BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en Vol. 30 No. 2. 2025

UDC 338.43
DOI: 10.62660/bcstu/2.2025.77

Evaluating the effectiveness of image recognition systems for automatic detection of malicious files based on image metadata

Vitalii Yasenenko*
Master, Senior Software Developer
TP-Link
92618, 36 Technology St., Irvine, USA
https://orcid.org/0009-0004-4801-9541

Abstract. The relevance of the study was determined by the increasing threat of covert distribution of malicious software through the metadata of digital images, which complicated the detection using standard methods. The aim of the work was to develop a new approach to detecting malicious files through the analysis of image metadata using artificial intelligence methods. To achieve this, a detailed analysis of the main metadata standards was carried out, and vulnerable fields capable of hiding malicious code and being ignored by traditional security methods were identified. The results of the theoretical study showed that the most informative characteristics for threat detection were metadata features such as timestamps, geolocation coordinates, and device data. It was also established that non-standard values in the fields, such as abnormal timestamps or suspicious code markers, could serve as indicators of malicious activity. A comparison of traditional threat detection methods was conducted, which revealed the low effectiveness when working with metadata, as these methods were mainly focused on identifying malicious elements in the visual part of the file rather than on analysing the accompanying structure. The developed conceptual model, oriented towards the specified characteristics, demonstrated significant potential for effectively detecting anomalies and hidden malicious code in metadata. This approach made it possible to reduce the number of false positives, as it focused not only on detecting obvious deviations but also on subtler changes in the structural layer of images. The conclusions confirmed that the analysis of accompanying information was an important tool for detecting new forms of threats. The practical significance of the study lay in the possibility of using the proposed concept as a basis for the development of specialised systems for monitoring and preventing cyber incidents

Keywords: steganography; malware; digital images; machine learning; metadata anomalies

INTRODUCTION

The issue of covert distribution of malicious software through digital images remained one of the most complex challenges in the field of cybersecurity. One of the latest directions in this domain was the use of image metadata to conceal malicious code, which significantly complicated the process of detecting threats using standard file-checking methods. A particular threat was posed by subtle modifications in the pixels or metadata structure that did not cause visual changes and remained invisible to traditional security systems. Due to the combination of classical and multi-level methods of hiding information, traditional detection approaches proved to be ineffective. With the advancement of steganographic capabilities, there was a growing need for new systems of automatic threat recognition capable of analysing not only the visual component but also hidden information in the structure of images. Therefore, the development and evaluation of the effectiveness of

Article's History: Received: 24.12.2024; Revised: 17.04.2025; Accepted: 16.06.2025

Suggested Citation:
Yasenenko, V. (2025). Evaluating the effectiveness of image recognition systems for automatic detection of malicious files based on image metadata. *Bulletin of Cherkasy State Technological University*, 30(2), 77-87. doi: 10.62660/bcstu/2.2025.77.

*Corresponding author

such systems based on artificial intelligence (AI) was a pressing task for modern cybersecurity.

Against the backdrop of active development of steganographic techniques in metadata, the task of creating such methods and developing techniques for detecting hidden information became important. Y. Fernando *et al*. (2024) presented an innovative approach to steganography via image metadata, which ensured a high level of concealment of transmitted information. However, the authors focused on improving hiding methods while leaving the issue of detection without thorough analysis. In the work of D.R. Setiadi *et al*. (2025), a wide range of modern steganographic methods were reviewed, including image, textual and 3D data, but the need to explore practical aspects of detecting malicious modifications in metadata was also noted.

The challenges of increased data stealth and detection difficulty were addressed in the study by O. Kuznetsov *et al*. (2024a), which showed that using a wide range of signals in steganography significantly increased detection complexity. This, in turn, required the adaptation of more complex analytical models, including AI applications. In this context, the works of L. Caviglione & W. Mazurczyk (2022) outlined the danger of stego-malware as hidden threats, emphasising the ineffectiveness of traditional antivirus tools in detecting new types of attacks. V. Verma *et al*. (2022) supported this conclusion by describing the difficulties of identifying malicious code disguised as legitimate digital objects in Windows environments.

The use of deep learning for image analysis was explored by J. El Abdelkhalki *et al*. (2022), where the successful application of neural networks for detecting malicious content was highlighted. However, attention was paid to the analysis of the visual part of images, while the issue of hidden threats in metadata remained underexplored. The study by A.I. Iskanderani *et al*. (2021) demonstrated the effectiveness of using AI for steganalysis of digital objects, but the authors did not address the modelling of threats at the metadata level.

Additionally, C. Ahmadi *et al*. (2024), in their work, analysed the use of steganography in backdoor attacks against AI models, emphasising the difficulty of detecting such intrusions, which further underscored the need to create adaptive protection mechanisms. Meanwhile, reviews by I.H. Sarker (2023) and A.H. Salem *et al*. (2024) confirmed the growing role of intelligent systems in ensuring cybersecurity, but the issue of metadata analysis remained insufficiently developed in the works.

The comparison of traditional and intelligent solutions for threat detection tasks conducted by F. Wang & Y. Tang (2024) showed the advantages of AI methods in complex scenarios, although without detailing the specific features of working with supplementary information in media files. Regarding Ukrainian research, A. Kashtalian *et al*. (2024) developed a multi-computer system for detecting malicious software considering metamorphic changes, focusing on code transformations and behavioural characteristics of programs. However, metadata structure remained outside the scope of analysis. At the same time, A. Kobozieva *et al*. (2023) focused on developing steganalysis methods for digital video and image sequences, paying attention to changes at the bit-sequence level. However, the approaches were primarily oriented towards content analysis rather than structural metadata features, leaving open the question of integrating such solutions into the task of detecting malicious attachments in image-associated data.

