Explainable artificial intelligence for cyber-resilient smart cities: enhancing trust, transparency, and adversarial robustness

Vlad Stanescu,
Beia Consult International, Bucharest, Romania
vlad.stanescu@beia.ro

Maria Niculae,
Beia Consult International, Bucharest, Romania
maria.niculae@beia.ro

George Suciu,
Beia Consult International, Bucharest, Romania
george@beia.eu

Mari-Anais Sachian,
Beia Consult International, Bucharest, Romania
anais.sachian@beia.ro

Abstract
The rapid integration of Artificial Intelligence (AI) into smart city infrastructures has transformed urban services, enabling intelligent transportation, energy optimization, predictive governance, and advanced cybersecurity mechanisms. However, the increasing complexity of AI-driven systems has also introduced critical vulnerabilities related to adversarial attacks, opaque decision-making, and reduced public trust. Within the context of the AIAS project, which focuses on AI-assisted cybersecurity and adversarial AI resilience, this paper investigates the role of Explainable Artificial Intelligence (XAI) in strengthening cyber-resilient intelligent cities. The research explores how explainability mechanisms can improve the transparency, accountability, and interpretability of AI-based cybersecurity systems operating in smart urban environments. The study examines state-of-the-art XAI techniques, including feature attribution methods, surrogate models, attention visualization, and human-centered interpretability frameworks, applied to cyber threat detection and anomaly monitoring systems. Particular attention is given to adversarial AI scenarios, where malicious manipulation of machine learning models may compromise decision integrity and urban critical infrastructure. The paper proposes an explainability-oriented framework capable of supporting both "AI for Cybersecurity" and "Cybersecurity for AI" approaches by enabling security analysts and public administration stakeholders to better understand automated decisions, identify anomalies, and validate threat intelligence outputs. The framework aims to reduce the "black-box" nature of deep learning systems while supporting compliance with emerging regulatory requirements regarding trustworthy and ethical AI governance. Expected outcomes include improved trust in AI-enabled public services, enhanced incident response capabilities, and increased resilience of smart city ecosystems against adversarial threats and information sabotage. Furthermore, the research contributes to the broader discussion on ethical AI deployment in public administration by emphasizing transparency, human oversight, and responsible data governance. The study aligns with current European priorities on trustworthy AI and cyber resilience, supporting the development of secure, explainable, and citizen-centered intelligent urban infrastructures.

Keywords: XAI, Cybersecurity, Deep Learning, Trustworthy AI

1. Introduction
Artificial Intelligence (AI) has become an essential component of contemporary smart city infrastructures, supporting the development of intelligent transportation systems,

automated public services, energy optimization, predictive governance, and advanced cybersecurity mechanisms. As urban environments become increasingly digitalized and interconnected, AI-based systems are expected to improve efficiency, responsiveness, and decision-making in public administration and critical infrastructure management [1, 2]. In this context, smart cities are no longer defined only by the deployment of digital technologies, but also by their capacity to process data, anticipate risks, and provide secure, transparent, and citizen-centered services.

However, the growing dependence on AI-driven systems also introduces new cybersecurity and governance challenges. Smart city infrastructures rely on complex cyber-physical systems, Internet of Things devices, communication networks, cloud services, and automated decision-making tools. These components expand the attack surface and create opportunities for cyber threats such as data manipulation, malware attacks, information sabotage, unauthorized access, and adversarial attacks against machine learning models [3, 4]. In particular, adversarial AI represents a significant concern, as malicious actors may attempt to manipulate input data or exploit model vulnerabilities in order to degrade the performance, reliability, or integrity of AI-based systems.

A major limitation of many AI and deep learning models used in cybersecurity is their black-box nature. Although these models can achieve strong performance in classification, anomaly detection, and threat identification tasks, their internal decision-making processes are often difficult to interpret. This lack of transparency can reduce trust among security analysts, public administration stakeholders, and decision-makers, especially when AI systems are deployed in sensitive domains such as smart city security, public services, and critical infrastructure protection. Explainable Artificial Intelligence (XAI) addresses this limitation by providing methods that make AI decisions more understandable, traceable, and accountable [5].

