MAD-OOD: A Deep Learning Cluster-Driven Framework for an Out-of-Distribution Malware Detection and Classification

Tosin Ige *, Christopher Kiekintveld , Aritran Piplai , Asif Rahman , Olukunle Kolade , Sasidhar Kunapuli

Posted Date: 22 December 2025

doi: 10.20944/preprints202512.1896.v1

Keywords: malware; deep learning; cluster analysis; out-of-distribution (OOD); in-distribution (ID); malware attack

 doi:10.20944/preprints202512.1896.v1

Article

MAD-OOD: A Deep Learning Cluster-Driven Framework for an Out-of-Distribution Malware Detection and Classification

Tosin Ige [1,*], Christopher Kiekintveld [1], Aritran Piplai [1], Asif Rahman [1], Olukunle Kolade [2] and Sasidhar Kunapuli [3]

[1] Dept. of Computer Science, The University of Texas at El Paso, Texas, USA
[2] Dept. of Computer Science, University of North Carolina, North Carolina, USA
[3] Independent, San Jose, CA, USA
* Correspondence: toige@miners.utep.edu

Abstract

Out-of-distribution (OOD) detection remains a critical challenge in malware classification due to the substantial intra-family variability introduced by polymorphic and metamorphic malware variants. Most existing deep learning–based malware detectors rely on closed-world assumptions and fail to adequately model this intra-class variation, resulting in degraded performance when confronted with previously unseen malware families. This paper presents MAD-OOD, a novel two-stage, cluster-driven deep learning framework for robust OOD malware detection and classification. In the first stage, malware family embeddings are modeled using class-conditional spherical decision boundaries derived from Gaussian Discriminant Analysis (GDA), enabling statistically grounded separation of in-distribution and OOD samples without requiring OOD data during training. Z-score–based distance analysis across multiple class centroids is employed to reliably identify anomalous samples in the latent space. In the second stage, a deep neural network integrates cluster-based predictions, refined embeddings, and supervised classifier outputs to enhance final classification accuracy. Extensive evaluations on benchmark malware datasets comprising 25 known families and multiple novel OOD variants demonstrate that MAD-OOD significantly outperforms state-of-the-art OOD detection methods, achieving an AUC of up to 0.911 on unseen malware families. The proposed framework provides a scalable, interpretable, and statistically principled solution for real-world malware detection and anomaly identification in evolving cybersecurity environments.

Keywords: malware; deep learning; cluster analysis; out-of-distribution (OOD); in-distribution(ID); malware attack

1. Introduction

The potency of malware to successfully infiltrate any system no matter how sophisticated made it an indispensable tool available to cybercriminals today, as malware had proven to be highly successful in the extraction of sensitive data which could be used by cybercriminals against their victim. Several machine and deep learning based approaches had been widely proposed to combat the rampant threat of malware attack but the close-world assumption of identical samples are quickly violated whenever state-of-the-art models are exposed to previously unseen out-of-distribution malware attack. Hence, the persistent proliferation of malicious software (malware) poses an ever-evolving challenge to cybersecurity systems worldwide. With the advent of increasingly sophisticated malware variants and the ubiquity of polymorphic and metamorphic transformations, traditional malware detection mechanisms face significant limitations in achieving high detection accuracy, particularly in the context of out-of-distribution (OOD) detection [1–3]. While deep learning and machine learning have emerged as dominant paradigms in malware classification tasks, most state-of-the-art models demonstrate

 doi:10.20944/preprints202512.1896.v1

considerable performance degradation when evaluated on previously unseen malware families or variants not represented in their training data [4–6]. This research is motivated by a critical observation: the existing approaches do not effectively exploit the variation in latent embedding spaces between individual variants within the same malware family—a characteristic that is both prevalent and underutilized [7–12].

Malware classification typically relies on the assumption that samples from a particular class provide a comprehensive representation of that class. However, in the domain of malware, this assumption falls short. Each malware family comprises multiple variants, often generated using obfuscation techniques such as encryption, packing, or polymorphism, which drastically alter the feature representations of these samples while preserving their malicious intent [4,13–16]. Consequently, a single malware variant does not sufficiently represent the diversity of the family in feature space. This insight is a central tenet of our approach, and we argue that it accounts for the poor performance of existing OOD detection methods when applied to malware datasets.

