

Deep Learning Intrusion Detection: A Critical Reflection on the Inherent Biases of Datasets and Evaluation Misconceptions

Zihao Lei*

DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian 116024, China

**Corresponding author: Zihao Lei.*

Abstract

Deep learning-based intrusion detection is an important means of network security protection, but the existing validation and evaluation mechanisms still have many systematic problems. This paper discusses three aspects: mainstream validation methods, inherent biases in public datasets, and the failure of evaluation metrics, and explains the reasons for the “false prosperity” in this field. Models cannot truly understand attack characteristics and only remember the statistical features of old datasets. When the sample classes are imbalanced, metrics such as accuracy and F1 can easily be misleading, and the good performance of models on test sets is difficult to reflect their actual detection effect. This paper points out that slow data updating, unreasonable evaluation standards, and insufficient model robustness are the main difficulties at present, and puts forward some ideas for establishing objective evaluation standards. This review hopes to remind researchers to pay attention to the reliability of model validation and promote deep learning-based intrusion detection from laboratory testing to real network protection scenarios.

Keywords

deep learning, intrusion detection, evaluation metric, class imbalance

1. Introduction

Network intrusion detection is an important technology for cybersecurity defense and is mainly used to identify different types of malicious attacks from massive amounts of network traffic. Today, the Internet has widely covered critical infrastructure and many aspects of social life. Once intrusion detection fails, it can easily lead to data leakage, service crashes, or even public safety problems. Therefore, research in this field has strong practical value. In recent years, deep learning has been widely used in intrusion detection research, and many existing studies have shown that model detection accuracy on various public datasets can exceed 99%.

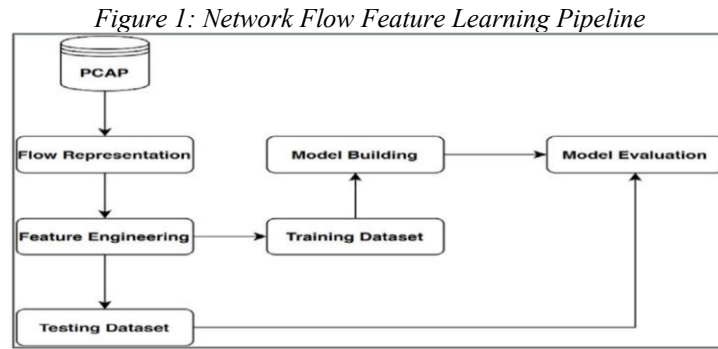
However, there is an obvious problem in current research: the extremely high detection accuracy measured in laboratory environments cannot be reproduced in real network scenarios. The main reason is that commonly used public datasets such as NSL-KDD and CIC-IDS2017 have problems such as outdated data and inherent biases. During model training, what the model learns is only the unique features of the dataset rather than the

real characteristics of attack behavior. At the same time, under imbalanced data conditions, commonly used evaluation metrics such as accuracy and F1 may be biased and cannot objectively reflect model performance, which can easily lead to an overestimation of the model's actual detection level. Based on this, this paper summarizes the mainstream validation methods for deep learning-based intrusion detection, analyzes the negative effects of dataset bias and evaluation metric failure on the reliability of research results, and provides evaluation optimization schemes combined with real network scenarios, offering a reference for standardized research in this field.

2. Typical Validation Paradigms of Deep Learning Intrusion Detection

2.1 Paradigm Based on Network Flow Statistical Features

According to differences in the representation form of input data and feature extraction methods, the validation paradigms of deep learning intrusion detection methods have formed a relatively clear classification pattern in academia. They mainly cover paradigms based on network flow statistical features, paradigms based on raw packet payloads, and paradigms based on traffic imaging [1]. The Network Flow Feature Learning Pipeline is depicted in Figure 1.