This paper investigates the role of XAI in strengthening cyber-resilient smart cities by focusing on the interpretability of AI-based cybersecurity systems. The main objective is to explore how explainability mechanisms can support both “AI for Cybersecurity”, by improving threat detection and analysis, and “Cybersecurity for AI”, by increasing the transparency and robustness of AI systems themselves. To support this objective, the paper presents an initial proof-of-concept experiment based on malware image classification using the MaleVis dataset, a ResNet18 transfer learning model, and Grad-CAM visual explanations.

The contribution of this work is threefold. First, it discusses the relevance of explainability for cybersecurity in smart city environments, where trust, transparency, and human oversight are essential. Second, it proposes an explainability-oriented workflow that connects malware classification with visual interpretation of model decisions. Third, it validates the feasibility of this workflow through an initial experimental implementation, showing how Grad-CAM can highlight image regions that contribute to malware classification outputs. The study therefore provides an initial step toward more transparent, interpretable, and resilient AI-based cybersecurity systems for intelligent urban infrastructures.

2. Explainable AI for Cybersecurity

2.1 AI and Cybersecurity Challenges in Smart Cities

Smart cities rely on highly interconnected digital infrastructures that combine sensors, Internet of Things devices, communication networks, cloud platforms, public administration systems, and AI-based decision-support tools. These technologies enable more efficient urban services, including intelligent mobility, energy management, public safety, environmental monitoring, and predictive governance. At the same time, the increasing digitalization of urban infrastructures makes cities more dependent on the availability, integrity, and confidentiality of data. As a result, cybersecurity becomes a core requirement for the safe operation of smart cities, rather than a secondary technical concern [6, 7].

Artificial Intelligence has become increasingly relevant in cybersecurity because it can support the detection of anomalies, malware, suspicious network behavior, and emerging attack patterns at a scale that is difficult to manage through traditional rule-based systems. AI-based threat detection mechanisms can process large volumes of heterogeneous data and identify correlations that may remain hidden to human analysts. Recent studies in the SCRD and TRUST research ecosystem emphasize that AI can improve cyber defense by enabling automation, adaptive detection, and faster response to complex threats [3, 8]. However, the same capabilities that make AI useful for defense can also be exploited by malicious actors, for example through adaptive malware, automated phishing, synthetic identities, and adversarial manipulation of machine learning models [8].

These risks are particularly relevant for smart cities, where AI-supported decisions may influence public services, critical infrastructure, mobility systems, and citizen-facing applications. In such environments, the performance of an AI model is not sufficient on its own; decision-makers and cybersecurity analysts also need to understand why a system has produced a specific output. The lack of transparency in complex machine learning and deep learning models may reduce trust, complicate incident response, and limit accountability. Therefore, explainability becomes an essential requirement for AI-based cybersecurity systems, especially in high-impact public sector environments where human oversight, regulatory compliance, and trust governance are necessary conditions for responsible deployment [9, 10].

2.2 Explainable AI Approaches for Threat Detection

Explainable Artificial Intelligence has become increasingly important in cybersecurity because many modern detection systems rely on complex machine learning and deep learning models whose decisions are difficult to interpret. In threat detection scenarios, explanations are necessary not only to improve trust, but also to support analysts in understanding why a specific event, file, network behavior, or anomaly was classified as malicious. Recent studies emphasize that XAI can be applied across multiple cybersecurity domains, including intrusion detection, malware detection, phishing detection, fraud detection, digital forensics, and zero-day vulnerability analysis [11, 12]. In these contexts, explainability can help transform AI outputs from simple alerts into actionable information that can be assessed by human experts.

Several XAI methods have been proposed to improve the interpretability of AI-based models. LIME explains individual predictions by approximating the local behavior of a complex model with an interpretable surrogate model [13], while SHAP assigns contribution values to input features based on game-theoretic principles [14]. These methods are particularly useful in tabular or feature-based cybersecurity tasks, such as intrusion detection, anomaly detection, and risk scoring. In contrast, visual explanation methods such as Grad-CAM are especially relevant for convolutional neural networks, as they generate class-discriminative heatmaps that highlight the regions of an input image that contributed most to a model prediction [5]. This makes Grad-CAM suitable for image-based malware classification, where binary files are converted into visual representations and analyzed using deep learning models.

