



(a) LIME method. (b) SHAP Method

(c) Occlusion sensitivity method. (d) LIME method.

(e) SHAP method. (f) Occlusion sensitivity method.

Figure 4: Object analysis with each XAI method as data is poisoned with, (a-c) Blurring and (d-f) Steganography.

coefficients of variation are larger for steganography than for blurring indicating that XAI methods can also provide clues about the nature of the error. We also investigated the effect between attack type and data poison level. Two-way ANOVA test between attack type and data poisoning level indicates 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 but 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, this requires more processing and more elaborate processing pipelines which drains the resources of the drone faster and limits their operations.

6 Discussion

Stakeholders Model diagnostics are particularly important for the companies and organizations that operate drones in urban settings. At the same time, municipalities and governmental authorities can require diagnostics to be integrated onto the drones before they issue permits as this can help ensure the drones are not vulnerable to targeted attacks. We would expect diagnostics to eventually become a legal requirement as well as this allows auditing the drone behaviour when accidents or other harmful behaviours occur.

Improvement As room to improve our work, it would be important to study also other drone functionalities, such as navigation, localization, or collision avoidance. As these functions can result in dangerous behaviours, it is essential to

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

design experiments that minimize risks, as well as to obtain the necessary ethical and legal permissions to carry out the experiments. We are also interested in exploring other environmental monitoring use cases, such as air quality [2] or water pollution monitoring [4]. These are examples of domains where drones could be used for emission accounting and thus there would be financial incentives to influence the performance of the AI models.

Practical Limitations Attacks on AI models are not the only aspect that changes the behavior of autonomous drones, as hardware failures and software malfunctions can also affect their behavior. This requires methods that can operate on the drones and differentiate in real-time between targeted attacks and other errors. Another essential aspect is to integrate the drone with rollback mechanisms, return to home protocols, or other procedures that allow them to react to situations where erroneous data or other abnormalities are observed. Exploiting infrastructure backdoors is another potential way to attack drones.

End-User Support Our work has demonstrated how changes in data can affect predictions and how XAI methods can be used to identify which parts of the data contributed to a given prediction. While this can help to detect abnormal behaviors, the output of the XAI methods can be difficult to interpret for end-users that analyze the drone behavior (e.g., technicians or drone operators). Further work is thus needed to develop tools and mechanisms that can translate the results of the XAI methods into insights that end-users that can use to fix potential problems.

Debugging Models Optimally the AI models could be debugged in the wild semi-autonomously. This, however, would require dedicated support mechanisms on the drone which may be difficult to implement. For example, detecting points of deviation would require maintaining a reference model and re-training the model updates by replaying individual data points. The model could then be analyzed with XAI methods after each iteration to identify potential abnormalities in the data and to support the detection of issues. This process easily becomes highly resource intensive and requires sufficient memory and physical storage on the drones. In parallel, there is a need to verify the integrity of the reference models and XAI tools, which requires dedicated mechanisms such as trusted execution environments.

7 Summary and Conclusions

Powerful AI models are important for enabling autonomous operations of drones, particularly in complex and highly varying environments such as cities. These models need to be accurate and perform robustly as otherwise the drones can cause harm to citizens or damage the environment. We presented a research vision of how to enable the AI models to be trustworthy, identifying key challenges and requirements, and arguing that methods for model diagnosis should be integrated to drones to verify that the operations are indeed safe. We experimentally demonstrated the importance of model diagnosis by showing how targeted attacks, such as data poisoning, can break the AI models integrated on the drones. What makes these attacks particularly problematic is that they do not require access to the device or the AI model, operating solely by manipulating the inputs that it uses. We also demonstrated that explainable AI (XAI) methods provide a foundation for identifying issues but that they are also subject to limitations. In particular, XAI methods are unable to distinguish between targeted attacks and other malfunctions, their performance depends on the sensor modality, the environment, and the characteristics of the input data.

References

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

[2] Naser Hossein Motlagh et al. Toward blue skies: City-scale air pollution monitoring using uavs. *IEEE Consum. Electron. Mag*, 2022.

[3] Zhigang Yin, Mayowa Olapade, Mohan Liyanage, Farooq Dar, Agustin Zuniga, Naser Hossein Motlagh, Xiang Su, Sasu Tarkoma, Pan Hui, Petteri Nurmi, and Huber Flores. Toward city-scale litter monitoring using autonomous ground vehicles. *IEEE Pervasive Comput.*, 2022.

[4] Huber Flores, Naser Hossein Motlagh, Agustin Zuniga, Mohan Liyanage, Monica Passananti, Sasu Tarkoma, Moustafa Youssef, and Petteri Nurmi. Toward large-scale autonomous marine pollution monitoring. *IEEE IoT Mag*, 4(1):40–45, 2021.

[5] Anna Wojciechowska et al. Designing drones: Factors and characteristics influencing the perception of flying robots. *Proceedings of IMWUT 2019*, 3(3):1–19.

[6] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box ai decision systems. In *AAAI Artif Intell.*, volume 33, pages 9780–9784, 2019.

[7] Afaf Taïk, Hajar Moudoud, and Soumaya Cherkaoui. Data-quality based scheduling for federated edge learning. In *2021 IEEE LCN*, pages 17–23. IEEE, 2021.

[8] Plamen Angelov and Eduardo Soares. Towards explainable deep neural networks (xdnn). *Neural Networks*, 130:185–194, 2020.

[9] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[10] Mathias Anneken, Manjunatha Veerappa, and Nadia Burkart. Anomaly detection and xai concepts in swarm intelligence. 2021.

[11] Muhammad K Shehzad et al. Artificial intelligence for 6g networks: Technology advancement and standardization. *IEEE Vehicular Technology Magazine*, 2022.

[12] Rahmi Arda Aral et al. Classification of trashnet dataset based on deep learning models. In *IEEE BigData*, pages 2058–2062. IEEE, 2018.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144, 2016.

[14] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*, 30, 2017.

[15] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.