The paradigm based on network flow statistical features is the most mainstream validation method at present. This type of method takes NetFlow or session-level aggregated statistical features, such as packet count, byte count, connection duration, and flag combinations, as model inputs, and uses architectures such as CNN, LSTM, and Transformer to complete automatic feature extraction and classification. The systematic review of network intrusion detection elaborates in detail the application progress of recurrent neural networks and their variants, autoencoders, convolutional neural networks, generative adversarial networks, and cutting-edge models such as Transformer and graph convolution in this paradigm [1]. The CRGT-SA model proposed by Chen et al. jointly learns spatiotemporal features through CNN, gated TCN, and LSTM modules, and introduces a self-attention mechanism to select salient features. Its binary-classification accuracy on the UNSW-NB15 dataset reaches 91.5%, and its multiclass accuracy reaches 90.5% [2]. The MemFGAN method of Zhang et al. introduces a memory module into a generative adversarial network so that the model learns feature vectors of normal samples and stores and strengthens them in memory. Its F1 value increases by an average of 0.147 on the four datasets MAWILab, ISCX2012, IDS2017, and IDS2018 [3]. Cao et al.'s L2-AMNN model introduces a two-layer attention mechanism into a bidirectional long short-term memory neural network to capture key byte-level information and packet-level information. Its accuracy and detection rate increase by an average of 4.05% and 2.48%, respectively, compared with models such as ensemble SVM, DNN, and LSTM [6]. This paradigm has the advantages of controllable feature dimensions and relatively high computational efficiency, but its limitation is that it relies too much on statistical attributes preset in public benchmark datasets and can easily amplify the inherent biases of datasets.

2.2 Paradigm Based on Raw Packet Payloads

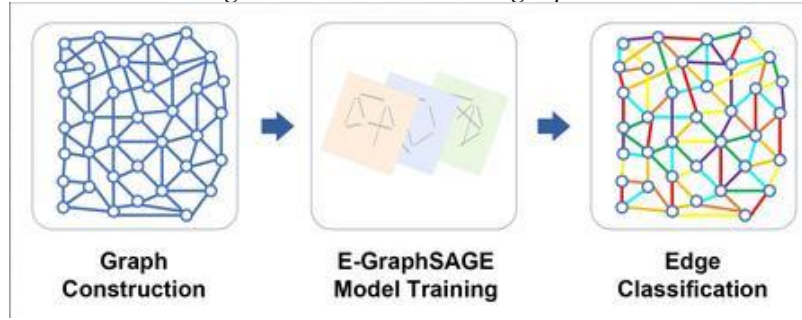
The paradigm based on raw packet payloads has gradually emerged in recent years. This type of method directly takes raw traffic byte sequences or packet payload data as input, avoids the information loss caused by manual feature engineering, and is theoretically closer to the detection requirements of real network environments. Cao et al.'s L2-AMNN model directly extracts effective payload data from raw network traffic

as samples without complex feature engineering [6]. Ding et al. innovatively proposed a method of converting network data into images and designed a deep convolutional neural network model for training and recognition [7]. However, this type of method has relatively high requirements for computing resources, and encrypted traffic still restricts its practical application.

2.3 Paradigm Based on Traffic Imaging

The paradigm based on traffic imaging converts network traffic into grayscale images or two-dimensional matrices and uses mature computer vision deep learning models for detection. Ding et al.'s method converts network data into images, uses two convolution and pooling processes to reduce the dimensions of the images, uses the ReLU function to accelerate network convergence, and uses Dropout to prevent overfitting [7]. Yang et al. proposed an intrusion detection model based on an improved convolutional neural network (ICNN), namely IBIDM, and used a cross-layer aggregation mechanism to replace the traditional “convolution-pooling-fully connected” cascading structure [8]. This form essentially converts structured data into images, which may additionally introduce unnecessary assumptions about spatial relationships.

Figure 2: Visualized Learning Pipeline



Overall, the three paradigms follow different research ideas, but they all assume one premise: public datasets can reflect the attack distribution and traffic characteristics in real networks. This assumption is exactly the root of the current “false prosperity” in related research.

