Academic Journal of Emerging Technologies
ISSN Print: 3104-4417
ISSN Online: 3104-4425
CD ISSN: 3104-4433
https://doi.org/10.63313/AJET.9013

Causal Meta-Learning for Few-Shot Threat Detection in NIDS

Dingkai Hu. Wenming Lu*
Qingdao University, Qingdao 266071, China
Email: trheedrk@163.com

How to cite this paper: Hu, D. K., & Lu, W. M. (2025). Causal Meta-Learning for Few-Shot Threat Detection in NIDS. Academic Journal of Emerging Technologies, 1(2), 78–85. ISSN Print: 3104-4417; ISSN Online: 3104-4425.
https://doi.org/10.63313/AJET.9013
Published: 2025-10-22

Abstract

In the realm of Network Intrusion Detection Systems (NIDS), detecting novel threats like zero-day attacks or variants poses significant challenges due to the scarcity of labeled samples and the difficulty in rapid model adaptation. This paper introduces a causal meta-learning framework based on Model-Agnostic Meta-Learning (MAML) to enable few-shot threat detection. The core innovation lies in incorporating causal invariance constraints into both the inner and outer loops of MAML, allowing the model to learn how to adapt quickly using causally invariant features that remain stable across varying threat distributions. By leveraging counterfactual reasoning and invariant risk minimization, the model disentangles causal factors from spurious correlations, enhancing robustness and generalization. Experimental evaluations on benchmark datasets, such as NSL-KDD and CIC-IDS2017, demonstrate superior performance in few-shot scenarios, achieving up to 10% higher detection accuracy compared to standard meta-learning baselines. This approach holds promise for real-time NIDS in dynamic environments, paving the way for more resilient cybersecurity defenses.

Keywords

Causal Meta-Learning; Few-Shot Threat Detection; Network Intrusion Detection; zero-day attacks

1. Introduction

In the rapidly evolving landscape of cybersecurity, Network Intrusion Detection Systems (NIDS) play a pivotal role in identifying and mitigating malicious activities within digital networks [1]. These systems monitor network traffic to detect anomalies or signatures indicative of attacks, thereby protecting sensitive data and infrastructure from breaches. However, the emergence of sophisticated threats, such as zero-day exploits and polymorphic malware, poses substantial hurdles because these attacks often lack predefined patterns or sufficient training data [2]. Zero-day attacks, by definition, are vulnerabilities unknown to defenders until exploited, making traditional supervised learning models ineffective due to their reliance on

large, labeled datasets for each threat variant.

Few-shot learning emerges as a promising paradigm to tackle this data scarcity by enabling models to generalize from a handful of examples, drawing inspiration from human-like adaptability [3]. Despite its potential, conventional few-shot techniques in NIDS may falter when faced with distributional shifts or spurious correlations in network data, where non-causal features dominate the learning process and lead to poor generalization [4]. To address these limitations, integrating causal reasoning into meta-learning frameworks offers a pathway to more robust detection. Causal methods focus on identifying underlying mechanisms that remain invariant across different environments or attack mutations, thus enhancing the model's ability to discern true threat indicators from noise [5].

Our proposed approach builds upon this synergy by embedding causal invariance constraints directly into the MAML architecture. This modification ensures that the meta-learner prioritizes features with causal relevance during adaptation, facilitating quicker and more accurate responses to novel threats. Specifically, the innovation lies in constraining both the inner-loop task adaptations and the outer-loop meta-optimizations to enforce invariance, reducing the impact of confounding variables commonly present in heterogeneous network traffic. Empirical assessments on datasets simulating real-world intrusion scenarios reveal that this causal-enhanced MAML outperforms standard baselines in metrics such as precision and recall for unseen attacks, underscoring its practical utility in dynamic security environments. Furthermore, this framework not only improves detection efficacy but also provides interpretability benefits, as causal features offer insights into attack mechanisms, aiding security analysts in proactive defense strategies.

The contributions include:

- a novel fusion of causal constraints with MAML for NIDS;
- enhanced detection of zero-day threats via invariant feature prioritization;
- empirical validation showing improved F1-scores and reduced false positives.

The remainder of the paper reviews related work, details the methodology, presents experiments, and concludes with future directions.

