Detecting Unknown Attacks: Transformer-Based Image Classifiers with an Out-of-Distribution Detector

Yong-Syuan Chen, Hsiang-Yin Lien, Jo-Yu Li, and Chia-Yu Lin(✉)

Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan
sallylin0121@ncu.edu.tw

Abstract. Traditional attacks such as viruses, trojans, and backdoors remain significant security challenges, especially as systems face both known and emerging threats. Current detection methods often struggle to identify unknown attacks, leaving systems vulnerable. To address this issue, we propose a transformer-based image classifier with an Out-of-Distribution (OOD) detector that uses the Swin Transformer for known attacks and the local outlier factor (LOF) for unknown attacks. Our approach utilizes the Swin Transformer's attention mechanism to capture intricate attack patterns while LOF identifies outliers indicative of new, unseen threats. Through evaluation, our method achieved a 0.97 accuracy in classifying known attacks and about 0.7 accuracy in detecting unknown attacks, demonstrating its potential to significantly improve system security and deepen our understanding of traditional attack behaviors.

Keywords: Intrusion detection · out-of-distribution detection · open set · image classification

1 Introduction

In today's computing environments, system flexibility and reliability are critical as software systems are increasingly adopted. However, this widespread use also exposes systems to significant security risks through exploitable vulnerabilities. As attack vectors evolve and new threats emerge, detecting these unidentified malicious activities has become a crucial aspect of maintaining system integrity and security. Effective detection mechanisms are essential to protect against the ever-changing landscape of cyber threats.

Despite advances in security, traditional detection methods often fall short when it comes to identifying novel or unknown attacks. As attack techniques become more sophisticated, the need for robust systems capable of recognizing and addressing these new threats grows. Strengthening detection capabilities is vital not only for responding to known vulnerabilities but also for preemptively identifying potential security breaches that could exploit system weaknesses.

W.-T. Chu et al. (Eds.): TAAI 2024, CCIS 2415, pp. 319–329, 2025.
https://doi.org/10.1007/978-981-96-4596-1_24

Recent approaches have aimed to improve malware detection and system security. For example, Rahali et al. introduced "MalBERTv2 [10]", a BERT-based model designed for proactive malware detection. Although effective for known attacks, MalBERTv2 faces challenges with new, previously unseen attacks and requires substantial computational resources. Similarly, Seneviratne et al. developed "SHERLOCK [11]", a self-supervised model that converts malware binaries into images, preserving their structure and semantic information. Although this method maintains valuable data integrity, it can still struggle with highly obfuscated or polymorphic malware that significantly alters its appearance. This limitation can result in false negatives and missed detections of new or heavily modified malware variants.

To address these limitations, we propose a supervised Out-of-Distribution (OOD) detector utilizing the Virus MNIST dataset [9], which features image representations of various malicious attack sequences. Our approach employs a deep learning model trained on this dataset to detect and classify both known and unknown attack methods. In addition, we introduce a hybrid intrusion detection system designed to enhance the identification and categorization of unseen threats. By integrating advanced machine learning techniques, our method aims to improve the adaptability and generalization capabilities of the model, addressing the gaps left by current approaches.

We have published the initial idea of a supervised OOD detector at the 2024 IEEE International Conference on Consumer Electronics - Taiwan, as a two-page conference paper [3]. The difference between the previous paper and the current manuscript lies in focusing on a more comprehensive evaluation of various classifiers and OOD detection methods. The previous paper [3] relied only on a Swin Transformer-based classifier paired with selective feature extraction techniques for anomaly detection. The updated system in this paper explores a broader range of classifiers, including ResNet, EfficientNet, and Swin Transformer, to classify known malware. Furthermore, we improve the OOD detection process by employing EfficientNet as a feature extractor integrated with the Local Outlier Factor (LOF) model.

We also demonstrate significant improvements in the updated approach to detecting unknown attacks, achieving a classification accuracy of 0.97 for known attacks and 0.69 for unknown attacks. This enhanced detection capability contributes to the overall robustness of the system by providing a more comprehensive understanding of evolving attack techniques. The adaptability of the system ensures that it remains effective against new and emerging threats, offering long-term security and resilience. Using advanced machine learning techniques, our approach provides a more effective and dynamic solution to modern cybersecurity challenges.

2 Related Works

2.1 Supervised Intrusion Detection

In the past, many rule-based intrusion detection methods were used. With the rise of Artificial Intelligence (AI) technologies, numerous studies have developed AI models to detect malicious attacks. Rahali et al. proposed the "MalBERTv2 model," which uses a public dataset to train a BERT semantic model, with a fully connected layer added at the end of the model as a classifier. This approach utilizes natural language processing (NLP) techniques to proactively detect malware threats, achieving an F1-score of 82% [10]. However, building a semantic model for malware detection requires significant computational resources for training, and semantic models face challenges in identifying new types of attacks.