3. The Problem at the Source: Analysis of the Obsolescence and Inherent Biases of Public Datasets

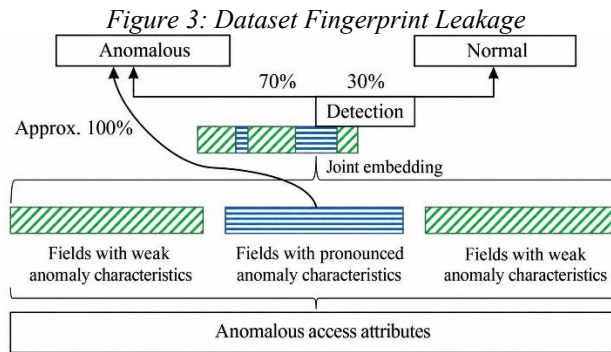
3.1 Obsolescence of Datasets

The obsolescence of datasets is the most obvious problem. KDD Cup 1999 was born in 1999, and its simulated network topology, attack tools, and traffic patterns have long been disconnected from the current network environment. Although the study of Alrawashdeh and Purdy in 2016 still obtained a detection rate of 97.9% on the KDDCUP 99 dataset, it also frankly stated that the defects of the KDDCUP 99 dataset had been widely recognized [17]. Even the relatively newer CIC-IDS2017 has difficulty reflecting the dynamic evolution characteristics of advanced persistent threats and zero-day attacks because its attack scenarios are statically preset.

3.2 The “Fingerprint Leakage” Problem of Datasets

A more hidden problem is the “fingerprint leakage” of datasets. Because there are systematic differences in statistical features between attack traffic and normal traffic during the construction of public datasets, deep learning models can achieve extremely high detection accuracy by learning these “dataset fingerprints” rather than truly understanding attack semantics. Vanlalruata Hnamte and Jamal Hussain used the DCNNBiLSTM model and achieved an accuracy of 100% on the CICIDS2018 dataset and 99.64% on the Edge_IIoT dataset [18]. However, there is an obvious problem behind these excellent experimental results: achieving the best performance on a single dataset does not mean that the model can adapt to real network scenarios. In binary classification experiments, all attack samples are uniformly classified as abnormal samples, and the model can achieve an almost perfect detection result by relying on the majority of attack samples. This situation, in which

the majority samples hide the identification defects of minority attacks, makes the high accuracy lose its reference value.



3.3 The Gap Between Static Dataset Distributions and the Dynamics of Real Networks

In addition, there is a clear gap between static datasets and the dynamic changes of real networks. Real network traffic is constantly affected by concept drift, and users' daily network behavior and attackers' attack methods change continuously with different scenarios and over time. Zhao et al. mentioned that this field currently faces many problems, including poor data quality, sample imbalance, difficulties in processing high-dimensional data, and insufficient real-time performance [1]. Kasongo also confirmed through experiments that when facing completely new attack samples, an increase in feature dimensions can lead to a clear decrease in the testing accuracy of intrusion detection models [19]. Although datasets such as CIC-IDS2017 contain sufficient amounts of data, their fixed collection, labeling, and release process is too static to effectively test the continual learning and adaptive capabilities of models.

4. Failure of the Effectiveness of Evaluation Metrics: Deceptiveness of Metrics under Imbalanced Data

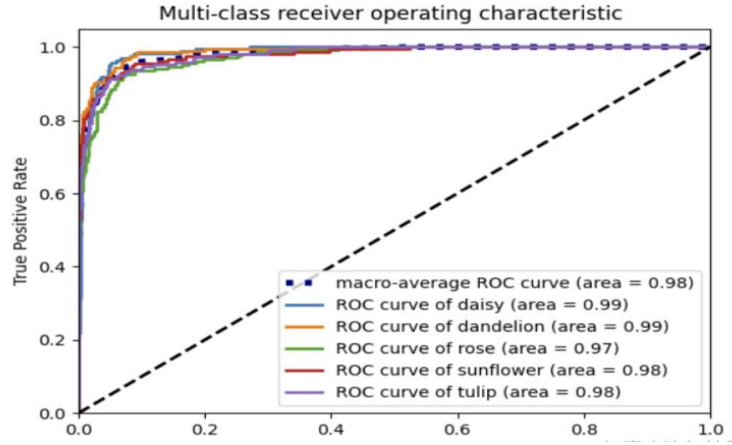
4.1 Deceptiveness of Accuracy under Imbalanced Data