Other explainability approaches include attention visualization and intrinsically interpretable machine learning models. Attention-based mechanisms can indicate which parts of an input sequence, log entry, or feature representation receive greater importance during prediction, which is useful for threat analysis and security monitoring workflows. Interpretable models, such as decision trees, rule-based systems, or generalized additive models, can also be used when transparency is more important than maximum predictive performance. In operational environments such as Security Operations Centers, these techniques may support alert prioritization, incident triage, analyst decision-making, and model auditing. Nevertheless, existing research also shows that explainability in cybersecurity remains challenging, because explanations must be accurate, understandable to analysts, robust against adversarial manipulation, and suitable for real-time use [12, 15].

Within this context, the present work focuses on visual explainability for malware classification as an initial proof-of-concept. The use of Grad-CAM is motivated by the image-based nature of the selected malware dataset and by the need to provide interpretable visual evidence for deep learning decisions. Rather than proposing a new malware classification architecture, this paper investigates how an explainability layer can be integrated into an AI-based cybersecurity workflow in order to support transparency, trust, and human oversight in cyber-resilient smart city environments.

2.3 Research Motivation

Although AI-based cybersecurity systems have demonstrated strong potential for detecting threats, anomalies, and malicious behavior, their practical adoption in sensitive smart city environments is still limited by the difficulty of interpreting their decisions. In many cases, deep learning models can provide accurate classifications without offering sufficient information about the reasoning behind those outputs. This creates a gap between technical performance and operational trust, especially when the results of an AI system must be evaluated by cybersecurity analysts, public administration stakeholders, or decision-makers responsible for critical urban services.

Explainability is still insufficiently integrated into many cybersecurity workflows. Most AI-based threat detection systems focus primarily on predictive accuracy, while interpretability is often added only as a secondary component or omitted entirely. This

limits the ability of human experts to validate alerts, identify false positives, investigate suspicious patterns, or understand whether a model relies on meaningful features or on irrelevant correlations. In the context of adversarial AI, this limitation becomes even more important, as attackers may attempt to exploit hidden model weaknesses or manipulate inputs in ways that are difficult to detect without interpretability mechanisms.

Smart cities represent a particularly sensitive application domain because AI-based decisions may affect public services, mobility systems, administrative processes, critical infrastructures, and citizen trust. A cyber incident in such an environment may have consequences that extend beyond technical disruption, influencing safety, service continuity, institutional credibility, and public confidence. Therefore, cybersecurity systems deployed in smart city contexts should not only be accurate, but also transparent, auditable, and understandable to the human operators involved in the decision-making process.

For this reason, the present paper introduces an initial proof-of-concept focused on integrating explainability into an AI-based cybersecurity workflow. The experimental implementation uses malware image classification as a practical case study, combining a ResNet18-based deep learning model with Grad-CAM visual explanations. This approach is relevant because it demonstrates how visual explanation techniques can help interpret model decisions and support the development of more transparent cybersecurity tools. While the experiment represents an initial validation step, it provides a foundation for future work involving additional malware classes, other XAI methods, adversarial robustness evaluation, and integration into broader smart city security scenarios.

3. Proposed Framework and Experimental Setup

3.1 Explainability-Oriented Framework

The proposed framework is designed as an initial explainability-oriented workflow for AI-based cybersecurity analysis in smart city environments. Its purpose is to connect automated threat classification with visual interpretation mechanisms, allowing the output of a deep learning model to be analyzed not only in terms of prediction accuracy, but also in terms of transparency and interpretability. The framework follows a sequential pipeline composed of four main stages: data preparation, malware classification, model evaluation, and explainability generation.

In the first stage, malware samples are represented as images and prepared for deep learning processing. This approach enables the use of convolutional neural networks, which are suitable for identifying visual patterns and structural similarities in image-based malware representations. The input images are resized to a fixed resolution and converted into tensors before being passed to the classification model. This preprocessing step ensures consistency across samples and allows the model to process the dataset in a standardized format.

The second stage consists of malware classification using a convolutional neural network based on transfer learning. In the current proof-of-concept implementation, a ResNet18 architecture was selected due to its relatively low complexity, availability of pretrained

weights, and suitability for image classification tasks. The final classification layer was adapted to the number of malware classes included in the experiment. The model was then trained on a reduced subset of the MaleVis dataset in order to validate the end-to-end pipeline before extending the analysis to more complex scenarios.