Contribution

The primary contributions of this research are threefold:

- **Novel Framework for OOD Detection:** We propose a two-stage framework for OOD detection and malware classification that exploits the variation in embedding space between variants of the same malware family. The first stage uses spherical decision boundaries defined through Gaussian discriminant analysis to evaluate the likelihood of a test sample being in-distribution, while the second stage integrates this information into a deep neural network for final prediction.
- **Distance-Based Confidence Modeling:** Our method utilizes statistical measures, including Z-score and the coefficient of variation, to determine how far a test sample lies from the centroid of a class distribution. A key insight is the use of multiple z-score comparisons across class boundaries, where a sample must lie within a standard range relative to at least one centroid to be considered in-distribution. This multi-boundary approach accounts for the multidimensional nature of malware distributions.
- **Empirical Validation:** We provide a comprehensive evaluation of our model using malware datasets containing 25 known malware families and multiple out-of-distribution variants. Our model outperforms existing baseline methods in OOD detection, achieving superior AUC scores, particularly for novel malware types.

2. Background and Related Work

Out-of-distribution detection refers to the task of identifying inputs that do not belong to the distribution of the training data. This is critical for deploying machine learning systems in real-world environments, where they frequently encounter data that differs from what they were trained on. Traditional OOD detection methods in computer vision, such as those based on softmax confidence scores, Mahalanobis distances, and reconstruction error from autoencoders, have demonstrated promising results in standard benchmark datasets like CIFAR-10 and ImageNet [17–21]. However, the application of these techniques in the malware domain has been limited and underwhelming due to the distinct characteristics of malware data, including high intra-class variance and distributional overlap among different malware families.

Prior works in malware detection using deep learning have typically employed convolutional neural networks (CNNs) trained on image representations of binaries [10], recurrent neural networks on opcode sequences [22], or graph-based models leveraging control-flow information [23]. Although these approaches achieve high classification accuracy under closed-world settings, their performance declines significantly when tasked with recognizing novel, unseen malware families. One-class classification methods, such as One-Class SVM and Support Vector Data Description (SVDD) [24], have been explored to address this issue [25,26]. SVDD, for example, constructs a hypersphere around training data to capture the normal class distribution and flag samples falling outside the sphere as

 doi:10.20944/preprints202512.1896.v1

anomalous. However, these methods often lack scalability or robustness when confronted with the complex embedding distributions exhibited by modern malware datasets.

Recent advances have sought to integrate distance-based anomaly detection techniques into deep learning pipelines. For instance, works like Deep SVDD [27] and Geometric Transformations for OOD detection [25] attempt to reshape the representation space to better separate in-distribution and out-of-distribution samples. Nonetheless, these methods rarely consider the granularity of variation within individual classes, especially in datasets where class boundaries are inherently fuzzy due to polymorphic behaviors. Our work builds upon this direction by explicitly modeling the intra-class variance within malware families through spherical decision boundaries derived from Gaussian discriminant analysis.

3. Research Methodology

3.1. Spherical Boundary Modeling for Malware Classification

since each family of malware has different variants leading to variation in the embedding spaces **??**. Finding the spherical decision boundary for each class can be instrumental to determines whether a test sample had been previously seen during training or not by integrating the concept of Gaussian discriminant analysis 1 into deep neural networks. This enables it to learn class-conditional distributions that are explicitly modeled as separable Gaussian distributions using the distance of a test sample from each class-conditional distribution to define the confidence score which is eventually used to identifying OOD samples.

$$\widehat{L}(u,v) = \begin{cases} 1 & \text{if } u = v \text{ and } d_v \neq 0 \\ -\dfrac{1}{\sqrt{d_u d_v}} & \text{if } (u,v) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

To effectively model the various spheres within the dataset, each sphere is treated as distinct class-conditional distributions, specifically assuming isotropic Gaussian distributions to ensure that OOD samples are positioned farther from all predefined class distributions without relying on OOD samples for validation.