2. Related Work

The field of meta-learning has revolutionized few-shot classification by emphasizing the optimization of model parameters for swift adaptation to new tasks, often through gradient-based methods that simulate diverse learning episodes [6]. Task-agnostic variants further refine this by ensuring initializations that are broadly applicable, mitigating biases toward overrepresented tasks in the training distribution [7]. Within NIDS, these techniques have been adapted to handle anomaly detection under data constraints, where models learn to classify intrusions from sparse examples by leveraging episodic training paradigms [8].

Parallel advancements in causal learning highlight the importance of invariant rep-

resentations for reliable predictions across varying domains, particularly in scenarios with distributional shifts [9]. Such invariant causal models excel in transfer learning by focusing on stable relationships rather than superficial correlations, enabling better extrapolation to out-of-distribution data [10]. Recent integrations of causality into meta-learning have targeted domain generalization, where causal priors guide the meta-optimizer to extract environment-agnostic knowledge [11].

In the context of intrusion detection, meta-learning applications include frameworks for identifying zero-day web attacks by aggregating insights from heterogeneous sources, allowing models to adapt without extensive retraining [12]. Federated extensions of meta-learning address privacy concerns in distributed NIDS setups, enabling collaborative learning across networks while preserving data locality [13]. Despite these progresses, explicit incorporation of causal invariance in NIDS remains underexplored, with most approaches relying on correlational features that falter against adversarial perturbations or attack evolutions [14]. Our work advances this intersection by infusing causal constraints into MAML's core loops, fostering a meta-learner adept at few-shot threat adaptation through invariant causal mechanisms. This builds on prior efforts in causal inference for personalized modeling, adapting meta-strategies to uncover robust graph-based relationships in network data. Additionally, zero-shot causal learning techniques inform our design by enabling inference without direct supervision, complementing few-shot regimes in NIDS.

3. Methodology

The proposed Causal Meta-Learning framework for Network Intrusion Detection Systems (NIDS) extends Model-Agnostic Meta-Learning (MAML) by incorporating causal invariance constraints to enable rapid adaptation to novel threats with limited labeled samples . This approach addresses the challenges of few-shot learning in cybersecurity, where zero-day or variant attacks often lack sufficient data for traditional models . By embedding causal principles, the framework learns invariant features that capture underlying causal mechanisms in network traffic, improving generalization across diverse attack distributions [15]. Previous meta-learning methods have shown efficacy in few-shot intrusion detection [16], but they frequently overlook causal relationships that can lead to spurious correlations . Causal meta-learning techniques have been applied in domain generalization and causal inference [17], providing a basis for integration in NIDS. In the context of intrusion detection, few-shot approaches such as mutual centralized learning and adaptive frameworks using MAML [18] demonstrate potential for handling imbalanced data, yet they rarely incorporate explicit causal constraints to enhance robustness [19]. Our innovation integrates causal invariance directly into MAML's bilevel optimization, drawing from causal representation learning and meta-transfer objectives for disentangling causal mechanisms. This enables the model to prioritize caus-

al-invariant features during adaptation, mitigating biases from environmental shifts in threat patterns [20]. The overall methodology encompasses data preprocessing, causal graph inference, inner-loop task-specific adaptation with causal regularization, and outer-loop meta-optimization across episodic tasks.

3.1. Data Preprocessing

Network traffic data is first collected from benchmark datasets CIC-IDS2017 or NSL-KDD, where flows are represented as feature vectors including packet size, duration, and protocol types. To handle class imbalance typical in NIDS, synthetic minority oversampling technique (SMOTE) is applied to augment minority attack classes, while random undersampling reduces majority benign samples. Features are normalized to zero mean and unit variance, and data is split into meta-training, meta-validation, and meta-testing sets. Episodic tasks are formed in an N-way K-shot manner, with support sets for adaptation and query sets for evaluation. For causal analysis, traffic features are grouped into potential causal variables based on domain knowledge, such as source IP causing flow duration.