The third stage evaluates the model using baseline performance indicators, such as training loss and validation accuracy. These metrics provide an initial view of the model's ability to distinguish between selected malware classes. However, performance metrics alone do not explain why the model assigns a specific class to a malware image. For this reason, the framework includes an additional explainability stage, which aims to make the model's decision process more interpretable.

The final stage integrates Grad-CAM as the first explainability method. Grad-CAM generates visual activation maps that highlight the image regions most relevant to the model's prediction. In the proposed workflow, these heatmaps are overlaid on the original malware images, allowing analysts to inspect which visual areas contributed most strongly to the classification result. This explainability layer supports the transition from a purely predictive model to a more transparent cybersecurity tool, where human analysts can better understand and validate AI-based decisions.

Overall, the framework demonstrates how malware classification and explainability can be combined into a single workflow. Although the current implementation represents an initial experimental step, it provides a practical foundation for future integration of additional XAI techniques, larger malware datasets, adversarial robustness testing, and broader smart city cybersecurity use cases.

3.2 Dataset and Preprocessing

The experimental implementation was carried out using the MaleVis malware visualization dataset, which provides malware samples represented as images. This type of representation allows binary malware files to be analyzed through visual patterns, making the dataset suitable for convolutional neural network-based classification. The use of malware images is also relevant for explainability analysis, as visual interpretation methods such as Grad-CAM can be applied directly to highlight image regions that influence the model's predictions.

For the current proof-of-concept, a reduced subset of malware classes was selected instead of using the full dataset. This decision was made in order to validate the complete processing pipeline before scaling the experiment to a larger number of malware families. The first experiment was conducted on two malware classes, while the second experiment extended the setup to multiple selected classes, including samples from families such as Adposhel, Agent, Allaple, and Amonetize. This incremental approach allowed the classification and explainability workflow to be tested under controlled conditions.

The dataset was organized into separate training and validation folders, following the directory structure required by the image classification pipeline. Each malware family was stored in a dedicated class folder, allowing automatic label assignment during dataset

loading. The images were loaded using an image folder-based procedure and transformed before being passed to the model.

The preprocessing stage consisted of resizing all input images to 224 × 224 pixels and converting them into tensor representations. This resolution was selected to match the input requirements of the ResNet18 architecture used in the experiment. No complex feature engineering was applied, as the objective was to evaluate the feasibility of an end-to-end deep learning and explainability workflow. The same preprocessing strategy was applied both during model training and during the Grad-CAM visualization stage in order to ensure consistency between classification and explanation outputs.

3.3 Deep Learning Model

The malware classification component of the proposed workflow was implemented using a convolutional neural network based on the ResNet18 architecture. ResNet18 was selected because it provides a good balance between classification capability and computational efficiency, making it suitable for an initial proof-of-concept experiment. Its residual connections help reduce the degradation problem that may appear in deeper neural networks, while its pretrained implementation allows rapid adaptation to image classification tasks.

A transfer learning strategy was adopted in order to reuse knowledge learned from large-scale image datasets and adapt it to the malware image classification task. The pretrained ResNet18 model was used as the backbone, while the final fully connected layer was replaced with a new classification layer corresponding to the number of malware classes included in each experiment. In the first experiment, the model was configured for binary classification, while in the second experiment the final layer was adapted dynamically according to the number of selected malware families.

The model was trained using the cross-entropy loss function, which is commonly applied in multi-class classification problems. The Adam optimizer was used with a learning rate of 0.001. The training process was performed for three epochs, using a batch size of 16. During training, the model received resized malware images as input and learned to associate visual patterns with their corresponding malware class labels.

After training, the model was evaluated on the validation subset. The validation process measured the number of correct predictions in relation to the total number of validation samples, producing an initial accuracy score. In the extended experiment, additional evaluation outputs were generated, including a classification report and a confusion matrix. These metrics were used to assess the baseline performance of the classifier before applying the explainability layer. However, the main objective of the model was not to achieve state-of-the-art malware classification performance, but to validate the feasibility of integrating deep learning-based detection with visual explanation mechanisms.