The aim of the study was to develop a conceptual approach to detecting malicious files by analysing image metadata using modern AI methods. To achieve this aim, the following objectives were formulated: to justify the selection of key metadata characteristics subject to analysis for detecting malicious software; to analyse typical scenarios of using steganography in image metadata to conceal malicious software; to assess the potential of using AI methods for automatic threat recognition in metadata; and to develop a basic concept for a model that detects malicious files based on metadata analysis.

MATERIALS AND METHODS

The assessment of the possibility of concealing malicious software in digital images was conducted through the analysis of the technical characteristics of metadata formats, the usage patterns in file processing, and potential risk aspects related to manipulation of structural elements. The study took into account scenarios of covert threat distribution via modification, addition, or encryption of information in metadata fields without affecting the visual integrity of the file. The examination was based on open technical documentation of metadata formats, such as the EXIF Specification (Camera & Imaging Products Association, 2012) and the International Press Telecommunications Council (2024), as well as cybersecurity analytical reports, such as the Internet Security Threat Report (2017), which explored the use of steganography for hiding malicious software. For the sake of analytical relevance, the materials were selected based on criteria of topicality, practical orientation, and alignment with the subject of malicious code detection in metadata. Open archives of digital images also played an important role, providing examples of typical metadata structures, such as the OpenImages Dataset (Krasin *et al.*, 2017), Break Our Steganographic System (Bas *et al.*, 2011), and StegoAppDB (Newman *et al.*, 2019).

Theoretical modelling of the malware detection system in image metadata was performed through the analysis of structural features of the exchangeable image file format (EXIF), the international press telecommunications council metadata standard (IPTC), and the

extensible metadata platform (XMP), involving open technical specifications and academic publications in the field of information security. To develop the conceptual design, data were used on metadata organisation, potential attack vectors, and known practices of manipulating supplementary information in digital files. Critical metadata characteristics were identified through the systematisation of fields that allowed unrestricted input or storage of large data volumes without mandatory validation. Particular attention was paid to timestamps, geolocation coordinates, device identifiers, software versions, structural consistency of fields, and total metadata volume. The list of features was formed based on comparative analysis of typical and anomalous values found in publicly available metadata examples of digital images.

The conceptual model of processing was built by determining the main stages of input data transformation: extraction of supplementary information, removal of secondary fields, normalisation of numerical values to a single scale, and encoding of categorical variables using one-hot encoding. All processed features were considered as an input vector for further analysis using machine learning. To ensure the model's resilience to data structure variability, consistency of feature formats was maintained regardless of the original metadata type. The classification component of the concept was developed with a focus on using a shallow multi-layer perceptron (MLP) neural network. This decision was based on the requirements for computational efficiency and the model's ability to adapt to complex interdependencies between features without manual pattern identification. The alternative use of gradient boosting models was considered for situations requiring increased classification accuracy with limited input features. To manage classification sensitivity, the system proposed a probabilistic interpretation of model outputs with the ability to adjust threshold values.

The conceptual verification of the system's architecture was carried out by logical modelling of potential scenarios for concealing malicious code in metadata, based on typical anomalies and non-standard structural modifications described in the literature. The modelling was theoretical in nature and involved risk analysis arising from manipulations with metadata fields (UserComment, MakerNote, XMP custom fields, etc.). All assumptions were based on publicly available technical documentation and analytical reports, which avoided the experimental influence of variable environments and ensured the generalisability of the concept for future implementation.

RESULTS

Metadata of digital images represent a critically important element of modern information systems, providing contextual descriptions of content, facilitating indexing, copyright management, and automation of content processing. However, alongside the widespread use of metadata, significant information security threats arise due to the peculiarities of implementation and functioning of the main metadata storage standards. The most common formats are EXIF, IPTC, and XMP, each with specific data organisation characteristics and corresponding vectors of potential attacks.

The EXIF standard, initially developed for digital cameras, involves embedding metadata directly into image files, primarily in JPEG or TIFF formats. Since EXIF was primarily aimed at preserving technical information, such as shooting parameters, camera model, or geolocation data, the issue of metadata structure security was not a priority during the design of the standard. Fields that allow free input of data, such as "UserComment" and "MakerNote", may store arbitrary text or binary arrays without proper validation. This creates a possibility for injecting hidden code, including scripts or objects, for subsequent system exploitation during file processing by inadequately protected software (Monika & Eswari, 2023).

IPTC metadata, used mainly in professional journalism and media production, allow the description of textual image attributes: titles, keywords, author names, usage rights, etc. The IPTC format provides stricter data structuring compared to EXIF, but several fields still allow the input of large volumes of text without format or length limitations. Accordingly, if the receiving system does not implement adequate mechanisms for checking or cleaning textual data, such fields may be exploited to inject malicious content, including SQL injections, scripts, or other forms of attacks on system integrity or confidentiality (Jian *et al.*, 2021).

The XMP format, proposed by Adobe Systems, implements a more flexible metadata concept based on the principles of extensibility and integration. XMP uses standardised XML structures to describe metadata fields, allowing the inclusion of an unlimited number of hierarchically organised records. Compared to EXIF and IPTC, XMP offers greater flexibility and expansion potential by adding user-defined schemas and namespaces. At the same time, the use of XML as a foundation introduces additional vulnerabilities specific to the processing of structured data. Since most XML parsing libraries do not impose default restrictions on document depth or size, inadequate handling of XMP metadata may become a serious attack vector on applications interacting with digital images. For illustration, Table 1 presents an example of a typical EXIF metadata set and a simulated example with anomalous values that may potentially contain hidden payloads or indicate steganographic use of the field.