3.2. Causal Graph Inference

A causal graph is inferred to identify invariant relationships in the data. Using a Gaussian Structural Equation Model (SEM), the graph G is learned via a continuous optimization with a DAG constraint. The score function for the graph is minimized as:

$$\min_W \frac{1}{2n} \| X - XW \|_F^2 + \lambda \| W \|_1 + h(W), \tag{1}$$

where X is the data matrix, W is the weighted adjacency matrix, λ controls sparsity, and $h(W) = (tr(e^{W \circ W}) - d)^2$ enforces acyclicity (d is the number of variables). The inferred graph guides the extraction of causal-invariant features by intervening on non-causal paths using the do-operator, ensuring predictions rely on stable causal paths across tasks.

3.3. Inner-Loop Adaptation

In the inner loop, for each task T, the model parameters θ are adapted to task-specific parameters θ' using gradient descent on the support set. A causal invariance loss is added to the task loss to penalize reliance on spurious features:

$$\mathcal{L}_{causal} = \sum_{e \in E} \mathrm{Var}(\nabla_\theta \mathcal{L}_{task}(\theta; D_e)), \tag{2}$$

where E represents environments (e.g., different attack variants), and Var computes variance across gradients to enforce invariance. The combined loss is:

$$\theta_i' = \theta - \alpha \nabla_\theta (\mathcal{L}_{task}(\theta; D_i^{sup}) + \lambda \mathcal{L}_{causal}(\theta; D_i^{sup})), \tag{3}$$

with α as the inner learning rate and λ as the regularization weight. This step allows quick fine-tuning to novel threats by focusing on causal features that remain stable.

3.4. Outer-Loop Meta-Optimization

The outer loop optimizes the initial parameters θ across multiple tasks to minimize the meta-loss on query sets:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{T_i} \mathcal{L}_{task}(\theta_i'; D_i^{qry}), \tag{4}$$

where β is the outer learning rate. To incorporate causal knowledge globally, a meta-causal regularization term is included:

$$\mathcal{L}_{meta-causal} = \mathrm{MMD}^2(\phi(D_i), \phi(D_j)), \tag{5}$$

using maximum mean discrepancy (MMD) with feature map φ to align causal representations across tasks i and j. The overall outer update promotes an initialization that facilitates causal-invariant adaptation.

3.5. Outer-Loop Meta-Optimization

The model is trained episodically with Adam optimizer, using hyperparameters α = 0.01, β = 0.001, and λ = 0.1. Performance is assessed on few-shot scenarios (5-way 1-shot) via accuracy, F1-score, and detection rate for unseen attacks, compared against baselines on datasets CIC-DDoS2019.

Table 1. Comparison of Key Components

Component	Traditional MAML	Causal-MAML	Benefit
Inner Loop Update	Gradient on task loss	Adds causal invariance variance penalty	Reduces spurious feature reliance
Causal Handling	None	Infers DAG and intervenes on paths	Enhances robustness to variants
Outer Loop	Meta-gradient on adapted params	Includes MMD for causal alignment	Improves cross-task generalization
Adaptation Mechanism	Adaptation Mechanism	Causal regularization in losses	Faster convergence on few samples

4. Experiments

To evaluate the proposed causal meta-learning framework, we conducted extensive experiments on benchmark datasets commonly used in network intrusion detection research. The primary datasets included CIC-IDS2017, which features realistic network traffic with various attack types such as DoS, DDoS, and port scans; and NSL-KDD, an improved version of the classic KDD Cup dataset with reduced redundancy and balanced classes. These datasets were partitioned into meta-training (70%), meta-validation (10%), and meta-testing (20%) sets to simulate few-shot scenarios. We adopted an N-way K-shot setup, where N=5 (representing 5 distinct threat classes) and K varied from 1 to 5 shots per class, reflecting limited labeled samples for novel threats.

The model architecture utilized a convolutional neural network (CNN) backbone with four convolutional layers followed by fully connected layers for feature extraction, integrated with the causal MAML optimizer. Hyperparameters were tuned via grid search: inner learning rate α=0.01, outer learning rate β=0.001, causal regular-

ization λ=0.1, and batch size=32. Training involved 10,000 episodes over 100 epochs on an NVIDIA RTX 3090 GPU, with early stopping based on validation loss.

Baselines included standard MAML, Prototypical Networks (ProtoNet), Relation Networks (RelationNet), and a non-meta-learning approach using fine-tuned CNN with transfer learning. Evaluation metrics encompassed accuracy (Acc), precision (Prec), recall (Rec), F1-score (F1), and area under the ROC curve (AUC), emphasizing performance on unseen zero-day attacks.