3.4 Grad-CAM Explainability Integration

After the malware classification model was trained and evaluated, Grad-CAM was integrated as the first explainability method in the proposed workflow. The objective of

this stage was to provide a visual interpretation of the model's predictions and to identify the image regions that contributed most strongly to the classification decision. This is particularly relevant for malware image classification, where the input is represented visually and where convolutional neural networks learn spatial patterns that are not directly interpretable by human analysts.

Grad-CAM was applied to the last convolutional layer of the ResNet18 model, corresponding to the final residual block. This layer was selected because it captures high-level spatial features that are relevant for class discrimination. For each selected malware image from the validation set, the trained model produced a prediction, and Grad-CAM generated a class activation map associated with that decision. The resulting heatmap highlights the areas of the image that had the greatest influence on the model output.

The generated activation maps were then overlaid on the original malware images in order to obtain visual explanations. Warmer regions in the Grad-CAM visualization indicate stronger contributions to the model prediction, while less highlighted regions indicate lower relevance. This allows the analyst to inspect whether the classifier focuses on meaningful visual structures rather than irrelevant or noisy areas of the image.

The integration of Grad-CAM transforms the output of the deep learning model from a simple class label into a more interpretable decision-support result. Instead of only knowing which malware class was predicted, the analyst can also observe where the model focused during classification. In a broader cybersecurity context, such explanations can support alert validation, model auditing, analyst trust, and human oversight. Although Grad-CAM does not provide a complete explanation of the internal model reasoning, it offers an initial and practical mechanism for improving transparency in AI-based malware detection workflows.

4. Experimental Results and Discussion

4.1 Classification Results

The experimental evaluation was conducted on the selected subset of the MaleVis dataset using the ResNet18-based transfer learning model described in the previous section. The model was trained for three epochs using a batch size of 16 and the Adam optimizer with a learning rate of 0.001. During training, the loss decreased progressively from 37.8946 in the first epoch to 21.2076 in the second epoch and 17.1617 in the third epoch. This downward trend indicates that the model was able to learn discriminative visual patterns from the selected malware image classes during the initial training phase.

After training, the model was evaluated on the validation subset, which contained 539 malware images from four selected classes: Adposhel, Agent, Allaple, and Amonetize. The model achieved a validation accuracy of 95.18%, suggesting that the selected convolutional neural network was able to distinguish between the malware families included in the proof-of-concept experiment. However, this result should be interpreted as an initial baseline rather than as a final performance benchmark, since the experiment was conducted on a reduced subset of malware classes and with a limited number of training epochs.

The classification report provides a more detailed view of the model performance across the selected classes. Adposhel obtained the strongest results, with precision, recall, and F1-score equal to 1.00. Amonetize also achieved high performance, with a precision of 0.99, recall of 0.93, and F1-score of 0.96. Allaple obtained a precision of 0.91, recall of 0.97, and F1-score of 0.94, while Agent achieved a precision of 0.89, recall of 0.91, and F1-score of 0.90. These results indicate that most classification errors were concentrated in a limited number of cases, especially between visually closer malware families.

Table 1 summarizes the classification performance obtained on the validation set.

Table 1. Classification report for the selected MaleVis classes

Class	**precision**	**recall**	**f1-score**	**support**
Adposhel	1.00	1.00	1.00	144
Agent	0.89	0.91	0.90	120
Allaple	0.91	0.97	0.94	128
Amonetize	0.99	0.93	0.96	147
accuracy			0.95	539
macro avg	0.95	0.95	0.95	539
weighted avg	0.95	0.95	0.95	539

Source: Authors' own processing based on the experimental results.

In addition to the classification report, the confusion matrix was generated in order to provide a visual overview of the model's prediction behavior across the selected malware classes. As shown in Fig. 1, the model correctly classified all Adposhel samples, while most misclassifications occurred between Agent, Allaple, and Amonetize. More specifically, 10 Agent samples were classified as Allaple, 4 Allaple samples were classified as Agent, and 9 Amonetize samples were classified as Agent. This suggests that these malware image classes may contain more visually similar patterns, making them more difficult to separate compared to Adposhel.