Thus, digital image metadata can be used to hide malware using a variety of techniques aimed at manipulating the structural elements of the file. Table 2 lists the main methods of hiding threats through metadata.

Table 1. Examples of typical and anomalous values in digital image metadata

Field	Typical value	Abnormal value
Make	Canon	$$%null%%
Model	EOS 600D	bash_exec:systemctl.sh
Software	Adobe Photoshop CC 2019	aHR0cDovL21hbGljaW91cy5leGU=
UserComment	Summer vacation, Aug 2022	<script>eval('malware')</script>
XMP Payload	-	{"payload":"execute","flag":true}

Source: created by the author

Table 2. Typical methods of hiding malware in digital image metadata

Hiding method	Metadata field	Risk description
Inserting encrypted text	UserComment, MakerNote	Hidden payload among legitimate data
Creating redundant fields	XMP custom fields	Injection of non-standard metadata
Manipulation of geolocation data	GPSInfo	Transferring malicious commands
Paste into file header	File Header	Code masking in service segments
Mutation of standard fields	Make, Model, Software	Imitation of legitimate data to hide commands
Using non-standard encodings	UserComment, XMP Description	Data masking via Base64, Hex and other encodings

Source: created by the author based on data from the Camera & Imaging Products Association (2012), Internet Security Threat Report (2017), and International Press Telecommunications Council (2024)

The analysis of metadata storage standards demonstrated that the existing solutions did not provide an adequate level of security without additional measures of validation, cleaning, and input data restriction. In this regard, it became necessary to investigate the effectiveness of existing mechanisms for detecting hidden malicious objects in digital images, focusing both on traditional antivirus approaches and on specialised steganalysis methods.

The main methods used in practice remained general-purpose antivirus systems. The functioning was based on searching for known signatures, behavioural analysis of objects, or heuristic anomaly detection. However, when working with images, antivirus systems demonstrated significant limitations. Image files by nature contained numerous permissible variations in content and metadata, complicating the construction of unambiguous signatures for malicious modifications. In addition, antivirus products were mostly content-oriented and did not conduct in-depth analysis of accompanying data, which allowed attackers to use metadata for transferring hidden code without hindrance (Sarker, 2024).

Steganography in metadata differs from classical methods of hiding visual information in that it uses textual or binary metadata fields to embed hidden content. Therefore, another approach to detecting concealed changes in images became steganalytic methods based on pixel structure analysis. Such methods aimed to detect traces of steganography by examining statistical anomalies in pixel distribution, histogram analysis, or the calculation of specific signal characteristics. The use of such approaches allowed for the highly sensitive identification of classical forms of information hiding in visual data (Płachta *et al.*, 2022). However, in the case of working with metadata, the effectiveness of steganalytic methods significantly decreased, as the accompanying image information could not be analysed via visual features (Kuznetsov *et al.*, 2024b). Thus, traditional pixel-based analysis did not allow for the detection of hidden threats embedded at the level of an image's structural data.

It is also worth noting attempts to use simple machine learning models to detect changes in file data structure. These models were generally built on classification or clustering approaches using limited sets of features, such as metadata size, number of fields, or types of attributes used. Although the results of such analysis made it possible to detect obvious anomalies, these methods remained insensitive to complex and well-camouflaged threats. The models often failed to consider deep interrelations between different metadata characteristics and were prone to high levels of false positives in cases of non-standard but legitimate changes to file structure. For a visual comparison of the effectiveness of traditional methods for detecting hidden malicious software in metadata of digital images, Figure 1 illustrates the dependence of detection accuracy on the level of attack complexity.

As a result of the analysis of existing solutions, it was established that none of the traditional approaches provided an adequate level of effectiveness in detecting concealed malicious code in image metadata. The lack of adaptability in models and the limitations of classical analysis methods at the level of content or simple characteristics of accompanying data created significant obstacles to detecting new forms of threats. Thus, the analysis showed that existing methods had limited potential for identifying modern threats, especially those masked within the metadata layer of digital images. Consequently, there arose a need to develop a solution focused specifically on analysing the structural information of files.

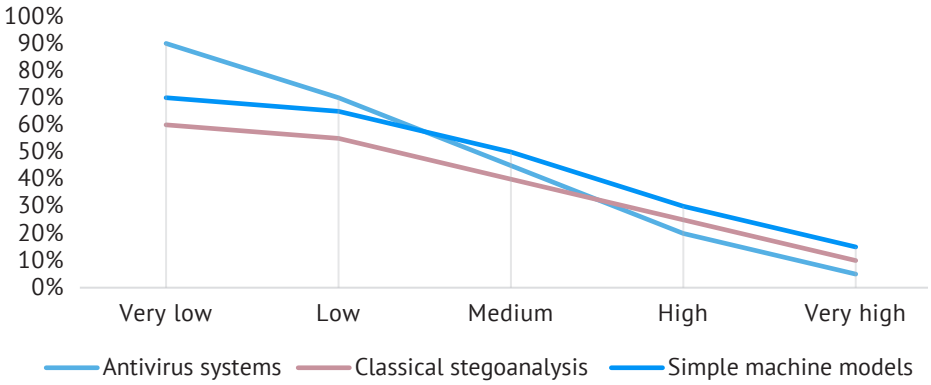

Figure 1. Dependence of the effectiveness of various methods for detecting malware in metadata on the complexity of attacks

Source: created by the author

Based on the analysis conducted, it was expected that the input to the proposed system could be a digital image file in JPEG, PNG, or another common graphic format containing a metadata structure. Given the specific task – detecting concealed malicious elements – it would be reasonable to focus not on the full visual analysis of the image, but rather on processing its structural part. Metadata are considered a potentially vulnerable carrier of hidden information, which justifies the prioritisation in the analysis. To ensure a structured approach to detecting malicious software in image metadata, the proposed system envisaged the sequential execution of several stages of data processing. Figure 2 presents the general system architecture, illustrating the main stages of analysis of the input object.