In supervised malware detection models, experts require that a large amount of labeled data or feature engineering be used to extract malware features for model training. To address this, S. Seneviratne et al. combined self-supervised learning to propose the "SHERLOCK" detection model [11]. SHERLOCK first uses the byte-to-pixel mapping of byteplot to convert the binary files of malware into images, preserving the structural and semantic information of the malware. The converted images are randomly masked by 75% and fed to the encoder of a ViT model. The encoder extracts optimal embeddings, which are then passed to the decoder. The decoder uses these embeddings to map back to the image's pixel values, following a self-supervised learning approach, with the goal that the encoder can reconstruct the original image using only 25% of the information. Finally, the study employs the same ViT encoder classifier used in self-supervised learning as the classification model, with a linear layer added on top to map the embeddings to the number of output classes. Each task is then fine-tuned to adapt to specific classification tasks (binary classification, malware type classification, and malware family classification).

2.2 Out-of-Distribution Detection

One-class classification (OCC) methods [1] have been widely applied in the detection of malicious attacks. As attack techniques evolve, researchers have started to look for methods capable of detecting unknown attacks. To this end, out-of-distribution detectors (OOD detectors) have been developed that focus on distinguishing between in-distribution (ID) and out-of-distribution (OOD) data. This has become a key focus of current research [13]. There are currently two main research directions for applying OCC in OOD detectors. The first approach involves enhancing the feature extraction capabilities of the OCC during the training phase, making it easier for the classifier to identify OOD samples. The second approach focuses on using the feature space or the prediction confidence during the inference phase of the OCC model to determine whether a sample belongs to the OOD.

In research aimed at improving the feature extraction capabilities of OCC, Zheng Zhang et al. proposed the Deep Support Vector Data Description (Deep-SVDD) model [14]. In the improved Deep-SVDD model, data are mapped to a

high-dimensional space where average data are enclosed within a hypersphere, while anomalous data lies outside the hypersphere. The model's objective function is to minimize the volume of the hypersphere (i.e., minimize the distance of data points from the center of the hypersphere) and the reconstruction error. The objective function is as follows, where $f(x_i; W)$ is the output of the network, $g()$ is the reconstruction function, c is the center of the hypersphere, and λ is the regularization parameter. The model achieves more stable training results using the newly designed objective function. Experimental results on the MNIST, Fashion-MNIST, and MVTec datasets demonstrate that this approach effectively enhances the model's ability to identify OOD samples.

$$\min_{W,c} \frac{1}{n}||(f(x_i, W) - c)|| + \lambda||g(f(x_i, W)) - x_i|| \tag{1}$$

During a model's training process, the scale and range of features can significantly impact the performance of the feature extractor. When there is a significant disparity in the range or scale of different features in the dataset, the model tends to overly focus on features with more extensive numerical ranges while neglecting those with smaller ranges. This issue, known as model confusion caused by varying feature scales, can degrade the performance of the feature extractor. To address this problem, W. Hu et al. combined L2-norm regularization and holistic regularization (H-regularization) and proposed the HRN (H Regularization with 2-Norm instance-level normalization) deep classification network training method [4]. L2-norm regularization addresses the confusion caused by different feature scales in the data, ensuring a uniform distribution of feature values for each data point. This helps the model to better learn and distinguish different features, improving overall performance and accuracy. H-regularization controls the output of the encoder network, reducing feature bias caused by extreme values in the input data. For a given one-class dataset, the training loss function of HRN is as follows:

$$L = \mathbb{E}_{x \sim p_x}[-log(Sigmoid(En(x)))]NLL + \lambda \mathbb{E}_{x \sim p_x}[||En(x)||_2^2] \tag{2}$$

This equation emphasizes minimizing the Negative Log Likelihood (NLL) to enhance the importance of training the encoder network on single-class data distributions. Compared to commonly used L2-norm regularization [12], the regularization proposed by W. Hu et al. has been shown to be highly effective in improving the performance of OOD detection in one-class classification.