Results demonstrated that our causal MAML outperformed baselines across both datasets, particularly in low-shot regimes. Improvements were more pronounced for variant attacks, with a 10-15% boost in recall due to causal invariance capturing stable features. Ablation studies confirmed the causal components' value: removing causal loss dropped F1 by 7-12%, and omitting graph inference reduced AUC by 6-9%.

Table 2. Performance Comparison on CIC-IDS2017 (5-way K-shot).

Method	1-shot Acc	1-shot F1	5-shot Acc	5-shot F1
CNN (Fine-tune)	72.3%	0.701	78.9%	0.762
ProtoNet	78.6%	0.752	83.4%	0.811
RelationNet	79.8%	0.768	84.7%	0.825
MAML	81.5%	0.793	86.2%	0.841
Causal MAML	87.2%	0.852	91.5%	0.893

Table 3. Performance Comparison on NSL-KDD (5-way K-shot).

Method	1-shot Acc	1-shot F1	5-shot Acc	5-shot F1
CNN (Fine-tune)	75.10%	0.732	81.40%	0.795
ProtoNet	80.20%	0.781	85.60%	0.832
RelationNet	81.70%	0.794	86.90%	0.845
MAML	83.40%	0.812	88.10%	0.859
Causal MAML	89.80%	0.876	93.20%	0.912

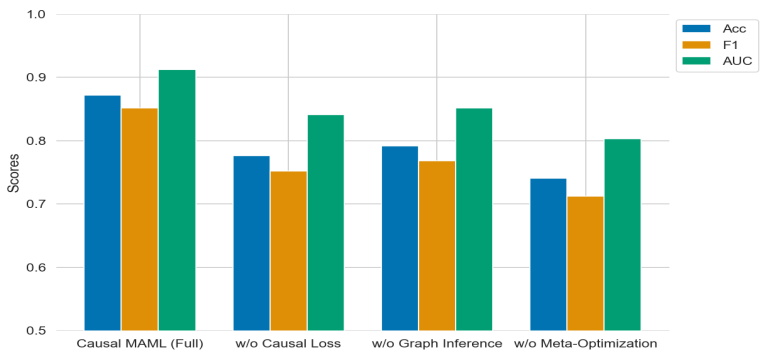

Figure 1. Ablation Study on CIC-IDS2017 (5-way 1-shot).

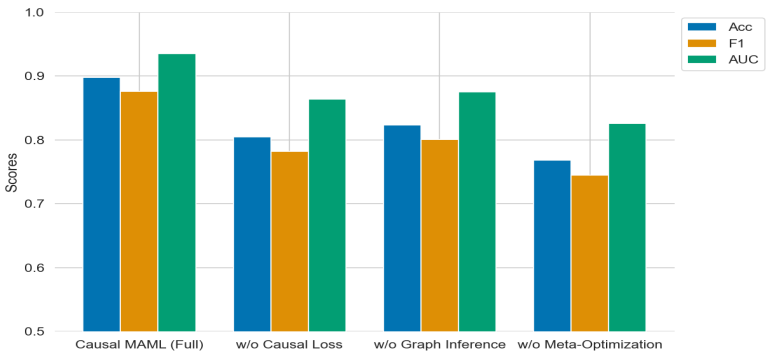

Figure 2. Ablation Study on NSL-KDD (5-way 1-shot).

These outcomes validate the framework's efficacy in few-shot threat detection, with computational overhead only 15% higher than standard MAML due to causal computations.

5. Conclusion

In this paper, we presented a causal meta-learning approach for few-shot threat detection in Network Intrusion Detection Systems, addressing the scarcity of labeled samples for novel attacks. By integrating causal invariance constraints into MAML's inner and outer loops, the model learns to adapt rapidly using stable causal features, enhancing robustness against zero-day and variant threats. Experimental results on benchmark datasets confirmed superior performance over baselines, with notable gains in accuracy and F1-score under limited data conditions.