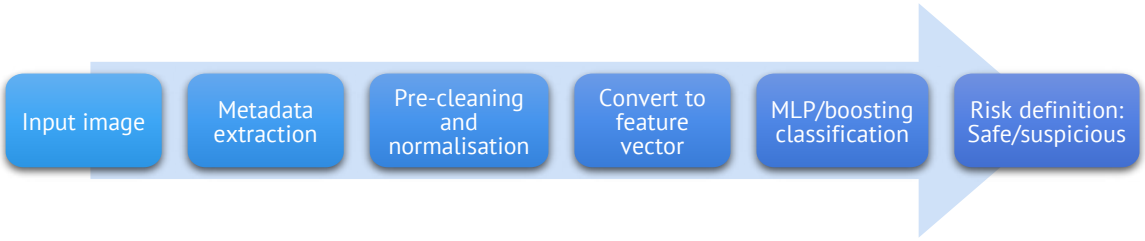

Figure 2. General scheme of digital image processing in the proposed threat detection system

Source: created by the author

It was envisaged to perform preliminary parsing of the input file to extract supplementary data, such as the EXIF section and other accessible fields, characteristic of the respective format. This approach was determined by the fact that metadata usually contains static textual or numerical values, which simplifies the processing compared to the analysis of complex pixel structures. Metadata analysis was expected to include the study of a number of characteristics that could potentially serve as indicators of malicious activity. Specifically, it was proposed to evaluate the overall size of the EXIF section, the number of present tags, the types of tags used, and to check for anomalous or unusual values in the metadata fields. Such anomalies could include excessively large or negative numerical values, incorrect date formats, unusual or suspicious codes identifying device manufacturers or shooting parameters.

To increase system reliability, a superficial analysis of basic pixel structure characteristics was allowed – for example, noise levels or microdeformations – however, the main focus would remain on processing metadata as a more probable vector for threat concealment. The most informative elements for analysis proved to be the timestamps for file creation and editing, geolocation coordinates, recording device data, software version numbers, and the total metadata size. Anomalies in these fields, especially in cases of inconsistency or excessive size, could indicate attempts to conceal a malicious payload (Zuppelli *et al.*, 2021).

In selecting the type of machine learning model, it would be appropriate to focus on the need to ensure a balance between high analytical sensitivity and moderate computational demands. Accordingly, one of the most suitable architectures was considered to be an MLP with some layers. This structure allows for efficient operation with medium-dimensional input vectors and provides sufficient flexibility for adaptive learning. Alternatively, gradient boosting methods could have been chosen, due to demonstrating high accuracy on tabular data. However, the neural network was considered the preferred option due to its ability to self-update and better suitability for processing new types of threats

without requiring full model reconstruction. Compared to decision trees or support vector machines, MLPs offer a higher learning capacity on complex interdependencies between features, which is critically important for detecting non-standard and masked malicious patterns in metadata structures (Yadav *et al.,* 2022).

The data processing concept in the system was expected to involve several sequential stages. First, the structural information extracted from the input image would be cleansed of secondary attributes, then converted into a feature vector. When forming input feature vectors for the model, preprocessing steps such as normalisation of numerical values (e.g. scaling to a 0 to 1 range) and one-hot encoding for categorical fields like tag types or device identifiers were to be applied. Such data preparation would allow the model to more effectively identify patterns in mixed feature sets. This vector would then be fed into the model, which would compute internal representations and classify the file into two classes: safe object or object with potential threat. To improve prediction accuracy, a probabilistic interpretation of the model output would be used, enabling flexible adjustment of trigger thresholds depending on the acceptable risk level in specific application conditions.

At the system's output, a decision on the analysed object would be generated in the form of a percentage probability that the metadata contained hidden malicious code. For practical use, an integrated interpretation of the result was also foreseen, in the form of a simple label – "safe" or "suspicious" – which would simplify operational decision-making on whether to further process or isolate the file. The proposed model was expected to ensure adaptability to new methods of malicious code concealment in metadata, while maintaining a low rate of false positives when processing atypical but safe changes. Thanks to the ability of neural networks to learn from complex multidimensional data, it was anticipated that such a system could significantly enhance the reliability and timeliness of detecting modern cyber threats that remain beyond the reach of traditional analysis methods.

The expected results of the model's functioning were clarified through an analysis of typical use-case scenarios. For instance, in cases of processing images with EXIF metadata of standard size-approximately 1,000 bytes – containing only typical tags such as camera model, exposure parameters, etc., a high probability of safe classification at 98-99% was anticipated. Conversely, if anomalies were present in the metadata – non-standard date formats, atypical or excessively large numerical values in geolocation or timestamp fields – the system would be expected to demonstrate a significant increase in threat detection probability. In such cases, the likelihood of classifying the image as suspicious was expected to rise to 80-90%, depending on the nature and intensity of the detected deviations. Thus, the model could effectively support decisions regarding the further processing or isolation of potentially dangerous files, which is critically important for ensuring information security amid growing cyber threats.