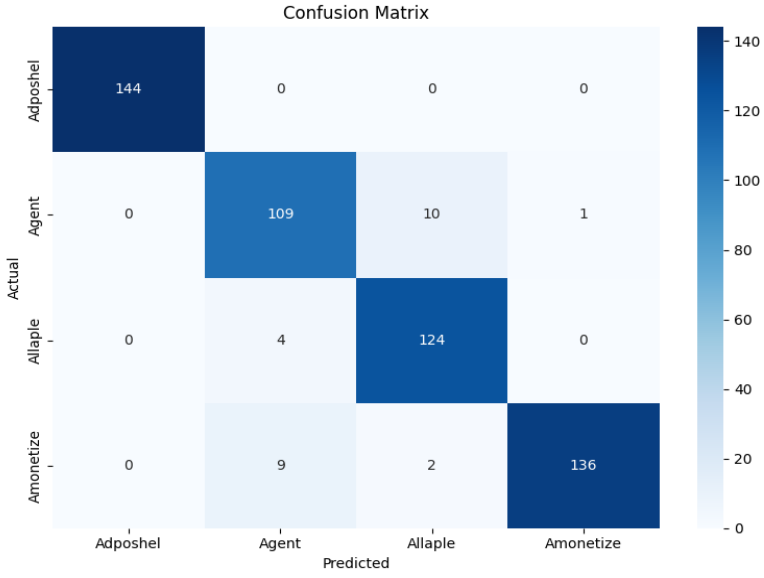


Fig. 1. Confusion matrix for the selected MaleVis malware classes.
Source: Authors' own processing based on the experimental results.

The confusion matrix confirms that the model provides a suitable baseline for the explainability workflow. The observed errors are useful for the next stage of analysis, where Grad-CAM visualizations are used to inspect the image regions that influenced the model predictions. In this way, classification results and visual explanations complement each other: the quantitative metrics indicate how the model performs, while the explainability outputs help interpret why specific decisions may have been produced.

4.2 Grad-CAM Visualization Results

After evaluating the classification performance, Grad-CAM was applied to selected validation images in order to generate visual explanations for the model predictions. The objective of this stage was to inspect which regions of the malware image contributed most strongly to the classification decision. For each selected sample, the original malware image was compared with the corresponding Grad-CAM heatmap overlaid on the same image.

The visual explanations show that the model does not rely uniformly on the entire image, but focuses more strongly on specific regions. In several cases, the highlighted regions correspond to areas with visible structural patterns, such as dense horizontal bands, texture variations, or transitions between visually distinct areas of the malware image. These regions may represent discriminative visual characteristics learned by the convolutional neural network during training.

Fig. 2 presents an example of the Grad-CAM explanation generated for a malware sample from the validation set. The left side shows the original malware image, while the right side

shows the Grad-CAM visualization. The warmer regions indicate the areas that contributed more strongly to the model's classification decision.

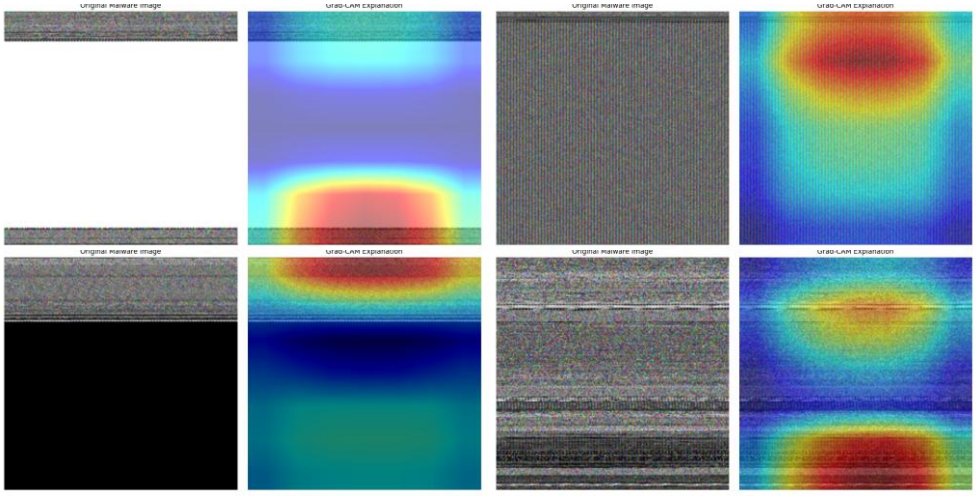


Fig. 2. Grad-CAM visualization examples for selected malware images from the MaleVis validation subset.
Source: Authors' own processing based on the experimental results.