The deceptiveness of accuracy under imbalanced data has long been discussed in the field of machine learning. In intrusion detection scenarios, majority attack classes have sufficient samples, while minority attack classes are extremely rare. A model only needs to perform well on the majority classes to push the overall accuracy close to 100%, while in fact it has almost no identification capability for low-frequency, high-risk attacks. The RNN-IDS model proposed by Yin et al. showed classification accuracy better than that of traditional machine learning methods on benchmark datasets [16]. In intrusion detection scenarios, majority attack classes have sufficient samples, while minority attack classes are extremely rare. A model only needs to perform well on the majority classes to push the overall accuracy close to 100%, while in fact it has almost no identification capability for low-frequency, high-risk attacks. The RNN-IDS model proposed by Yin et al. showed classification accuracy better than that of traditional machine learning methods on benchmark datasets [16]. However, this conclusion of being "better than traditional methods" is based on the evaluation system of a specific dataset, and overall accuracy conceals differences in classification performance across different attack categories.

4.2 The Balance Illusion of F1-Score

The "balance illusion" of F1-Score is also worthy of vigilance. As the harmonic mean of precision and recall, F1 does suppress the deceptiveness of a single overall accuracy metric to a certain extent. However, when datasets are extremely imbalanced, a classifier biased toward the majority class can still obtain a seemingly good F1 value. The STL-IDS method proposed by Al-Qatf et al. combines a sparse autoencoder and SVM, and obtained performance better than methods such as J48, Naive Bayes, and random forest on the NSL-KDD dataset [20]. However, this type of comparison report often lacks separate F1 reports for each attack category, making it impossible for readers to judge the model's actual performance on rare attacks such as U2R and R2L.

Figure 4: Confusion Matrix and ROC Comparison on Imbalanced Data



4.3 The Hidden Misleading Nature of ROC-AUC

The misleading nature of ROC-AUC is even more hidden. AUC measures a model's ranking ability for positive and negative samples. Under imbalanced data, because the number of negative samples is very large, a model only needs to rank most negative samples behind positive samples to obtain a relatively high AUC, even if the recall rate for minority attacks is extremely low. Taking the study of Vanlalruata Hnamte and Jamal Hussain as an example, their DCNNBiLSTM model achieved 100% accuracy on the CICIDS2018 dataset [18], and the AUC of this result must approach the theoretical upper limit. However, the paper does not separately report the model's precision and recall on each minority attack category, making it difficult for this nearly perfect overall metric to validate its actual performance on different attack categories. What is more worthy of reflection is that current studies commonly use two reporting methods, macro-averaging and micro-averaging. In multiclass settings, macro-averaging gives equal weight to each class and can better reflect the model's performance on rare classes. In contrast, micro-averaging is mainly influenced by classes with large numbers of samples and can easily hide the model's failure to detect minority attack classes. Although Zhao et al. systematically reviewed the use of different deep learning models in network intrusion detection, further critical discussion of the evaluation metrics themselves is still needed [1]. A systematic summary of deep learning intrusion detection studies in recent years shows that current studies generally tend to use overall accuracy or aggregate F1 as the main evaluation metrics, while separate performance reports for each attack category are insufficient [1].

5. Toward Robust Evaluation: Suggestions for a Validation Framework Oriented to Real Dynamic Network Environments

5.1 "De-fingerprinting" at the Dataset Level and a Dynamic Updating Mechanism

Researchers should be encouraged to use cross-validation on multiple datasets to validate the generalization of models and avoid relying on a single dataset. Kasongo's study used the two datasets NSL-KDD and UNSW-NB15 for validation. The binary-classification test accuracy of XGBoost-LSTM on NSL-KDD reached 88.13%, and that of Simple-RNN on UNSW-NB15 reached 87.07% [19]. This idea of validation on multiple datasets is worthy of promotion. At the same time, a temporal-slice validation method should be explored: earlier data are used for training and later data are used for testing, so as to simulate a model's performance in a concept-drift scenario.

