Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-11, November 2025
DOI: 10.53469/jrse.2025.07(11).13

Lightweight Intrusion Detection Model Using Adaptive Knowledge Distillation

Shurui Yan, Xin Liu*, Fengbiao Zan*

School of Intelligence Science and Engineering, Qinghai Minzu University, Xining 810007, Qinghai, China
**Correspondence Author*

Abstract: ***To address the trade-off between detection accuracy and computational overhead in network intrusion detection systems within resource-constrained environments, this paper proposes a lightweight model based on adaptive knowledge distillation. Traditional knowledge distillation methods often exhibit low knowledge transfer efficiency and insufficient learning of hard samples when processing network traffic data. The proposed model achieves collaborative optimization of lightweight architecture and detection performance through two core mechanisms. First, a dynamic weight allocation strategy based on the Euclidean distance between teacher and student model outputs is designed to adaptively adjust the weights of soft targets and hard labels in the loss function, thereby enhancing the stability of knowledge transfer. Second, Focal Loss is introduced to strengthen the model's ability to learn hard samples, improving the recognition of complex attack patterns and rare threats. Experimental results on the NSL-KDD dataset demonstrate that the proposed method, while compressing the model parameters by nearly two orders of magnitude, still outperforms traditional knowledge distillation methods in detection performance, providing a feasible technical pathway for efficient intrusion detection in resource-constrained environments.***

Keywords: Knowledge distillation, Intrusion detection, Adaptive learning, Model compression, Lightweight deployment**.**

1. Introduction

Driven by the deepening of digital transformation and the widespread adoption of 5G technology, cyberspace security is facing unprecedented challenges. Attack methods are becoming increasingly complex and diverse, evolving from traditional port scanning and denial-of-service attacks to advanced persistent threats and zero-day exploits. In this context, network intrusion detection systems (NIDS), as a key component of the cybersecurity defense architecture, have grown in importance [1,2].

Deep learning based network intrusion detection approaches have received increasing attention from both academia and industry in recent years owing to their powerful feature learning capability and promising generalization performance. Compared with conventional rule based or classical machine learning methods, deep models can automatically extract high level representations from raw network traffic, enabling effective identification of previously unseen attack patterns and sophisticated threat behaviors [3-5]. Representative architectures include convolutional neural networks, recurrent neural networks, long short term memory networks, and autoencoders, all of which have demonstrated superior detection accuracy on several public datasets.

Nevertheless, high-performance deep-learning models are inherently coupled with substantial computational and memory costs. A seven-hidden-layer multilayer perceptron, for instance, typically entails more than 800 k trainable parameters and millions of floating-point operations during a single forward pass. Such complexity precludes real-time deployment on resource-constrained edge devices, IoT endpoints, or high-throughput network appliances. Moreover, operational environments impose stringent latency and energy budgets, further compounding the deployment challenge.

Knowledge Distillation supplies a promising solution to the aforementioned problem. First presented by Hinton et al. in 2015, the method compresses a large teacher network into a compact student by transferring the dark knowledge embedded in the teacher's output distribution, namely the inter class similarity relationships, thereby maintaining accuracy while reducing size. Conventional KD introduces a temperature parameter to soften the teacher's posterior, generating soft targets that convey rich similarity information and guide the training of the smaller model.

However, directly applying traditional knowledge distillation to network intrusion detection systems still faces several key challenges, mainly due to the mismatch between the characteristics of network traffic data and the assumptions of conventional KD mechanisms. Specifically, network traffic data are high dimensional, non stationary, and exhibit severe class imbalance, leading to a mixture of easy and hard samples, with hard samples especially those from rare attack categories being particularly difficult to learn. Second, traditional KD methods typically rely on fixed loss weight combinations, which fail to adapt to the evolving state of the student model during training, thereby affecting the stability of knowledge transfer and the final performance. Although some studies have attempted to introduce adaptive weighting or sample reweighting strategies to improve KD, these methods are mostly designed for image data and depend on strong independent and identically distributed assumptions and clear class semantic structures, making them unsuitable for NIDS environments characterized by temporal correlations, concept drift, and extreme class imbalance. Therefore, designing a distillation framework that can adapt to the unique properties of network traffic data and dynamically optimize the knowledge transfer process is of great significance for achieving high performance and lightweight NIDS. To address these challenges, this paper proposes a lightweight intrusion detection model based on adaptive knowledge distillation, aiming to enhance the efficiency of knowledge transfer and the robustness of the model. Specifically, we design a dynamic weight adjustment mechanism based on the Euclidean distance between the outputs of the teacher and student models, which adaptively balances the loss contributions of soft targets and hard label supervision in response to changes in the model state during training. Additionally, Focal Loss is introduced to increase the focus on hard samples, thereby improving the discrimination of complex attack patterns.