Some other studies focus on determining whether a sample belongs to OOD by analyzing the feature space or the prediction confidence of a trained OCC model. Peyman Morteza et al. [8] proposed the Gaussian Mixture-based Measurement (GEM) method. GEM models the feature space as class-conditional multivariate Gaussian distributions. For a given input sample x, its energy score $E(x)$ is calculated as:

$$E(x) = -\log \sum_{i=1}^{M} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i), \tag{3}$$

Table 1. Example of Malware Classification

Class	Count	Group	Type	Example
0	2516	Beneware	Good	putty.exe
1	7684	Malware	Adware	IESettings
2	3037	Malware	Trojan	Supreme.exe
3	2404	Malware	Trojan	myfile.exe
4	796	Malware	Installer	myfile.exe
5	6662	Malware	Backdoor	myfile.exe
6	15377	Malware	Crypto	Powershell
7	7494	Malware	Backdoor	BitTorrent.exe
8	2571	Malware	Downloader	myfile.exe
9	3339	Malware	Heuristic	myfile.exe

where $\mathcal{N}(x|\mu_i, \Sigma_i)$ is the multivariate Gaussian distribution with mean μ_i and covariance Σ_i, and π_i is the mixture component weight. The energy score represents the degree of similarity between the input sample and the distribution of the training data. A low energy score indicates that the sample closely matches the training data distribution, while a high energy score suggests that the sample is likely to belong to OOD.

Ryo Kamoi et al. explored the effectiveness of Mahalanobis distance in anomaly detection [5]. The formula for Mahalanobis distance is as follows, where x is the observation, μ is the mean, and Σ^{-1} is the inverse of the covariance matrix.

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{4}$$

Mahalanobis Distance considers the covariance structure of the data, allowing it to identify anomalies in multi-dimensional spaces with varying variances and features.

3 Methods

3.1 Dataset

Virus-MIST [9] is a dataset used for virus image classification. It contains processed and labeled virus images with ten categories, including nine types of malware and one type of benign executable file, amounting to 50,000 image samples. This dataset provides a standardized benchmark for researchers to train and test malware recognition models using machine learning or deep learning methods. Since image data contains structural and semantic information and many well-established image processing and analysis techniques can be applied, converting executable files into images is an effective method for feature extraction and dimensionality reduction.

3.2 Transformer-Based Image Classifier

To enhance the effectiveness of detecting malicious attacks, we propose utilizing attention mechanisms with the Swin Transformer [7], a model designed to improve the computational complexities of the Vision Transformer (ViT) [2], as shown in Fig. 1. Developed by Liu et al., the Swin Transformer introduces a hierarchical structure by progressively merging image patches, allowing it to scale more efficiently.

The Swin Transformer operates in two key steps: W-MSA (window-based multi-head self-attention) and SW-MSA (shifted window-based multi-head self-attention). In the W-MSA step, the input image is divided into windows of varying sizes, and self-attention is applied within each window. The SW-MSA step then shifts these windows, applying self-attention to the patches within the new windows. This shifting mechanism allows the model to capture information across nonoverlapping windows, thereby learning a richer set of features.

Each block in the Swin Transformer integrates both W-MSA and SW-MSA, with a two-layer MLP (multi-layer perceptron) following each MSA module. LayerNorm (LN) is applied before each MSA and MLP module, and residual connections are used after each module to maintain the integrity of the learned features.

Our classifier leverages this architecture to distinguish known attacks while categorizing normal and unknown instances into an "others" class for out-of-distribution (OOD) detection. The Swin Transformer's ability to capture fine-grained details within and across windows makes it a powerful tool for identifying and classifying various types of attacks.

3.3 Out-of-Distribution Detector

In our out-of-distribution (OOD) detector, we further enhance the model's ability to classify instances within the "Others" category, building upon the techniques described in [6]. To improve generalization and mitigate the limitations imposed by data distribution, we incorporate pre-trained models as feature extractors. Specifically, we utilize Swin Transformer, EfficientNet, and ResNet18 to extract diverse features from the input data. These models are chosen for their distinct architectures and strengths. Swin Transformer excels in capturing hierarchical features through attention mechanisms. EfficientNet balances performance and efficiency. ResNet18 offers a straightforward yet powerful residual learning framework. By comparing the effectiveness of these models, our goal is to balance the sensitivity of individual detectors while capturing unique data characteristics, thereby laying a strong foundation for effective anomaly detection.

To address the challenge of feature-scale inconsistencies, we apply H-Regularization with 2-Norm instance-level normalization (HRN) [4]. HRN ensures that the multiple feature sets extracted from different models are normalized and consistent, which is crucial for robust OOD detection. This normalization process improves the comparability of features, making it easier to distinguish between in-distribution (ID) and out-of-distribution (OOD) samples.