5.2 Hierarchical and Fine-Grained Evaluation Metric System

In addition to overall F1 and AUC, precision, recall, and F1 should be reported by category. For rare attack categories, metrics applicable to resource-limited scenarios, such as precision at k, should be introduced. Liu et al., in response to the problem that deep learning in industrial wireless sensor networks is easily affected by adversarial attacks, constructed an adversarially enhanced model by using a generative adversarial network and increased model training strength by increasing the number of training repetitions and the attack strength

of generated samples [4]. This evaluation from an adversarial perspective provides a new idea for improving the system of evaluation metrics.

5.3 Introducing Robustness Evaluation and Adversarial Testing

Liu et al.'s study pointed out that deep learning models are vulnerable to adversarial attacks in scenarios such as industrial wireless sensor networks, and the detection performance of existing models seriously declines under adversarial perturbations [4]. The model proposed in that study improved the defense effect against various adversarial attacks compared with existing methods [4]. It is recommended that adversarial-sample testing and data-poisoning simulations be included in the standard validation process.

5.4 Standardized Reporting of Inference Efficiency and Real-Time Metrics

Most current studies focus on detection accuracy, while few systematically report model inference latency, throughput, and resource consumption. Ding et al.'s ensemble deep forest method increases the convergence speed by nearly 50% while ensuring classification accuracy [10]. The IL-SVM-KNN proposed by Fu et al. requires less time and has lower resource consumption than a two-layer convolutional neural network [9]. The NDNN model designed by Wang et al. further improves network intrusion detection accuracy on the KDD Cup 99 dataset [11]. The DCNN-DSAE proposed by Li et al. has high precision in classification and shortens the processing time for detection and classification [12]. The DBN-IDS model of Gao et al. is applicable to the classified processing of massive high-dimensional nonlinear intrusion data [14]. The algorithm of Jia and Kong based on a convolutional neural network effectively improves classification accuracy on the KDD 99 dataset [13]. Therefore, future validation frameworks should make processing capacity per unit time and the number of model parameters mandatory reporting metrics.

6. Conclusion

This paper analyzes three validation paradigms of deep learning-based intrusion detection, inherent dataset biases, and the failure of evaluation metrics, and identifies the reasons for the “false prosperity” in this field: models cannot understand attacks themselves and only remember the statistical features of static datasets. In addition, accuracy and F1 can be misleading under class imbalance, so good-looking test results cannot represent the actual detection level. The main problems in current research are slow dataset updates, failure of the evaluation system, and the lack of adversarial robustness and real-time detection capability. Future research should use cross-validation on multiple datasets, report experimental metrics for each class, establish reliable evaluation rules, build dynamic datasets, and study continual learning methods that can deal with concept drift. Only by recognizing these problems can deep learning-based intrusion detection be applied to real network protection scenarios.

References

- [1] Zhao, P., Wang, H., Liu, Y., Zhang, S., Zhao, X., & Chi, Z. (2026). Review of deep learning in network intrusion detection. *Computer Engineering and Applications*, 62(10), 74-88. <https://doi.org/10.3778/j.issn.1002-8331.2510-0117>
- [2] Chen, J., Liu, W., Qiu, X., Lv, W., & Xiong, Y. (2025). CRGT-SA: An interlaced and spatiotemporal deep learning model for network intrusion detection. *Frontiers of Information Technology & Electronic Engineering*, 26(7), 1115-1130. <https://doi.org/10.1631/FITEE.2400459>
- [3] Zhang, H., Liang, Y., Lan, J., & Zhao, Q. (2024). Intrusion detection method based on memory module and filtered generative adversarial network. *Computer Engineering*, 50(6), 197-207. <https://doi.org/10.19678/j.issn.1000-3428.0068157>
- [4] Liu, Y., & Sun, Z. (2024). IWSN intrusion detection enhancement model for adversarial defense. *Journal of Chinese Computer Systems*, 45(8), 1980-1986. <https://doi.org/10.20009/j.cnki.21-1106/TP.2023-0124>
- [5] Wang, Z., Zhang, L., & Li, D. (2021). Survey of intrusion detection systems for Internet of Things based on machine learning. *Computer Engineering and Applications*, 57(4), 18-27. <https://doi.org/10.3778/j.issn.1002-8331.2010-0457>