2. Related Work

Research on network intrusion detection has evolved from rule-based to data-driven paradigms. Early systems relied on handcrafted rule sets for pattern matching; although effective against known attacks, they generalize poorly to novel or mutated threats.

Advances in machine learning have shifted the focus to statistical learning and feature engineering, making them the mainstream paradigm [6,7]. Conventional algorithms such as support-vector machines, decision trees, and random forests have been extensively applied to intrusion detection. These techniques extract traffic statistics, protocol fields, and behavioral descriptors to train classifiers that separate legitimate from malicious traffic. Nevertheless, their performance is tightly coupled to the quality of hand-crafted features, and their capacity to model high-dimensional nonlinear data remains limited.

In recent years, deep learning has achieved remarkable progress in network intrusion detection. Convolutional neural networks have been employed to extract spatially local patterns from traffic, demonstrating particular strength when processing packet payloads [8]. Recurrent neural networks and their variants such as long short term memory and gated recurrent units effectively capture temporal dependencies, making them well suited for detecting time sensitive attacks including distributed denial of service [9]. Autoencoders perform anomaly detection through reconstruction error and offer unique advantages in unsupervised and semi supervised learning scenarios [10]. Although deep-learning models deliver superior detection accuracy, their computational complexity and memory footprint severely restrict real-world deployment. This limitation is especially acute in resource-constrained scenarios such as Internet-of-Things and edge-computing environments, where model lightweighting has become a critical imperative.

Knowledge distillation has attracted broad attention as an effective model compression technique. Conventional KD originates from the pioneering work of Hinton et al. [11], where the Kullback Leibler divergence between the teacher and student output distributions is minimized to transfer inter class similarity information, the so called dark knowledge embedded in soft labels, to the student and thus improve generalization. Subsequent studies have extended the KD framework along multiple directions and have applied it to network intrusion detection. For example, a cooperative framework combining federated learning with distillation was introduced to address data heterogeneity [12], while self knowledge distillation was employed to design the lightweight TBCLNN model [13], and a distilled BERT framework tailored for efficient IoT intrusion detection was proposed in [14]. All of these approaches maintain or even improve detection accuracy while achieving significant model compression. Nevertheless, most of these enhancements still rely on the traditional KD transfer paradigm and do not fully account for the high dimensionality, severe class imbalance, and dynamic evolution that characterize network traffic data. Consequently, when facing rare attack categories or highly variable traffic patterns, both the efficiency and the robustness of knowledge transfer are often limited. Therefore, developing adaptive knowledge distillation methods specifically designed for NIDS scenarios has become essential for advancing detection performance in this domain.

3. Model Analysis

To overcome the accuracy bottleneck of lightweight network intrusion detection systems, this paper proposes a lightweight intrusion detection model based on adaptive knowledge distillation. The core innovation is a cooperative optimization mechanism driven by dynamic feedback throughout the learning process, which introduces adaptive strategies to refine and intelligently regulate knowledge transfer. Specifically, the model continuously measures the discrepancy between teacher and student predictions and dynamically adjusts the relative weight of soft targets and hard labels in the loss function. Simultaneously, it integrates Focal Loss to emphasize difficult samples, enabling the student to learn discriminative knowledge from the high-performance teacher more efficiently and robustly, with particular gains in detecting complex and stealthy threats. The complete pipeline consists of three stages: data preprocessing, teacher model training, and student model distillation. The architecture of the knowledge distillation model is illustrated in Figure 1.