$$\mathcal{N}(x \mid \mu_k, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right) \tag{2}$$

In the proposed approach, an innovative two-stage framework is introduced to address the limitations of current OOD detection methods in malware datasets. The first stage involves modeling class-conditional distributions using spherical decision boundaries derived from Gaussian Discriminant Analysis (GDA) [28]. This approach assumes isotropic Gaussian distributions for each class, enabling the estimation of distances between test samples and class centroids to determine likelihood and confidence scores for OOD detection.

 doi:10.20944/preprints202512.1896.v1

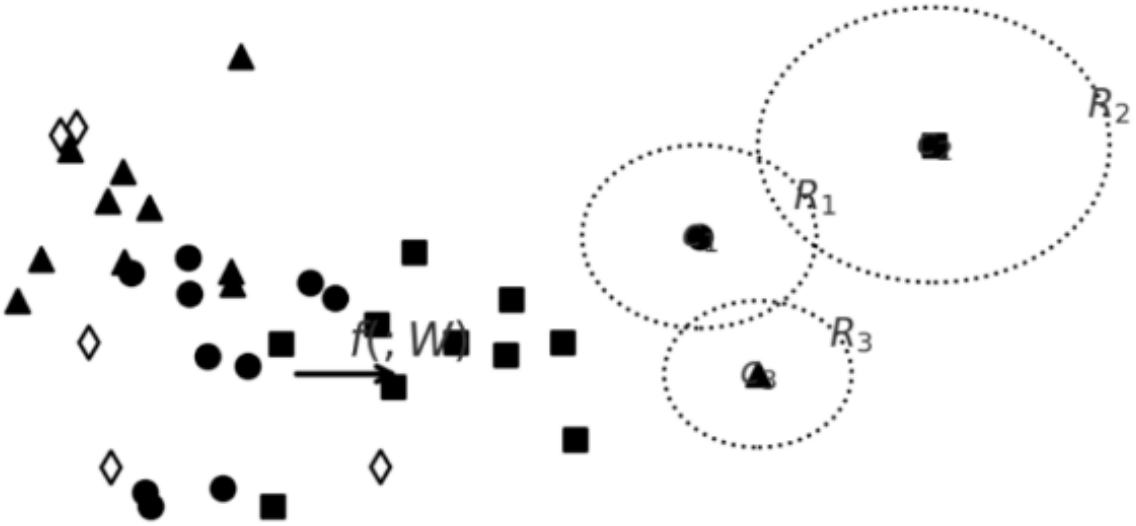

Figure 1. Class conditional probability through Gaussian Discriminant Analysis [28].

This approach provides a statistical basis for OOD classification using metrics such as the z-score, which measures the deviation of a sample from the mean in standard deviation units. By using a standard range (e.g., -1 to +1) to classify samples as in-distribution or out-of-distribution, the model effectively introduces a robust, interpretable decision framework. While similar to prior attempts at distance-based classification, this method innovatively combines these statistical insights with deep learning embeddings tailored to the unique characteristics of malware datasets [29,30]. This approach is closely related to recent advancements in Deep SVDD [5], which integrates the SVDD loss function into neural network training to learn a compact representation of in-distribution data. Similarly, the proposed method leverages the structure of latent embedding spaces to form well-defined spherical clusters, offering a more granular understanding of family-wise variations.

By measuring the distance between a test sample and the derived class-conditional distributions, we can determine the probability and confidence level of the sample belonging to each class. The theoretical foundation for integrating the distance-based classifier into deep neural networks (DNNs) is provided by Gaussian discriminant analysis (GDA).The qualitative analyses on the latent space obtained by this method provide the strong evidence that it places OOD samples more clearly distinguishable from ID samples in the space compared to the others. The task of data description, also referred to as one-class classification, involves defining a representation of the training data and utilizing it to determine whether a test sample belongs to the same distribution. Earlier research in this area primarily explored kernel-based methods, which sought to establish decision boundaries through identified support vectors. One notable approach, supporting vector data description (SVDD), constructs a closed spherical boundary (hypersphere) that encompasses most of the data.