Additional Grad-CAM examples were generated for malware samples belonging to the selected classes, including Adposhel, Agent, Allaple, and Amonetize. These examples show that the model attention varies depending on the visual structure of each malware family. For some samples, the activation is concentrated in compact regions, while for others it is distributed across broader image areas. This variation suggests that the model identifies different discriminative patterns depending on the malware class and image structure.

From an interpretability perspective, these visual explanations are useful because they provide additional information beyond the predicted label. A cybersecurity analyst can inspect whether the model appears to focus on meaningful visual structures or whether the prediction may be influenced by irrelevant regions. This can support model validation, error analysis, and trust-building in AI-based cybersecurity workflows.

However, the Grad-CAM results should be interpreted as an initial explainability layer rather than as a complete explanation of the model reasoning. Grad-CAM indicates where the model focused spatially, but it does not fully explain the semantic meaning of the learned features. Therefore, the visual results should be used together with classification metrics, confusion matrix analysis, and future experiments involving additional XAI methods.

4.3 Discussion

The experimental results show that the integration of explainability mechanisms can add an important layer of transparency to AI-based cybersecurity workflows. While the classification metrics indicate that the model was able to distinguish between the selected

malware classes with good baseline performance, the Grad-CAM visualizations provide additional insight into how these predictions were produced. This is relevant because, in operational cybersecurity contexts, the predicted label alone is often insufficient. Analysts need to understand whether an AI system focuses on meaningful patterns or whether the decision may be influenced by irrelevant visual artifacts.

The generated heatmaps indicate that the model concentrates its attention on specific regions of the malware images rather than treating the entire image uniformly. These highlighted regions suggest that the convolutional neural network learned discriminative visual structures associated with the selected malware classes. From an interpretability perspective, this is useful because it allows human analysts to inspect the areas that contributed most strongly to the prediction. Such visual evidence can support alert validation, help identify suspicious or unexpected model behavior, and provide a starting point for more detailed error analysis.

The results also highlight the importance of explainability for trust-building in smart city cybersecurity systems. In intelligent urban infrastructures, AI-based security tools may support decisions related to incident detection, service continuity, and protection of critical systems. If these tools operate as black boxes, their outputs may be difficult to justify to technical teams, public administration stakeholders, or decision-makers. By adding an explainability layer, AI-based cybersecurity systems become more transparent and easier to audit, which is essential in environments where accountability, reliability, and human oversight are required.

However, the current experiment also has several limitations. First, the evaluation was conducted on a reduced subset of the MaleVis dataset, which means that the results cannot be generalized to all malware families. Second, the model was trained for a limited number of epochs, as the objective was to validate the explainability workflow rather than to optimize the classifier. Third, Grad-CAM provides spatial visual explanations, but it does not fully describe the semantic meaning of the features learned by the neural network. Therefore, the heatmaps should be interpreted as supportive evidence, not as complete explanations of the model's internal reasoning.

Despite these limitations, the proof-of-concept confirms the relevance of integrating XAI mechanisms into AI-assisted cybersecurity pipelines. For cybersecurity analysts, Grad-CAM visualizations can provide a practical way to examine model attention, compare correct and incorrect predictions, and better understand the behavior of deep learning models used for malware detection. In future work, this approach can be extended with additional XAI methods, larger datasets, adversarial robustness evaluation, and integration into broader security monitoring workflows relevant to cyber-resilient smart cities.

5. Future Work

The current study represents an initial proof-of-concept focused on validating the integration of deep learning-based malware classification with visual explainability mechanisms. Future work will extend the experimental setup by including a larger number of malware families from the MaleVis dataset and by evaluating the model on more

complex and diverse classification scenarios. This will allow a more detailed analysis of the model's ability to generalize across different malware categories and visual structures.