- [6] Cao, L., Li, Z., Yang, Y., & Zhao, L. (2021). Intrusion detection method based on two-layer attention networks. *Computer Engineering and Applications*, 57(19), 142-149. <https://doi.org/10.3778/j.issn.1002-8331.2006-0220>
- [7] Ding, H., Wan, L., Zhou, K., Long, T., & Xin, Z. (2019). Study on intrusion detection based on deep convolution neural network. *Computer Science*, 46(10), 173-179. <https://doi.org/10.11896/JSJKX.180801429>
- [8] Yang, H., & Wang, F. (2019). Network intrusion detection model based on improved convolutional neural network. *Journal of Computer Applications*, 39(9), 2604-2610. <https://doi.org/10.11772/j.issn.1001-9081.2019020327>
- [9] Fu, Z., Xu, Y., Wu, Z., Xu, D., & Xie, X. (2020). SVM-KNN network intrusion detection method based on incremental learning. *Computer Engineering*, 46(4), 115-122. <https://doi.org/10.19678/j.issn.1000-3428.0054701>
- [10] Ding, L., Wu, Z., & Su, J. (2020). Intrusion detection method based on ensemble deep forests. *Computer Engineering*, 46(3), 144-150. <https://doi.org/10.19678/j.issn.1000-3428.0053018>
- [11] Wang, M., Wang, Y., & Han, J. (2018). Application of deep learning in new intrusion detection system. *Microelectronics & Computer*, 35(7), 83-86. <https://doi.org/10.19304/j.cnki.issn1000-7180.2018.07.017>
- [12] Li, C., Wu, Y., Qian, Z., Sun, Z., & Wang, W. (2018). DDos attack detection and defense in SDN based on a hybrid deep learning model. *Journal on Communications*, 39(7), 176-187. <https://doi.org/10.11959/j.issn.1000-436x.2018128>
- [13] Jia, F., & Kong, L. (2017). Intrusion detection algorithm based on convolutional neural network. *Transactions of Beijing Institute of Technology*, 37(12), 1271-1275. <https://doi.org/10.15918/j.tbit1001-0645.2017.12.011>
- [14] Gao, N., Gao, L., & He, Y. (2016). Deep belief nets model oriented to intrusion detection system. *Systems Engineering and Electronics*, 38(9), 2201-2207. <https://doi.org/10.3969/j.issn.1001-506X.2016.09.33>
- [15] Wang, Z., Zhang, L., & Li, D. (2021). Survey of intrusion detection systems for Internet of Things based on machine learning. *Computer Engineering and Applications*, 57(4), 18-27. <https://doi.org/10.3778/j.issn.1002-8331.2010-0457>
- [16] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access*, 5, 21954-21961. <https://doi.org/10.1109/ACCESS.2017.2762418>
- [17] Alrawashdeh, K., & Purdy, C. (2016). Toward an online anomaly intrusion detection system based on deep learning. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 195-200). IEEE. <https://doi.org/10.1109/ICMLA.2016.0040>
- [18] Hnamte, V., & Hussain, J. (2023). DCNNBiLSTM: An efficient hybrid deep learning-based intrusion detection system. *Telematics and Informatics Reports*, 10, Article 100053. <https://doi.org/10.1016/j.teler.2023.100053>
- [19] Kasongo, S. M. (2023). A deep learning technique for intrusion detection system using a recurrent neural networks based framework. *Computer Communications*, 199, 113-125. <https://doi.org/10.1016/j.comcom.2022.12.010>
- [20] Al-Qatf, M., Lasheng, Y., Al-Habib, M., & Al-Sabahi, K. (2018). Deep learning approach combining sparse autoencoder with SVM for network intrusion detection. *IEEE Access*, 6, 52843-52856. <https://doi.org/10.1109/ACCESS.2018.2869577>

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).