3.2. Cluster-Aided Malware Detection and Final Prediction Network

Following the initial cluster-based classification, the framework employs a second deep learning model that integrates the initial prediction from the clustering algorithm, the output of a supervised DL model, and the input image itself to refine the final prediction. This ensemble-like design enhances the robustness of classification by combining complementary information sources. Notably, unlike conventional vision-based DL models used in malware detection [6], which typically rely solely on image-to-label mappings, this architecture leverages structural knowledge of the latent space through embedding-based features and prior statistical analysis. This strategy offers a unique advantage in handling polymorphic malware samples, which often defy simple visual patterns.

 doi:10.20944/preprints202512.1896.v1

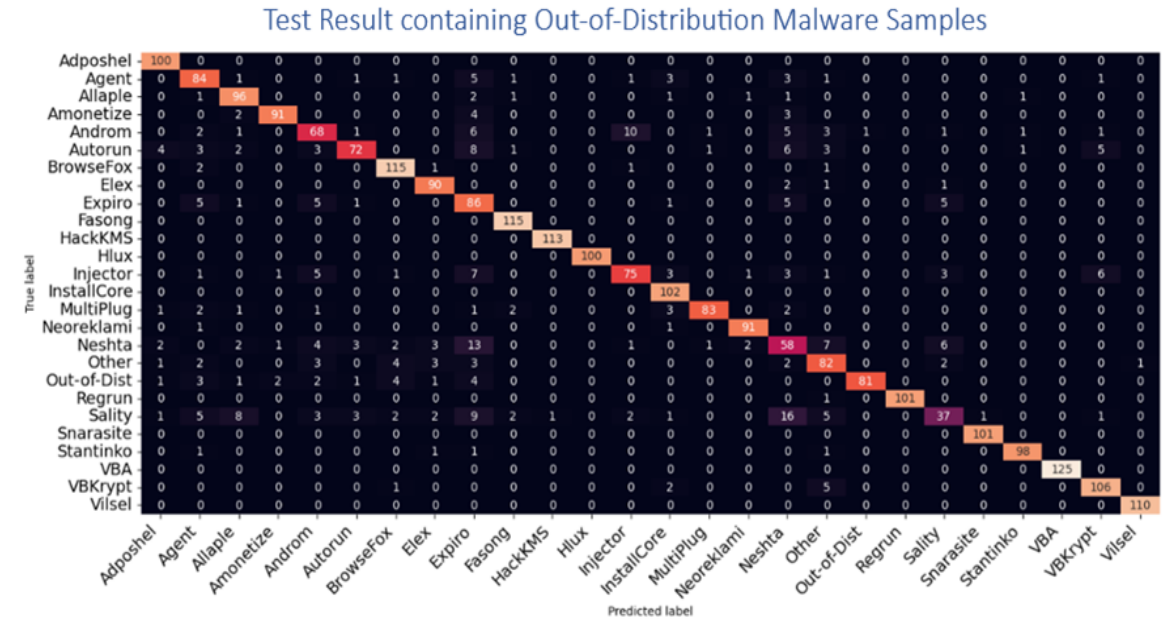

Figure 2. Confusion Matrix on sample test set containing previously unseen out-of-Distribution variants

Table 1. AUC Score on an Out-of-Distribution Malware Classification

Malware Family	AUC Score	Malware Family	AUC Score
Adposhel	0.0059	InstallCore	0.539
Agent	0.105	MultiPlug	0.576
Allaple	0.113	Neoreklami	0.629
Amonetize	0.166	Neshta	0.585
Androm	0.266	VBA	0.619
BrowseFox	0.257	Sality	0.573
Elex	0.305	Snarasite	0.830
Expiro	0.360	Stantinko	0.846
Fasong	0.383	Out-of-Dist. (Novel)	0.911
HackKMS	0.424	VBKrypt	0.916
Hlux	0.463	Vilsel	0.993
Injector	0.499		

Table 2. Comparison of MAD-OOD with other state-of-the-art Out-of-Distribution models on Benchmark Malware Dataset