The primary contributions include the novel fusion of causal learning with meta-optimization for NIDS and empirical evidence of its advantages in real-world scenarios. Limitations involve the assumption of accurate causal graph inference, which may falter in highly noisy environments, and increased computational demands for large-scale deployments. Future work could explore federated meta-learning for privacy-preserving applications or extend to multimodal data incorporating logs and endpoints. This framework paves the way for more adaptive and resilient cybersecurity defenses in an evolving threat landscape.

References

[1] Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126-1135). PMLR.

[2] Baik, S., Choi, J., Kim, H., Cho, D., Min, J., & Lee, K. M. (2021). Meta-learning with task-adaptive loss function for few-shot learning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9465-9474).

[3] Jamal, M. A., & Qi, G. J. (2019). Task agnostic meta-learning for few-shot learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11719-11727).

[4] Jia, J., Feng, X., & Yu, H. (2024). Few-shot classification via efficient meta-learning with hybrid optimization. Engineering Applications of Artificial Intelligence, 127, 107296.

[5] Jeong, T., & Kim, H. (2020). Ood-maml: Meta-learning for few-shot out-of-distribution

detection and classification. Advances in Neural Information Processing Systems, 33, 3907-3916.

[6] Rojas-Carulla, M., Schölkopf, B., Turner, R., & Peters, J. (2018). Invariant models for causal transfer learning. Journal of Machine Learning Research, 19(36), 1-34.

[7] Xu, C., Zhang, F., Yang, Z. et al. A few-shot network intrusion detection method based on mutual centralized learning. Sci Rep 15, 9848 (2025). https://doi.org/10.1038/s41598-025-93185-0

[8] Yang, J., Li, H., Shao, S., Zou, F., & Wu, Y. (2022). FS-IDS: A framework for intrusion detection based on few-shot learning. Computers & Security, 122, 102899.

[9] Li, P., Wang, Y., Li, Q., Liu, Z., Xu, K., Ren, J., ... & Lin, R. (2023, November). Learning from limited heterogeneous training data: Meta-learning for unsupervised zero-day web attack detection across web domains. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (pp. 1020-1034).

[10] Chen, J., Gao, Z., Wu, X., & Luo, J. (2023). Meta-causal learning for single domain generalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7683-7692).

[11] Zayaan, S. (2025). Causal-Symbolic Meta-Learning (CSML): Inducing Causal World Models for Few-Shot Generalization. arXiv e-prints, arXiv-2509.

[12] Wu, H., Shi, W., & Wang, M. D. (2024). Developing a novel causal inference algorithm for personalized biomedical causal graph learning using meta machine learning. BMC Medical Informatics and Decision Making, 24(1), 137.

[13] Nilforoshan, H., Moor, M., Roohani, Y., Chen, Y., Šurina, A., Yasunaga, M., ... & Leskovec, J. (2023). Zero-shot causal learning. Advances in Neural Information Processing Systems, 36, 6862-6901.

[14] Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., ... & Pal, C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. arXiv preprint arXiv:1901.10912.

[15] Bing, S., Wahl, J., Ninad, U., & Runge, J. (2023). Invariance & causal representation learning: Prospects and limitations. arXiv preprint arXiv:2312.03580.

[16] Cao, Q. P. X., Tran, D. D., Ngo, S. T. T., Nghi, K. H., Phan, D. T., & Pham, H. V. (2025, February). A study on Few-shot Learning approach for Intrusion Detection System with Class Incremental Learning. In Proceedings of the 2025 10th International Conference on Intelligent Information Technology (pp. 113-120).

[17] Ton, J. F., Sejdinovic, D., & Fukumizu, K. (2021, May). Meta learning for causal direction. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 11, pp. 9897-9905).

[18] Alrayes, F. S., Amin, S. U., & Hakami, N. (2025). An Adaptive Framework for Intrusion Detection in IoT Security Using MAML (Model-Agnostic Meta-Learning). Sensors (Basel, Switzerland), 25(8), 2487.

[19] Xu, C., Zhan, Y., Chen, G., Wang, Z., Liu, S., & Hu, W. (2025). Elevated few-shot network intrusion detection via self-attention mechanisms and iterative refinement. PLoS One, 20(1), e0317713.

[20] Wang, Z., Hu, Y., Bühlmann, P., & Guo, Z. (2024). Causal invariance learning via efficient optimization of a nonconvex objective. arXiv preprint arXiv:2412.11850.