If image processing involved metadata with many tags and anomalous values – e.g. up to 2,000 bytes of data in the EXIF section, where strange or non-standard codes for camera manufacturers or shooting parameters were detected – this could result in a significantly higher probability of threat detection, reaching 95% or more. In such a case, the system would need to be capable of adjusting its trigger thresholds to prevent excessive false positives, ensuring that ordinary files with mistakenly incorrect but harmless metadata would not be blocked. To maintain the model's relevance, it was possible to foresee the implementation of a mechanism for periodic updates based on new data. The system could store anonymised feature vectors of new files and periodically undergo fine-tuning based on stored data without requiring full retraining. This would allow the model to adapt to evolving methods of malicious code obfuscation.

At the same time, understanding and awareness of potential issues proved essential for further improving the model. One such issue could be the processing of data with certain but frequently recurring metadata anomalies that are not harmful, but might be misclassified as threats. This phenomenon may occur due to non-standard yet safe user operations – for example, when metadata contains incorrect or unusual values that do not conform to general patterns. This situation, in particular, could lead to false positives. However, these errors could be minimised due to the possibility of continuous learning and model adjustment on new data (Carneiro *et al.,* 2023). An essential part of this process is the adaptation of algorithms for optimising classification based on new data, which reduces the likelihood of errors in future use.

Another possible issue could be the trade-off between data processing speed and result accuracy. In cases where metadata has a complex structure or an image contains a large volume of data, the processing time may increase, potentially affecting overall system speed. Nevertheless, thanks to innovative optimisation techniques such as parallel data processing or the use of more powerful computing resources, this issue could be significantly mitigated (Monteiro *et al.,* 2021). The prospects for integrating faster computational mechanisms would significantly improve processing speed without loss of accuracy, which would positively impact overall efficiency. Further important challenge might be the scalability of the model as the volume of data processed increases. However, this issue could be addressed by improving big data processing methods, applying intelligent algorithms for automatic resource allocation, and reducing the need for constant processing of all data via early-stage filtering. As a result, these measures would allow the system to remain effective even when working with large volumes of metadata. Thus, despite some potential challenges, the implementation of this model could demonstrate significant potential for improving threat detection accuracy and

speed. Thanks to its ability to adapt to new data and to optimised processing workflows, the model would maintain a high level of effectiveness even under complex or changing data conditions, offering a considerable advantage over traditional security methods.

DISCUSSION

Automated threat detection systems based on AI methods have become an important tool in countering steganographic techniques used by malicious actors to conceal harmful components within digital objects. Data analysis in such systems usually focuses on visual information; however, metadata remains a less explored, albeit promising, source of intrusion indicators. The study examined the potential for constructing a model to detect malicious images based on the analysis of file structural information. This approach allowed not only for reduced computational costs, but also for identifying anomalies that remain invisible to systems focused on image content. In this regard, it became necessary to compare the obtained results with other contemporary studies on the use of AI in the classification and detection of threats in graphical data.

The proposed approach to analysing metadata of digital images as the primary source of potential threat indicators proved relevant in the context of an increasing number of attacks using steganographic methods to disguise malicious code. The conducted study confirmed that it is the structural characteristics of fields, rather than the visual features of images, that represent a promising target for threat modelling. This line of inquiry continued the direction initiated by D. Puchalski *et al*. (2020), who demonstrated the viability of detecting malicious objects through structural analysis of media files. The application of detailed parsing of accompanying information allowed for increased classification accuracy of objects while reducing the risk of false positives during atypical but legitimate file modifications.

Particular attention should be paid to the work of S. Kiltz *et al*. (2024), which implemented metadata-based tracing of image changes to detect files with hidden malicious payloads. The study confirmed that even with minimal visual changes, metadata remains a sensitive indicator of interference. This reinforced the conclusions regarding the effectiveness of metadata as a vector of analysis, complementing classical pixel comparison methods. Equally significant for drawing conclusions was the analysis of approaches to constructing classification models that use images as a source of input features. In the work of R. Chaganti *et al*. (2022), malware was represented as images and classified using EfficientNet. This approach demonstrated high accuracy in threat detection due to the capability of convolutional neural networks to detect complex spatial patterns. At the same time, this study substantiated the relevance of focusing on the structural features of accompanying data, thus avoiding excessive dependence on the visual representation of the object.

A similar approach was implemented in the study by S.A. Roseline *et al*. (2020), where a deep model based on a random forest was used to classify images based on visual features. The authors highlighted the advantages of the ensemble approach in improving accuracy and reducing the likelihood of overfitting in the presence of large volumes of heterogeneous input data. However, the model proved sensitive to visual deviations in pixel structure and was ineffective in cases where information concealment was carried out not through the visual channel, but at the level of the file's accompanying information. This confirmed the limitations of models focused exclusively on image analysis without considering contextual metadata. In this context, the structural analysis of metadata implemented in the current study made it possible to detect threats that remained outside the focus of visual-type models, while maintaining lower computational costs and greater adaptability to non-standard input object formats.

More advanced architectures were proposed by F. Ullah *et al*. (2022), who developed a hybrid model with visualisation of control-flow graphs and a multi-head self-attention mechanism. By combining structural features of program behaviour with deep analysis of instruction dependencies, the model achieved high accuracy in detecting complex threats. However, this approach required complex pre-processing, including decompilation and execution tracing, which rendered it unsuitable in scenarios with limited access to full code or in cases involving isolated media files. The concept proposed in the current study had the advantage that only the accompanying structural information was used for analysis, which is generally available without specialised tools. This ensured flexibility, scalability, and suitability for integration into high-performance environments with stringent response time requirements.