Method	**Model Comparison and Evaluation**					
	AUC	***AP-Id***	***AP-OOD***	***FPR***	***AR-OOD***	***ACC***
MSP	0.611	0.464	0.322	0.613	0.526	53.82
OE	0.247	0.634	0.709	0.751	0.592	0.407
EnergyOE	0.651	0.660	0.792	0.736	0.808	0.682
OCL	0.637	0.529	0.558	0.771	0.690	0.625
PASCL	0.209	0.405	0.392	0.229	0.393	0.592
OS	0.692	0.442	0.842	0.793	0.712	0.827
Class Prio	0.596	0.376	0.816	0.728	0.693	0.606
BERL	0.846	0.572	0.561	0.807	0.737	0.812
MAD-OOD	0.906	0.951	0.861	0.940	0.817	0.907

a

 doi:10.20944/preprints202512.1896.v1

Table 3. Model Evaluation on Benchmark Malware Dataset

OOD Malware	Proposed Model Generalization					
Dataset	***AUC***	***AP-Id***	***AP-OOD***	***TPR***	***AR-OOD***	***ACC***
MaleVis	0.911	0.864	0.822	0.813	0.926	93.82
BODMAS	0.847	0.834	0.809	0.851	0.792	0.907
Virus-MNIST	0.851	0.860	0.792	0.936	0.908	0.882
Stamina	0.837	0.929	0.858	0.871	0.890	0.925
MalImg	0.906	0.951	0.861	0.940	0.817	0.907

[a]Summary of OOD-MAD model against benchmark malware dataset

4. Metric Evaluation

Our metric evaluation is based on the centroid of each decision boundary as I can now measure the distance between a test sample and the derived class-conditional distributions and utilize it to determine the probability and confidence level of any particular sample belonging to a class to determine whether it is an in-distribution or out-of-distribution sample. In this, utilizing the class conditional distributions to determine the centroid of each decision boundaries. After this comes the challenge of how far new data points could be to any of the centroid before classifying it as in-distribution or an out-of-distribution sample.