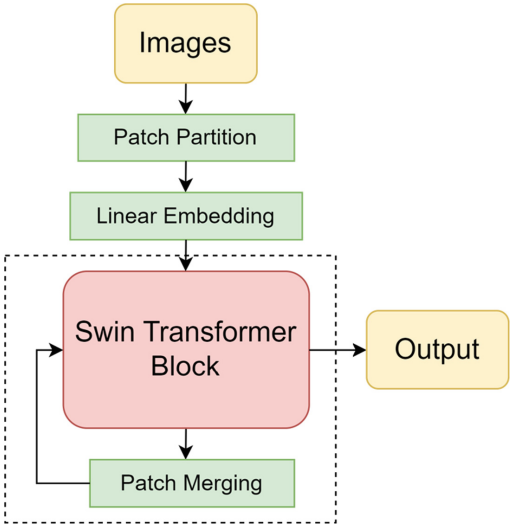

Fig. 1. The architecture of Swin Transformer [7].

Additionally, we improve distance measurement in the feature space by employing enhanced techniques such as Deep Support Vector Data Description (Deep-SVDD) [14] and Mahalanobis Distance (MD) [5]. Deep-SVDD is utilized to learn a compact representation of ID data, effectively drawing a boundary that maximizes the margin between ID and OOD samples. On the other hand, Mahalanobis Distance, known for its effectiveness in anomaly detection, is used to amplify the separation between OOD and standard samples by measuring the distance between a point and the distribution's mean, considering the covariance structure of the data.

By integrating these advanced techniques, our OOD detector not only improves the model's ability to identify anomalies but also enhances the overall robustness and reliability of the detection process.

4 Experiment

For our experiments, we used the Google Colab platform equipped with an A100 GPU to train and evaluate our Swin Transformer model. The Virus-MNIST dataset was used for this study, consisting of images that represent different malicious attack sequences. The dataset includes six categories of known malware samples, which serve as our primary focus for classification.

As shown in Fig. 2, our experiment framework involves two main components: an image classifier and an Out-of-Distribution (OOD) detector. Our image classification system utilizes ResNet, EfficientNet, and Swin Transformer to categorize input images into one of six known virus classes or label them as "other." Simultaneously, in the OOD detection component, we use each of these three classifiers

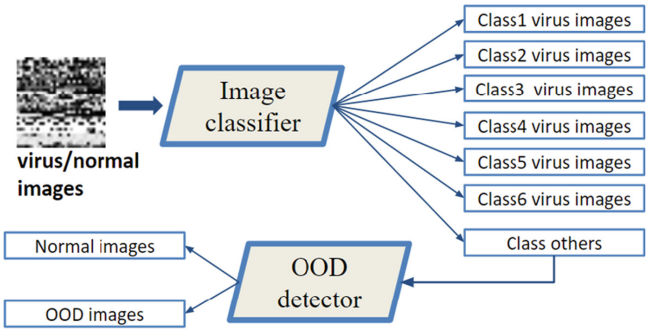

Fig. 2. System architecture

Table 2. Comparison of Classification Models on Virus MNIST Dataset

Model	Dataset	Classification Accuracy
Swin Transforme_t	Virus_MNIST	0.97
EfficientNet_b0	Virus_MNIST	0.71
ResNet18	Virus_MNIST	0.70

as feature extractors. The extracted features are then analyzed through three different OOD detection methods, enhancing our ability to distinguish between normal and unknown samples. These OOD methods leverage the maximum softmax probability as the primary decision criterion. For each validation data point, we record the softmax probability of the correct prediction and set a threshold based on the average softmax probability across the six known categories. If the softmax probability for a sample falls below this threshold, the sample is classified as unknown or OOD; otherwise, it is classified into one of the known malware categories.

The test results confirm the effectiveness of this method, with the Swin Transformer achieving an impressive accuracy of 0.97 in classifying known malware samples while successfully excluding unknown samples Table 2. This high accuracy underscores the model's capability to generalize well across different malware categories, making it a reliable tool for identifying and filtering out unknown threats in real-world applications.

In our OOD detection framework, we utilize EfficientNet as a feature extractor and apply the Local Outlier Factor (LOF) method to identify anomalies within the extracted features. EfficientNet is chosen for its efficiency and ability to extract high-quality features that capture the intricate patterns present in the data. Given that the normal samples in the training set represent a variety of normal behaviors, the resulting feature space is expected to be less dense and more dispersed. This variability requires a method capable of identifying subtle deviations from the norm, which is where LOF comes into play.