A significant contribution to model optimisation was offered in the study by A. El-Ghamry *et al*. (2023), where an effective IoT threat detection system was developed based on an optimised image-based classifier structure. The authors focused on reducing resource demands without sacrificing accuracy by combining relevant feature selection with adaptive learning. This approach reinforced the importance of focusing on compact yet meaningful feature sets, which aligns with the current study's orientation toward analysing a limited but informative segment of metadata. The proposed model did not require processing the visual content of the image, thereby lowering computational requirements and allowing effective detection of hidden anomalies in metadata structures, particularly through combining EXIF, IPTC, and XMP fields.

In the study by D. Vasan *et al*. (2024), a broad learning architecture was proposed for GPU-free classification, ensuring efficiency under limited resources by reducing model depth and employing multidimensional feature representation. This approach was evaluated as suitable for deployment in applied scenarios lacking

access to powerful graphics processors. The concept proposed in the current work, with a similar emphasis on resource efficiency, differed in that operations were conducted on tabular rather than visual vectors. This simplified data pre-processing and enabled system scalability without significantly burdening the runtime environment. The use of an MLP model with a low number of layers was supplemented by normalisation of input values, which collectively ensured efficient and adaptive processing of structural metadata features with the ability to accurately detect hidden threats.

The issue of algorithm explainability was also of particular importance in the context of applying intelligent systems for threat detection. The work by A. Galli *et al*. (2024) analysed methods for improving transparency in behavioural models for malware detection. The authors emphasised the importance of result interpretability for the practical implementation of AI solutions in cybersecurity. In light of this, the conceptual model proposed in this study provided for the possibility of outputting a probabilistic estimate and a flexible classification threshold, which would support better understanding of the analysis results even without a full "black box" model. This approach aligns with the position that algorithm transparency is a critical factor of trust in automated threat detection systems.

A general overview of the prospects of machine learning for malware detection was presented by J. Ispahany *et al*. (2024). This study pointed to limitations related to excessive dependence on signature databases, insufficient adaptability of existing systems to new types of attacks, and the difficulty of ensuring scalability. The conclusions obtained in the current study aligned with these statements, as the metadata-based detection concept aimed to overcome these limitations through structural analysis of accompanying data without reliance on known threat patterns. A particular prospect was the potential for continuous model updating without the need for full retraining, which met the challenges outlined by the authors of the review.

Another direction worth noting is the application of metadata in a broader sense – not only for images but also for network flows. M. Russo *et al*. (2021) demonstrated the effectiveness of network metadata analysis for detecting illegal cryptocurrency mining. The authors proved that even without access to full packet content, traffic metadata could serve as a sufficient source of indicators for threat detection. In this respect, the study reaffirmed a common idea: metadata, despite its auxiliary nature, can act as the primary indicator of malicious activity. This further justified the need to develop detectors focused on analysing the structural layer of data in various formats, including graphical ones. Comparison of the results obtained with modern approaches to malware detection showed that structural information analysis, particularly of digital image metadata, fills the gaps characteristic of methods focused solely on visual or behavioural features. The rationale for this approach was confirmed by several studies highlighting the importance of additional information layers for improving threat detection accuracy. Unlike traditional models, metadata analysis proved more informative in cases of complex or non-standard concealment forms and demonstrated better consistency with requirements for interpretability, speed, and adaptability.

CONCLUSIONS

The results of the study demonstrated the importance of metadata as a potential vector for concealing malicious programs, significantly complicating the detection using classical cybersecurity tools. It was established that standard threat detection approaches, mainly focused on visual features, are ineffective because of overlooking anomalous changes in the image's service structures. In particular, it was found that metadata in EXIF, IPTC, and XMP formats may contain hidden elements used to disguise malicious payloads. This significantly increases the vulnerability of information systems, as most traditional mechanisms analyse only the image content, ignoring its structural wrapper. Analysis of typical and anomalous field values, such as UserComment, Software, or XMP Payload, made it possible to identify characteristic injection vectors that require dedicated monitoring by detection systems.

Based on a comparative review and logical modelling, a conceptual model for detecting changes in digital image metadata using AI methods was proposed. The architecture, based on MLP, accounted for both numerical and categorical parameters after appropriate pre-processing, including scaling and encoding. The model demonstrated potential in detecting structural anomalies, such as incorrect date formats or suspicious geolocation coordinates, positioning it as a promising solution for practical implementation in the context of modern cyber threats. The results obtained underscored the advisability of focusing not only on file content but also on accompanying technical data, which may serve as an independent source of critically important information.

At the same time, the study identified a number of limitations, including the likelihood of false positives when processing non-standard but legitimate metadata changes, as well as challenges associated with system scalability when working with large volumes of data. To improve the efficiency of future implementations, it would be advisable to consider integrating additional approaches, such as behaviour-based file analysis at higher levels, and optimising computational algorithms to ensure rapid response without sacrificing accuracy. A promising area for further research is the adaptation of the model to new steganographic techniques and its integration into comprehensive cybersecurity systems, including automated tools for monitoring and countering multi-level threats.

ACKNOWLEDGEMENTS

None.

FUNDING
None.

CONFLICT OF INTEREST
None.