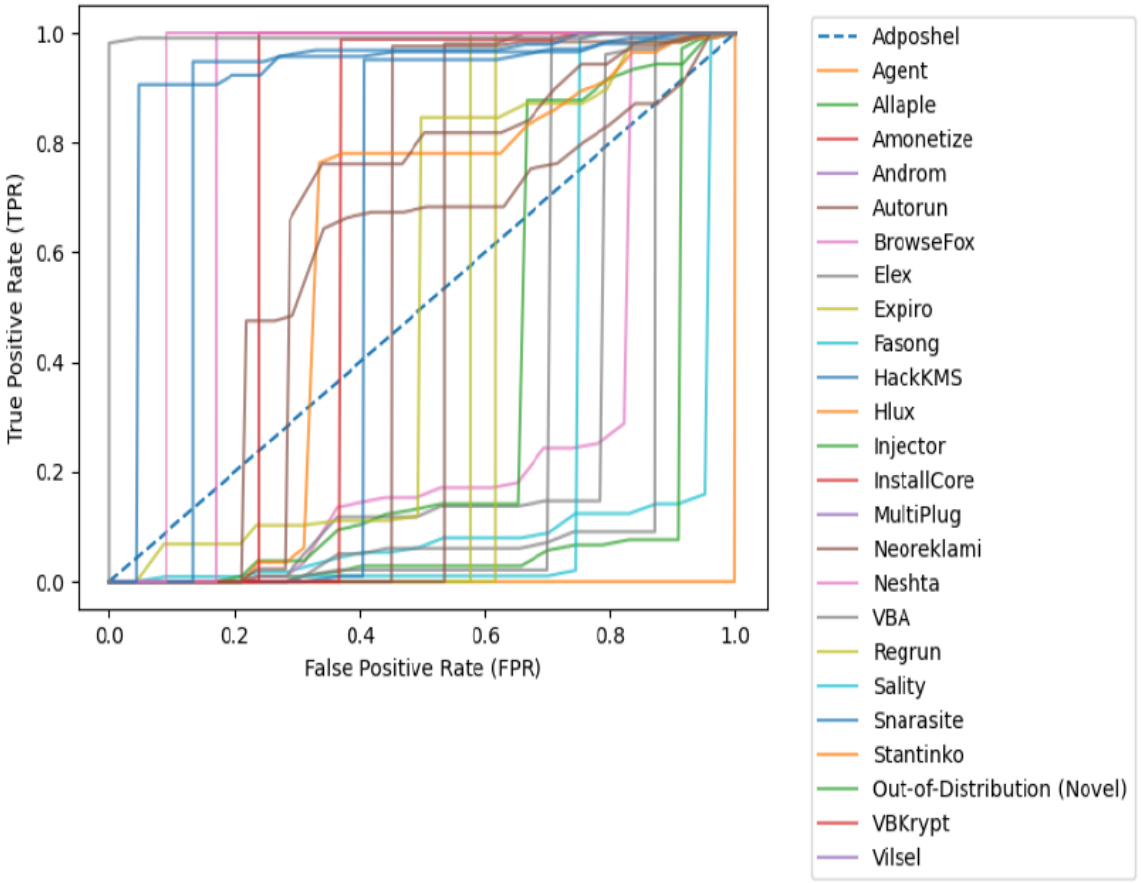

Figure 3. ROC Performance Evaluation of the Proposed Framework on the prediction of Previously Unseen Out-of-Distribution Classification

We computed the standard deviation of data point from any new samples to see how far apart it is to each of the centroid of the spherical boundary, the results was that of a low standard deviation between data points of news sample to the centroid and high standard deviation for others, then arises

 doi:10.20944/preprints202512.1896.v1

the problem of no single rule to determine a low standard deviation across several centroids. This lead to computation of the CV (coefficient of variation) which is the ratio of the distance between the mean and standard deviation considering it has specific range to determine and outliner. Z-score as a statistical method can detect outliers by measuring how far a data point deviates from the mean in terms of standard deviations using the formula:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

Where:

- z z-score (the number of standard deviations a data point is from the mean)
- x is the observed value
- μ is the mean of the data point within the cluster
- σ is the standard deviation of the new data point to the centroid cluster standard deviation

- **Computation of the Mean and Standard Deviation:**Calculate the mean (μ) and standard deviation (σ) of the selected feature across a dataset of benign and potentially malicious samples.
- **Z-score computation:**Determine the Z-score for each family represented by the centroid.
- **Pre-Determine in-distribution Threshold:** Common thresholds for outlier detection
  Z > 1 (or Z < -1) → Possible outlier
  Z > 2 → Highly suspicious
- **Initial Input Classification:** If a sample has a Z-score beyond the chosen threshold, it is flagged as an outlier and so is initially classify a possible out-of-distribution sample

So, to compare new data point from the test set, I use z-score because it has a standard range to determine an outlier data point. z-score (z) = (new data point - mean) / standard deviation The standard to determine outlier in z-score is that anything outside the range of -1 and +1 is an outliner. Hence, for any sample to be classify as an in-distribution sample, at least one of the z-score to each of the centroid of the spherical boundary must be within the range of -1 and +1. If all of the z-scores falls outside the -1 and +1 range, then it is consider as an out of distribution sample. If the z-score is less than -1 or greater than 1, the data point is an outlier as it is further away from the mean than majority of the other data points.

4.1. Training a Computer Vision Based-Deep Learning Classifier for Initial family Prediction

Splitting Data Train, validation, and test sets are carefully divided into: Train & Validation Sets: Contain only in-distribution (known malware and benign) data along with sample representation from out-of-distribution set. Test Set: Contains both in-distribution and OOD samples for evaluation.