Another important direction concerns the integration of additional explainability techniques. While Grad-CAM provides useful visual explanations for convolutional neural networks, future experiments may include complementary methods such as LIME, SHAP, and attention-based visualization. Combining multiple XAI techniques could provide a more complete interpretation of model behavior and help compare different types of explanations in cybersecurity tasks.

Future research will also address adversarial robustness. Since AI-based cybersecurity systems may themselves become targets of adversarial manipulation, it is important to evaluate how explainability mechanisms behave when malware images are modified or intentionally perturbed. This direction is particularly relevant for smart city environments, where the reliability of AI-based threat detection systems is essential for protecting critical services and infrastructures.

Finally, future work will explore how explainability-oriented cybersecurity models can be integrated into broader operational workflows, such as Security Operations Centers and smart city monitoring platforms. In such contexts, explanations should support not only technical interpretation, but also alert prioritization, incident response, human oversight, and decision accountability.

6. Conclusion

This paper examined the role of Explainable Artificial Intelligence in supporting cyber-resilient smart city environments, with a focus on improving the transparency and interpretability of AI-based cybersecurity systems. The study proposed an explainability-oriented workflow that combines malware image classification with visual explanation mechanisms, aiming to support both automated threat detection and human understanding of model decisions.

The experimental implementation demonstrated the feasibility of this approach using a ResNet18-based transfer learning model and the MaleVis malware visualization dataset. The classification results showed good baseline performance on the selected malware classes, while the Grad-CAM visualizations provided interpretable heatmaps indicating the image regions that contributed most strongly to the model predictions.

Although the experiment represents an initial proof-of-concept, the results confirm that explainability can add value to cybersecurity workflows by supporting model validation, analyst trust, and decision accountability. This is particularly relevant for smart city infrastructures, where AI-based systems must be not only accurate, but also transparent, auditable, and aligned with human oversight requirements.

Future developments will extend the evaluation to additional malware classes, complementary XAI methods, and adversarial robustness scenarios. Overall, the work

provides a practical foundation for developing more trustworthy and explainable AI-assisted cybersecurity systems for intelligent urban environments.

Acknowledgements
This research has received funding from European Commission's Horizon Europe research and innovation program under grant agreement No. 101131292 (AIAS).

References

[1] Vrabie, Catalin. "Promisiunile Inteligenței Artificiale (AI) administrației publice și orașelor inteligente." Smart Cities International Conference (SCIC) Proceedings. Vol. 11. 2023.
[2] Vrabie, Catalin. "Smart Urban Governance. Administrația Publică în era Smart: Tehnologie, Date și Cetățeni." Smart Cities and Regional Development (SCRD) Preprints 1.1 (2024).
[3] Boce, Grigorina, and Besmir Kanushi. "Artificial Intelligence in Cyber Security." Smart Cities and Regional Development (SCRD) Preprints 2.1 (2025).
[4] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
[5] Selvaraju, Ramprasaath R., et al. "Grad-CAM: visual explanations from deep networks via gradient-based localization." International journal of computer vision 128.2 (2020): 336-359.
[6] Vrabie, Catalin. "Artificial intelligence promises to public organizations and smart cities." PLAIS EuroSymposium on Digital Transformation. Cham: Springer International Publishing, 2022. 3-14.
[7] Loo-See, B. E. H., et al. "Smart Cities and Digitalisation: A Research Agenda of Public Administration." Smart Cities and Regional Development (SCRD) Journal 6.3 (2022): 41-58.
[8] Virtosu, Ina. "Guardians or threats? The double-edged role of artificial intelligence in cyber security." International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings. Vol. 2. 2025.
[9] Waizel, Guy. "The Agentic AI Security Adoption Matrix: Understanding Readiness and Resistance Across Domains." (2026).
[10] Ai, Hleg. "High-level expert group on artificial intelligence." Ethics guidelines for trustworthy AI 6.15.3 (2019): 2024.
[11] Capuano, Nicola, et al. "Explainable artificial intelligence in cybersecurity: A survey." Ieee Access 10 (2022): 93575-93600.
[12] Charmet, Fabien, et al. "Explainable artificial intelligence for cybersecurity: a literature survey." Annals of Telecommunications 77.11 (2022): 789-812.
[13] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
[14] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
[15] Nazim, Sadia, et al. "Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM." Plos one 20.5 (2025): e0318542.