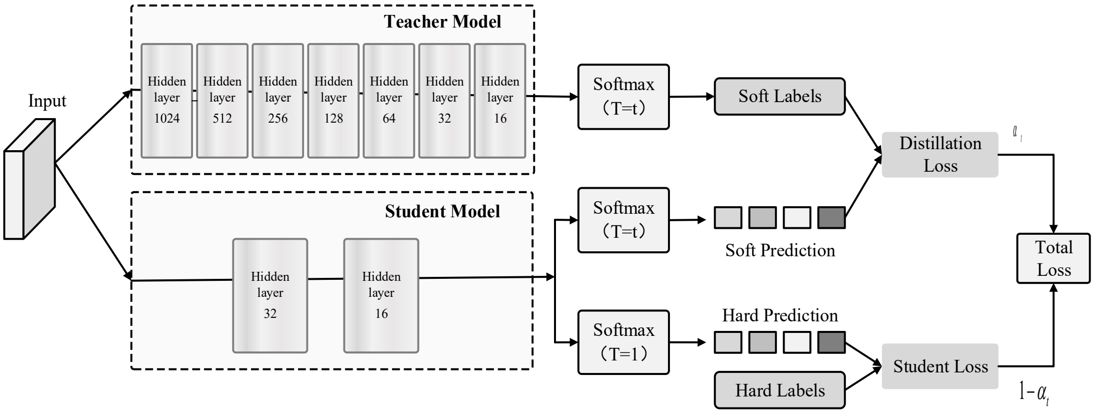

Figure 1: Knowledge distillation model structure

3.1 Data Preprocessing

To improve training efficiency and generalization, raw network traffic is systematically pre-processed. Categorical fields are converted to numeric form via one-hot encoding, while continuous variables are scaled to the interval [0, 1] using Min-Max normalization. The resulting uniform feature set is then used for both teacher and student training and evaluation.

3.2 Teacher and Student Model Design

The adaptive knowledge distillation intrusion detection framework strictly adheres to the teacher student paradigm; its efficacy hinges on a high quality knowledge source, namely the teacher, and an efficient recipient, the student.

The efficacy of distillation is governed by the quality of the teacher's knowledge. A teacher that generalizes well produces soft labels rich in inter class relations and uncertainty. We adopt a deep multilayer perceptron with seven hidden layers as the teacher. The network is trained end to end on the pre processed training set using cross entropy loss, Adam optimizer, and an initial learning rate of 1e-3 until full convergence. After training, all parameters are frozen, and a forward pass is performed on the entire training set to collect soft labels. These labels carry class probabilities together with the teacher's estimate of pairwise similarity and uncertainty, forming the primary supervisory signal for the student.

The student model is designed for extreme lightweight deployment to meet the stringent latency and power constraints of edge devices. We employ a compact two layer multilayer perceptron whose parameter count is orders of magnitude smaller than that of the teacher, enabling highly efficient inference. The objective is to approach or exceed the detection performance of conventional lightweight models while maintaining minimal footprint, thereby achieving high capability with a small architecture.

3.3 Dynamic Weighting via Teacher-Student Output Discrepancy

Traditional knowledge distillation employs a constant coefficient to balance soft label loss and hard label loss. This static allocation cannot accommodate the evolving state of training. To address this limitation we design a dynamic weight adjustment mechanism that is driven by the prediction similarity between teacher and student.

The core idea is that the relative weight of soft versus hard supervision should evolve with the agreement between the two models. When the student closely matches the teacher, it has already absorbed the dark knowledge in the soft labels, so the loss is shifted toward the hard target to consolidate learning of the true label and prevent bias inherited from the teacher. Conversely, when disagreement is large, soft label weight is increased so that the student can continue to extract knowledge from the teacher before relying on the hard signal.

Specifically, we employ the Euclidean distance D_{st} between the teacher and student output probability distributions as a measure of their prediction consistency.

$$D_{st} = \| P_t - P_s \|_2 \tag{1}$$

Here P_t and P_s denote the output probability distributions of the teacher and the student respectively. Building on this distance we define a dynamic weight function α_t that replaces the previously fixed coefficient α:

$$\alpha_t = 1 - exp(-\beta \cdot D_{st}) \tag{2}$$

In the equation, the parameter β governs the decay rate of the weight with respect to increasing distance; during training, an adaptive strategy $\beta = 1/\mathrm{median}(D_{\mathrm{st}})$ is employed to dynamically rescale the distance magnitude, thereby ensuring that α_t varies smoothly within a reasonable range.

This function is designed with the following mathematical properties:

Boundedness: since the exponential function maps to $(0,1]$, the weight is strictly confined to the desired interval $[0,1)$.