Training Procedure Step 1: Train on In-Distribution Data Loss Function: Use cross-entropy for classification (malware vs. benign). Optimizer: Adam and SGD with learning rate scheduling. Regularization: we applied dropout, and batch normalization

Evaluation Metrics In-Distribution Performance: Accuracy, and Confusion matrix for malware classification. OOD Detection Performance: AUROC (Area Under ROC Curve): Measures separation between in-distribution and OOD

4.2. Training a Computer Vision Based-Deep Learning Classifier for Final Prediction

Second Model is a deep neural network-based model which takes three arguments Argument 1-prediction from the cluster analysis Argument 2: prediction output from the first model Argument 3: Input image Unlike the first model, we make use of the new embedding spaces from the spherical boundary to train the second model since the final prediction comes from the second model. There are two stages to it here i. Initial prediction from the second model, since the second model is train on the new embedding spaces, We can expect a better predictive output from it ii. z-score Calculation

While previous work has explored embedding-based OOD detection [3][4], few have addressed the unique challenges posed by malware datasets. This study fills a critical gap by acknowledging

 doi:10.20944/preprints202512.1896.v1

and explicitly modeling the intra-family variation in malware, which is often overlooked in favor of simplifying assumptions. The integration of GDA-inspired decision boundaries, z-score analysis, and a multi-input deep learning classifier offers a holistic approach that outperforms existing methods in separating in-distribution malware from novel, unseen threats. Moreover, the empirical results support the efficacy of the proposed method. For instance, the Area Under the Curve (AUC) scores for known malware families such as Regrun (0.761), Snarasite (0.830), and Stantinko (0.846) indicate improved separability in the latent space, while the detection score for novel OOD samples reaches up to 0.911—highlighting its practical potential in real-world malware detection.

5. Contributions over Prior Work

This research introduces a novel two-stage framework for out-of-distribution (OOD) detection in malware classification by exploiting the inherent variability among malware variants within the same family. Unlike traditional models that assume class uniformity, this approach integrates unsupervised clustering with deep neural networks and Gaussian discriminant analysis (GDA) to establish spherical decision boundaries around malware family embeddings, effectively identifying unseen samples without requiring OOD data during training. A z-score–based statistical analysis enhances discrimination between in-distribution and outlier data. The second stage refines classification using enriched embeddings and multi-source inputs, resulting in superior predictive and OOD detection performance. Achieving an AUC score of 0.911, the model outperforms conventional methods, offering a robust, scalable, and statistically grounded solution for accurate malware detection, with potential extensions to real-time cybersecurity applications.

References

1. Zhou, Y. Rethinking reconstruction autoencoder-based out-of-distribution detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7379–7387.
2. Ige, T.; Kiekintveld, C.; Piplai, A.; Wagler, A.; Kolade, O.; Matti, B.H. An in-Depth Investigation Into the Performance of State-of-the-Art Zero-Shot, Single-Shot, and Few-Shot Learning Approaches on an Out-of-Distribution Zero-Day Malware Attack Detection. In Proceedings of the 2024 International Symposium on Networks, Computers and Communications (ISNCC). IEEE, 2024, pp. 1–6.
3. Ige, T.; Kiekintveld, C.; Piplai, A.; Wagler, A.; Kolade, O.; Matti, B.H. Towards an in-depth evaluation of the performance, suitability and plausibility of few-shot meta transfer learning on an unknown out-of-distribution cyber-attack detection. In Proceedings of the 2024 International Symposium on Networks, Computers and Communications (ISNCC). IEEE, 2024, pp. 1–6.
4. Ige, T.; Kiekintveld, C.; Piplai, A. An investigation into the performances of the state-of-the-art machine learning approaches for various cyber-attack detection: A survey. In Proceedings of the 2024 IEEE International Conference on Electro Information Technology (eIT). IEEE, 2024, pp. 135–144.
5. Ige, T.; Kiekintveld, C.; Piplai, A. Deep learning-based speech and vision synthesis to improve phishing attack detection through a multi-layer adaptive framework. *arXiv preprint arXiv:2402.17249* **2024**.
6. Ige, T.; Kiekintveld, C. Performance comparison and implementation of bayesian variants for network intrusion detection. In Proceedings of the 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings). IEEE, 2023, pp. 1–5.
7. Ige, T.; Kolade, A.; Kolade, O. Enhancing border security and countering terrorism through computer vision: A field of artificial intelligence. In *Proceedings of the Computational Methods in Systems and Software*; Springer, 2022; pp. 656–666.
8. Ige, T.; Sikiru, A. Implementation of data mining on a secure cloud computing over a web API using supervised machine learning algorithm. In Proceedings of the Computer Science On-line Conference. Springer, 2022, pp. 203–210.
9. Nguyen, A.T.; Lu, F.; Munoz, G.L.; Raff, E.; Nicholas, C.; Holt, J. Out of distribution data detection using dropout bayesian neural networks. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 7877–7885.
10. Wood, D.; Kapp, D.; Kebede, T.; Hirakawa, K. LMP-GAN: Out-of-Distribution Detection for Non-Control Data Malware Attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.