REFERENCES

[1] Ahmadi, C., Chen, J.L., & Lin, Y.T. (2024). Securing AI models against backdoor attacks: A novel approach using image steganography. *Journal of Internet Technology*, 25(3), 465-475. doi: 10.53106/160792642024052503012.
[2] Bas, P., Filler, T., & Pevný, T. (2011). Break our steganographic system – the BOSS contest. In *Proceedings of the 13th international conference on information hiding* (pp. 59-70). Berlin: Springer. doi: 10.1007/978-3-642-24178-9_5.
[3] Camera & Imaging Products Association. (2012). *Exchangeable image file format for digital still cameras: Exif Version 2.3*. Retrieved from https://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf.
[4] Carneiro, D., Guimarães, M., Carvalho, M., & Novais, P. (2023). Using meta-learning to predict performance metrics in machine learning problems. *Expert Systems*, 40(1), article number e12900. doi: 10.1111/exsy.12900.
[5] Caviglione, L., & Mazurczyk, W. (2022). Never mind the malware, here's the stegomalware. *IEEE Security & Privacy*, 20(5), 101-106. doi: 10.1109/MSEC.2022.3178205.
[6] Chaganti, R., Ravi, V., & Pham, T.D. (2022). Image-based malware representation approach with EfficientNet convolutional neural networks for effective malware classification. *Journal of Information Security and Applications*, 69, article number 103306. doi: 10.1016/j.jisa.2022.103306.
[7] El Abdelkhalki, J., Ahmed, M.B., & Abdelhakim, B.A. (2022). Image malware detection using deep learning. *International Journal of Communication Networks and Information Security*, 12(2). doi: 10.17762/ijcnis.v12i2.4600.
[8] El-Ghamry, A., Gaber, T., Mohammed, K.K., & Hassanien, A.E. (2023). Optimized and efficient image-based IoT malware detection method. *Electronics*, 12(3), article number 708. doi: 10.3390/electronics12030708.
[9] Fernando, Y., Darwis, D., Mehta, A.R., Wamiliana, W., & Wantoro, A. (2024). A new approach of steganography on image metadata. *International Journal on Informatics Visualization*, 8(2), 968-976. doi: 10.62527/joiv.8.2.2110.
[10] Galli, A., La Gatta, V., Moscato, V., Postiglione, M., & Sperlì, G. (2024). Explainability in AI-based behavioral malware detection systems. *Computers & Security*, 141, article number 103842. doi: 10.1016/j.cose.2024.103842.
[11] International Press Telecommunications Council. (2024). *IPTC photo metadata standard 2024.1*. Retrieved from https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata.
[12] Internet Security Threat Report. (2017). Retrieved from https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf.
[13] Iskanderani, A.I., Mehedi, I.M., Aljohani, A.J., Shorfuzzaman, M., Akther, F., Palaniswamy, T., Latif, S.A., & Latif, A. (2021). Artificial intelligence-based digital image steganalysis. *Security and Communication Networks*, 2021(1), article number 9923389. doi: 10.1155/2021/9923389.
[14] Ispahany, J., Islam, M.R., Islam, M.Z., & Khan, M.A. (2024). Ransomware detection using machine learning: A review, research limitations and future directions. *IEEE Access*, 12, 68785-68813. doi: 10.1109/ACCESS.2024.3397921.
[15] Jian, Y., Kuang, H., Ren, C., Ma, Z., & Wang, H. (2021). A novel framework for image-based malware detection with a deep neural network. *Computers & Security*, 109, article number 102400. doi: 10.1016/j.cose.2021.102400.
[16] Kashtalian, A., Lysenko, S., Savenko, O., Nicheporuk, A., Sochor, T., & Avsiyevych, V. (2024). Multi-computer malware detection systems with metamorphic functionality. *Radioelectronic and Computer Systems*, 2024(1), 152-175. doi: 10.32620/reks.2024.1.13.
[17] Kiltz, S., Dittmann, J., Loewe, F., Heidecke, C., John, M., Mädel, J., & Preißler, F. (2024). Forensic image trace map for image-stego-malware analysis: Validation of the effectiveness with structured image sets. In *Proceedings of the 2024 ACM workshop on information hiding and multimedia security* (pp. 125-130). doi: 10.1145/3658664.3659659.
[18] Kobozieva, A., Bobok, I., & Kushnirenko, N. (2023). Steganalysis method for detecting LSB embedding in digital video, digital image sequence. In *Information Control Systems and Technologies* (pp. 78-90). Odesa: CEUR.
[19] Krasin, I., *et al.* (2017). *OpenImages: A public dataset for large-scale multi-label and multi-class image classification*. Retrieved from https://github.com/openimages/dataset.
[20] Kuznetsov, O., Frontoni, E., & Chernov, K. (2024a). Beyond traditional steganography: Enhancing security and performance with spread spectrum image steganography. *Applied Intelligence*, 54(7), 5253-5277. doi: 10.1007/s10489-024-05415-z.
[21] Kuznetsov, O., Frontoni, E., Chernov, K., Kuznetsova, K., Shevchuk, R., & Karpinski, M. (2024b). Enhancing steganography detection with AI: Fine-tuning a deep residual network for spread spectrum image steganography. *Sensors*, 24(23), article number 7815. doi: 10.3390/s24237815.
[22] Monika, A., & Eswari, R. (2023). An ensemble-based stegware detection system for information hiding malware attacks. *Journal of Ambient Intelligence and Humanized Computing*, 14(4), 4401-4417. doi: 10.1007/s12652-023-04559-z.
[23] Monteiro, J.P., Ramos, D., Carneiro, D., Duarte, F., Fernandes, J.M., & Novais, P. (2021). Meta-learning and the new challenges of machine learning. *International Journal of Intelligent Systems*, 36(11), 6240-6272. doi: 10.1002/int.22549.