Table 3. Feature Extraction Model and Different Unknown Detection Techniques

	LOF	MD	GEM
Swin Transformer_t	0.67	0.60	0.56
EfficientNet_b0	0.69	0.57	0.55
ResNet18	0.51	0.54	0.52

Table 4. OOD Detection Accuracy Using LOF Method

Metric	Result
OOD Accuracy	0.69
OOD Normal Detection	52 / 170
OOD Malicious Detection	702 / 919

To accurately assess anomaly scores, we divide the standard samples from the original dataset into three distinct groups. This division ensures that each group represents a different subset of normal behavior, allowing the LOF model to fit each group separately and capture a wide range of normal variations. For each unknown sample, the LOF model calculates an anomaly score based on how much the sample deviates from the local density of its nearest neighbors within each group.

As a result, each unknown sample is assigned three different anomaly scores, one for each group. To determine the final anomaly score for that sample, we compute the average of these three scores. This averaging process helps to smooth out any potential noise or bias from a single group, providing a more robust and reliable assessment of the sample's likelihood of being an anomaly.

Our result is demonstrated in Table 3 and Table 4, where the calculated anomaly scores mostly differentiate between normal and unknown samples. This multistep approach not only enhances the detection of anomalies but also improves the overall robustness of the model by accounting for the inherent variability in normal sample distributions. Through extensive comparative experiments, the Swin Transformer combined with K-means clustering and LOF achieves superior accuracy in distinguishing both known and unknown attacks. This comparative analysis validates our proposed hybrid approach as the optimal solution to enhance system security.

5 Conclusion

We presented a transformer-based image classifier with an OOD detector to detect unknown attacks. Using the Virus-MNIST dataset as a benchmark, the proposed system combines the Swin Transformer to classify known malicious

samples and methods such as EfficientNet and LOF to detect unknown attacks. From the results of the experiment, our approach achieved an accuracy of 0.97 in classifying known attacks and approximately 0.7 in identifying unknown attacks. By integrating feature extractors and anomaly detection techniques, the generalization and accuracy of the model are enhanced, providing a robust system that addresses the limitations of existing methods and adapts to evolving security threats. Although progress has been made, further improvements in feature extraction and model selection are necessary to enhance detection accuracy and effectiveness.

Acknowledgements. This work is sponsored by the National Science and Technology Council (NSTC) under the project NSTC 113-2222-E-008-002 and NSTC 112-2622-8-A49-021.

References

1. Al-Qudah, M., Ashi, Z., Alnabhan, M., Abu Al-Haija, Q.: Effective one-class classifier model for memory dump malware detection. J. Sensor Actuator Netw. **12**, 5 (2023). https://doi.org/10.3390/jsan12010005
2. Alexey, D.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 (2020)
3. Chen, Y.S., Lien, H.Y., Li, J.Y., Lin, C.Y.: Supervised intrusion detection with out-of-distribution detection for microservices. In: 2024 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), pp. 121–122. IEEE (2024)
4. Hu, W., Wang, M., Qin, Q., Ma, J., Liu, B.: HRN: a holistic approach to one class learning. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
5. Kamoi, R., Kobayashi, K.: Why is the mahalanobis distance effective for anomaly detection? (2020). https://arxiv.org/abs/2003.00402
6. Lin, C.H., Lin, C.Y., Wang, L.J., Kuo, T.T.: Continual learning with out-of-distribution data detection for defect classification. In: 2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), pp. 337–338 (2023). https://doi.org/10.1109/ICCE-Taiwan58799.2023.10226969
7. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows (2021). https://arxiv.org/abs/2103.14030
8. Morteza, P., Li, Y.: Provable guarantees for understanding out-of-distribution detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 7831–7840 (2022)
9. Noever, D., Noever, S.E.M.: Virus-MNIST: a benchmark malware dataset. arXiv preprint arXiv:2103.00602 (2021)
10. Rahali, A., Akhloufi, M.: MalbertV2: Code aware BERT-based model for malware identification. Big Data Cognit. Comput. **7**, 60 (2023). https://doi.org/10.3390/bdcc7020060
11. Seneviratne, S., Shariffdeen, R., Rasnayaka, S., Kasthuriarachchi, N.: Self-supervised vision transformers for malware detection. IEEE Access **10**, 103121-103135 (2022). https://doi.org/10.1109/access.2022.3206445

12. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization (2017). https://arxiv.org/abs/1607.08022
13. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: a survey (2024). https://arxiv.org/abs/2110.11334
14. Zhang, Z., Deng, X.: Anomaly detection using improved deep SVDD model with data structure preservation. Pattern Recognit. Lett. **148** (2021). https://doi.org/10.1016/j.patrec.2021.04.020