Monotonicity: the weight α_t increases monotonically as the teacher-student discrepancy D_{st} increases. This ensures that when the discrepancy is large the soft-label loss L_{soft} receives higher weight to guide the student to imitate the teacher, whereas when the student performs well and closely matches the teacher the soft weight is reduced and the relative contribution of the hard-label loss L_{hard} is increased.

The final adaptively weighted overall loss function is formulated as:

$$L_{total} = \alpha_t \cdot L_{soft} + (1 - \alpha_t) \cdot L_{hard} \tag{3}$$

3.4 Hard-Label Focal Loss Optimization

In deep learning classification tasks network traffic data exhibit severe class imbalance and contain numerous ambiguous boundary instances as well as rare attack patterns. The standard cross entropy loss assigns equal weight to every sample so optimization is dominated by easy cases and the model fails to learn discriminative features for hard examples. To increase focus on difficult samples we adopt Focal Loss as the hard label loss term in place of traditional cross entropy. Focal Loss reweights each sample dynamically through its modulation mechanism thereby intensifying supervision for low confidence instances. The loss is expressed as follows:

$$L_{\text{focal}} = L_{\text{hard}} = -\overset{\overset{i=1}{C}}{\sum} \alpha_i (1 - p_i)^{\gamma} y_i \, log(p_i) \tag{5}$$

Here C denotes the total number of classes, y_i is the ground truth label in one hot form, and p_i represents the predicted probability for class i. The parameter γ acts as a focusing factor that controls the emphasis placed on low confidence samples: a larger γ increases the weight assigned to instances whose predicted probabilities are small, i.e., the hard examples. α_i is a class specific coefficient that counteracts bias introduced by imbalance in sample counts.

To further suppress the effect of class imbalance, we define the class weight coefficient as a function of sample frequency in the form

$$\alpha_i = \frac{\mathrm{sum}n}{C \cdot n_i} \tag{1}$$

Here n_i denotes the number of samples in class i and sum_n is the total number of training samples. This design assigns lower weights to more frequent classes, thereby reducing their dominance in the loss. In the experiments $\gamma = 2$ and α_i are set dynamically according to the training set distribution.

The core mechanism of this loss lies in the $(1 - p_i)^\gamma$ term that explicitly highlights hard samples. For easy cases whose p_i is close to 1, the factor approaches 0 and their contribution is sharply reduced. For hard cases with small p_i, the weight is amplified, forcing optimization to focus on these critical instances. Through this design the proposed method enhances recognition of complex and rare attack patterns in network traffic.

3.5 Chapter Summary

This chapter presents the complete design of the lightweight intrusion detection model based on adaptive knowledge distillation. To overcome the limitations of conventional methods when handling high dimensional and dynamic network traffic, the main contribution is a cooperative optimization loop with dynamic feedback. First, a prediction consistency driven weighting strategy is proposed to adaptively balance soft and hard supervision during distillation, stabilizing knowledge transfer. Second, an improved focal loss is introduced to enable fine grained control over the training process through the loss function itself.

4. Experimental Results and Analysis

4.1 Data Preprocessing

The experiments adopt NSL-KDD as the benchmark dataset. This revised version of KDD Cup 99 mitigates excessive redundancy and train-test distributional skew, making it a standard corpus for evaluating network intrusion detection systems. It contains diverse connection records that span multiple attack categories and normal traffic. Each instance is described by forty-one mixed-type features, including protocol, service, and duration, which are numerical or categorical.

To ensure consistent and numerically stable inputs, categorical variables are converted via one-hot encoding and continuous variables are min-max normalized to remove scale differences. The official train-test split is strictly followed to prevent information leakage.

4.2 Overall Performance Comparison

To validate the proposed approach we retain the large teacher small student paradigm described in Chapter 3. The teacher is a seven layer MLP with 824850 parameters while the student is a lightweight two layer MLP containing 4466 parameters yielding a reduction of almost two orders of magnitude.

The teacher model is first evaluated and achieves an F1 score of 86.58 percent on the test set, serving as a high performance baseline for knowledge transfer. Table 1 summarizes the test set results of all compared methods.