[24] Newman, J., Lin, L., Chen, W., Reinders, S., Wang, Y., Wu, M., & Guan, Y. (2019). StegoAppDB: A steganography apps forensics image database. In *Proceedings of the IS&T international symposium on electronic imaging: media watermarking, security, and forensics* (pp. 536-1-536-12). Springfield: Society for Imaging Science and Technology. doi: 10.2352/ISSN.2470-1173.2019.5.MWSF-536.
[25] Płachta, M., Krzemień, M., Szczypiorski, K., & Janicki, A. (2022). Detection of image steganography using deep learning and ensemble classifiers. *Electronics*, 11(10), article number 1565. doi: 10.3390/electronics11101565.
[26] Puchalski, D., Caviglione, L., Kozik, R., Marzecki, A., Krawczyk, S., & Choraś, M. (2020). Stegomalware detection through structural analysis of media files. In *Proceedings of the 15th international conference on availability, reliability and security* (article number 73). New York: Association for Computing Machinery. doi: 10.1145/3407023.3409187.
[27] Roseline, S.A., Geetha, S., Kadry, S., & Nam, Y. (2020). Intelligent vision-based malware detection and classification using deep random forest paradigm. *IEEE Access*, 8, 206303-206324. doi: 10.1109/ACCESS.2020.3036491.
[28] Russo, M., Šrndić, N., & Laskov, P. (2021). Detection of illicit cryptomining using network metadata. *EURASIP Journal on Information Security*, 2021, article number 11. doi: 10.1186/s13635-021-00126-1.
[29] Salem, A.H., Azzam, S.M., Emam, O.E., & Abohany, A.A. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *Journal of Big Data*, 11(1), article number 105. doi: 10.1186/s40537-024-00957-y.
[30] Sarker, I.H. (2023). Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. *Security and Privacy*, 6(5), article number e295. doi: 10.1002/spy2.295.
[31] Sarker, I.H. (2024). *AI-driven cybersecurity and threat intelligence: Cyber automation, intelligent decision-making and explainability*. Cham: Springer. doi: 10.1007/978-3-031-54497-2.
[32] Setiadi, D.R., Ghosal, S.K., & Sahu, A.K. (2025). AI-powered steganography: Advances in image, linguistic, and 3D mesh data hiding – a survey. *Journal of Future Artificial Intelligence and Technologies*, 2(1). doi: 10.62411/faith.3048-3719-76.
[33] Ullah, F., Srivastava, G., & Ullah, S. (2022). A malware detection system using a hybrid approach of multi-heads attention-based control flow traces and image visualization. *Journal of Cloud Computing*, 11(1), article number 75. doi: 10.1186/s13677-022-00349-8.
[34] Vasan, D., Hammoudeh, M., & Alazab, M. (2024). Broad learning: A GPU-free image-based malware classification. *Applied Soft Computing*, 154, article number 111401. doi: 10.1016/j.asoc.2024.111401.
[35] Verma, V., Muttoo, S.K., & Singh, V.B. (2022). Detecting stegomalware: Malicious image steganography and its intrusion in windows. In *Security, privacy and data analytics: Select proceedings of ISPDA 2021* (pp. 103-116). Singapore: Springer. doi: 10.1007/978-981-16-9089-1_9.
[36] Wang, F., & Tang, Y. (2024). Diverse intrusion and malware detection: AI-based and non-AI-based solutions. *Journal of Cybersecurity and Privacy*, 4(2), 382-387. doi: 10.3390/jcp4020019.
[37] Yadav, P., Menon, N., Ravi, V., Vishvanathan, S., & Pham, T.D. (2022). A two-stage deep learning framework for image-based android malware detection and variant classification. *Computational Intelligence*, 38(5), 1748-1771. doi: 10.1111/coin.12532.
[38] Zuppelli, M., Manco, G., Caviglione, L., & Guarascio, M. (2021). Sanitization of images containing stegomalware via machine learning approaches. In *Proceedings of the Italian conference on cybersecurity* (pp. 374-386). London: CEUR.

Оцінка ефективності систем розпізнавання зображень для автоматичного виявлення шкідливих файлів на основі метаданих зображень

Віталій Ясененко

Магістр, старший розробник програмного забезпечення
TP-Link
92618, вул. Технологій, 36, м. Ірвайн, США
https://orcid.org/0009-0004-4801-9541

Анотація. Актуальність дослідження обумовлена зростанням загрози прихованого розповсюдження шкідливих програм через метадані цифрових зображень, що ускладнює їх виявлення стандартними методами. Метою роботи було розробити новий підхід до виявлення шкідливих файлів через аналіз метаданих зображень із використанням методів штучного інтелекту. Для цього було проведено детальний аналіз основних стандартів метаданих, а також визначено уразливі поля, здатні приховувати шкідливий код та які ігноруються традиційними методами безпеки. Результати теоретичного дослідження показали, що найбільш інформативними для виявлення загроз є характеристики метаданих, такі як часові відмітки, геолокаційні координати та дані про пристрої. Окремо було встановлено, що нестандартні значення в полях, наприклад, аномальні часові позначки чи підозрілі кодові позначення, можуть виступати індикаторами шкідливої активності. Було здійснено порівняння традиційних методів виявлення загроз, яке виявило їх низьку ефективність при роботі з метаданими, оскільки ці методи здебільшого орієнтовані на виявлення шкідливих елементів у візуальній частині файлу, а не на аналіз супровідної структури. Розроблена концептуальна модель, орієнтуючись на зазначені характеристики, продемонструвала значний потенціал для ефективного виявлення аномалій і прихованого шкідливого коду в метаданих. Цей підхід дозволяє знижувати кількість помилкових спрацювань, оскільки фокусується не лише на виявленні очевидних відхилень, але й на більш тонких змінах у структурному шарі зображень. Висновки підтверджують, що аналіз супровідної інформації є важливим інструментом для виявлення нових форм загроз. Практична значимість дослідження полягає в можливості використання запропонованої концепції як основи для побудови спеціалізованих систем моніторингу та попередження кіберінцидентів

Ключові слова: стеганографія; шкідливе програмне забезпечення; цифрові зображення; машинне навчання; аномалії в метаданих