 doi:10.20944/preprints202512.1896.v1

11. Park, S.; Gondal, I.; Kamruzzaman, J.; Oliver, J. Generative malware outbreak detection. In Proceedings of the 2019 IEEE International Conference on Industrial Technology (ICIT). IEEE, 2019, pp. 1149–1154.
12. Ige, T.; Marfo, W.; Tonkinson, J.; Adewale, S.; Matti, B.H. Adversarial sampling for fairness testing in deep neural network. *arXiv preprint arXiv:2303.02874* **2023**.
13. Datta, E.; Hennig, J.; Domschot, E.; Mattes, C.; Smith, M.R. Topology of Out-of-Distribution Examples in Deep Neural Networks. *arXiv preprint arXiv:2501.12522* **2025**.
14. Shafiq, M.Z.; Khayam, S.A.; Farooq, M. Embedded malware detection using markov n-grams. In Proceedings of the International conference on detection of intrusions and malware, and vulnerability assessment. Springer, 2008, pp. 88–107.
15. Kan, Z.; Pendlebury, F.; Pierazzi, F.; Cavallaro, L. Investigating labelless drift adaptation for malware detection. In Proceedings of the Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, 2021, pp. 123–134.
16. Ige, T.; Adewale, S. AI powered anti-cyber bullying system using machine learning algorithm of multinomial naïve Bayes and optimized linear support vector machine. *arXiv preprint arXiv:2207.11897* **2022**.
17. Fort, S.; Ren, J.; Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems* **2021**, *34*, 7068–7081.
18. Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems* **2022**, *35*, 32598–32611.
19. Yang, J.; Zhou, K.; Li, Y.; Liu, Z. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision* **2024**, *132*, 5635–5662.
20. Adewale, S.; Ige, T.; Matti, B.H. Encoder-decoder based long short-term memory (lstm) model for video captioning. *arXiv preprint arXiv:2401.02052* **2023**.
21. Okomayin, A.; Ige, T. Ambient technology & intelligence. *arXiv preprint arXiv:2305.10726* **2023**.
22. Karunanayake, N.; Gunawardena, R.; Seneviratne, S.; Chawla, S. Out-of-distribution data: an acquaintance of adversarial examples-a survey. *ACM Computing Surveys* **2025**, *57*, 1–40.
23. Um, D.; Lim, J.; Kim, S.; Yeo, Y.; Jung, Y. Spreading Out-of-Distribution Detection on Graphs. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
24. Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S.A.; Binder, A.; Müller, E.; Kloft, M. Deep one-class classification. In Proceedings of the International conference on machine learning. PMLR, 2018, pp. 4393–4402.
25. Golan, I.; El-Yaniv, R. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems* **2018**, *31*.
26. Liang, S.; Li, Y.; Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* **2017**.
27. Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems* **2018**, *31*.
28. Lee, D.; Yu, S.; Yu, H. Multi-class data description for out-of-distribution detection. In Proceedings of the Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1362–1370.
29. Ige, T. Exploiting the In-Distribution Embedding Space with Deep Learning and Bayesian Inference for Detection and Classification of an Out-of-Distribution Malware (Extended Abstract). *PhilArchive* **2024**. Extended abstract.
30. Ige, T. Impact of Variation in Vector Space on the Performance of Machine and Deep Learning Models on an Out-of-Distribution Malware Attack Detection. *PhilArchive* **2025**. Forthcoming IEEE Conference Proceeding.