Table 1: Performance comparison of different methods on the NSL-KDD test set (%)

Module Name	ACC	PRE	REC	F1
Student Model	80.62	94.57	69.98	80.43
Traditional KD	82.05	94.91	72.35	82.11
Traditional+DW	83.02	96.72	72.63	82.96
Traditional+Focal	83.10	96.21	73.20	83.14
Proposed Method	83.84	96.85	74.02	83.91

The experimental results of the proposed adaptive knowledge distillation model on the NSL-KDD dataset demonstrate an exceptional balance between model lightweighting and detection performance. As shown in Table 1, the model achieves an F1-score of 83.91% and a precision of 96.85% while reducing the number of parameters by nearly two orders of magnitude compared to the teacher model. Its performance significantly outperforms traditional knowledge distillation methods (with an F1 improvement of 1.8 percentage points) and independently trained student models (with an F1 improvement of 3.48 percentage points). Moreover, it attains the highest recall rate of 74.02% among all compared methods, indicating a stronger capability in detecting real attack instances. These results confirm the synergistic effectiveness of the dynamic weight adjustment mechanism (DW) and Focal Loss: the former adaptively balances soft and hard label supervision based on the discrepancy between teacher and student outputs, ensuring stable knowledge transfer; the latter enhances discrimination of rare attacks by focusing on hard examples. Thus, the proposed approach offers a practical and efficient solution for deploying NIDS in resource-constrained environments.

5. Conclusion

The adaptive knowledge distillation model proposed in this paper achieves significant results on the NSL-KDD dataset through the synergistic optimization of dynamic weight adjustment and focal loss. It reduces the model parameters by nearly two orders of magnitude while maintaining superior detection performance, offering a feasible solution for intrusion detection systems in resource-constrained environments. Building on these findings, future work will focus on adapting the framework to more complex network architectures and further enhancing the model's adaptability and continual learning capability in dynamically evolving network environments.

Acknowledgements

This work was supported by Qinghai Province "Kunlun Talents·High-level Innovation and Entrepreneurship Talents" Project (2023).

References

[1] Chou D, Jiang M. A survey on data-driven network intrusion detection[J]. ACM Computing Surveys (CSUR), 2021, 54(9): 1-36.

[2] Choubisa M, Doshi R, Khatri N, et al. A simple and robust approach of random forest for intrusion detection system in cyber security[C]//2022 International conference on IoT and blockchain technology (ICIBT). IEEE, 2022: 1-5.

[3] Chamou D, Toupas P, Ketzaki E, et al. Intrusion detection system based on network traffic using deep neural networks[C]//2019 IEEE 24th international workshop on computer aided modeling and design of communication links and networks (CAMAD). IEEE, 2019: 1-6.

[4] Duan X, Fu Y, Wang K. Network traffic anomaly detection method based on multi-scale residual classifier[J]. Computer Communications, 2023, 198: 206-216.

[5] Vinayakumar R, Alazab M, Soman K P, et al. Deep learning approach for intelligent intrusion detection system[J]. IEEE Access, 2019, 7: 41525-41550.

[6] Wu T, Fan H, Zhu H, et al. Intrusion detection system combined enhanced random forest with SMOTE algorithm[J]. EURASIP Journal on Advances in Signal Processing, 2022, 2022(1): 39.

[7] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, et al. Anomaly-based network intrusion detection: Techniques, systems and challenges[J]. Computers & Security, 2009, 28(1-2): 18-28.

[8] Wang W, Zhu M, Zeng X, et al. Malware traffic classification using convolutional neural network for representation learning[C]//2017 International conference on information networking (ICOIN). IEEE, 2017: 712-717.

[9] Yin C, Zhu Y, Fei J, et al. A deep learning approach for intrusion detection using recurrent neural networks[J]. IEEE Access, 2017, 5: 21954-21961.

[10] Liu C, Antypenko R, Sushko I, et al. Intrusion detection system after data augmentation schemes based on the VAE and CVAE[J]. IEEE Transactions on Reliability, 2022, 71(2): 1000-1010.

[11] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[12] Shen J, Yang W, Chu Z, et al. Effective intrusion detection in heterogeneous Internet-of-Things networks via ensemble knowledge distillation-based federated learning[C]//ICC 2024-IEEE International Conference on Communications. IEEE, 2024: 2034-2039.

[13] Wang Z, Zhou R, Yang S, et al. A novel lightweight IoT intrusion detection model based on self-knowledge distillation[J]. IEEE Internet of Things Journal, 2025.

[14] Cao Z, Liu X, Zhou Z, et al. KD-BERT: A Lightweight Knowledge Distillation Bidirectional Encoder Representations from Transformers for IoT Network Intrusion Detection[J]. IEEE Transactions on Industrial Informatics, 2025.