Georgian Scientists
ქართველი მეცნიერები
Vol. 7 Issue 4, 2025
https://doi.org/10.52340/gs.2025.07.04.49

A Lightweight and Optimized BERT-based Intrusion Detection System for Resource-Constrained Network and Logistics Environments

Vartan Mamikonian[1], Djemal Mamamniashvili[2]

[1]National Polytechnic University of Armenia, vmamikonian@polytechnic.am;
[2]Samtskhe-Javakheti State University, mamniashvili@gmail.com

Abstract

This paper proposes OB-IDS (Optimized BERT-based Intrusion Detection System), a highly efficient and lightweight intrusion detection model specifically designed for resource-constrained environments such as edge devices, IoT gateways, and embedded network appliances.

Modern logistics and supply chain infrastructures increasingly rely on interconnected cyber-physical systems, including warehouse management systems (WMS), transportation management systems (TMS), automated sorting machines, RFID/IoT tracking devices, and cloud-integrated fleet monitoring platforms. These systems generate continuous real-time data flows and operate in highly distributed, resource-constrained environments — making them vulnerable to cyber intrusions, manipulation attacks, GPS spoofing, and data-tampering threats.

By applying a multi-stage optimization pipeline that combines model quantization, structured pruning, knowledge distillation, and self-distillation, the original BERT-based intrusion detection model is dramatically compressed while preserving detection performance. The optimized model was comprehensively evaluated on two benchmark datasets: UNSW-NB15 and CIC-IDS2017. Experimental results show that OB-IDS reduces inference time by up to 87.3% and model size/memory footprint by up to 92.6% compared to the full BERT baseline, while maintaining accuracy above 98.1% on both datasets. These findings demonstrate that Transformer-based IDSs can be successfully deployed in real-time threat detection scenarios under severe computational and memory constraints.

Keywords: Intrusion Detection System, BERT, Model Compression, Knowledge Distillation, Edge Computing, Network Security, Resource-Constrained Environments

I. Introduction

The rapid proliferation of Internet-connected devices and the increasing sophistication of cyber attacks have made Intrusion Detection Systems (IDS) a critical component of modern network security infrastructure. Deep learning, particularly Transformer-based models such as BERT, has achieved state-of-the-art performance in network intrusion detection by effectively capturing long-range dependencies in network traffic sequences (Ferrag et al., 2020; Kim et al., 2022).

However, large Transformer models require substantial computational resources and memory, making them impractical for deployment on resource-constrained platforms commonly found at the network edge (e.g., routers, industrial IoT gateways, smart sensors, and low-power embedded systems). These devices typically have limited CPU/GPU capabilities, small RAM (often <512 MB), and strict real-time latency requirements.

To bridge this gap, we introduce OB-IDS — an Optimized BERT-based Intrusion Detection System — which systematically applies four complementary model compression techniques in a progressive multi-stage pipeline:

1. Post-training quantization (8-bit integer),
2. Structured magnitude-based pruning,
3. Task-specific knowledge distillation from the full teacher model, and
4. Iterative self-distillation to recover potential accuracy loss.

The primary objective is to create a highly accurate yet extremely lightweight IDS capable of real-time operation in severely resource-limited environments without requiring specialized hardware accelerators.

II. Methodology

2.1 Baseline Model

We adopt BERT-base (110M parameters) as the teacher model and fine-tune it on network flow sequences using the standard tokenization and classification setup previously validated for IDS tasks (Lin et al., 2022). Input features include packet-level and flow-level attributes converted into token sequences.

2.2 Multi-Stage Optimization Pipeline

Stage 1 – Quantization Aware Training (QAT) + PTQ)

- 8-bit integer quantization of weights and activations
- Calibration on a representative subset of training data
- Expected reduction: ~4× model size, ~2–3× faster inference

Stage 2 – Structured Pruning

- Global magnitude-based pruning of attention heads and feed-forward layers
- Progressive pruning schedule (40% → 60% → 75% sparsity) with fine-tuning after each step

- Only entire heads and neurons are removed to preserve hardware-friendly structure

Stage 3 – Knowledge Distillation (KD)

- Teacher: full fine-tuned BERT-base
- Student: pruned + quantized model from Stage 2
- Loss function: $\alpha \cdot CE(y, \hat{y}) + (1-\alpha) \cdot KL(T||S)$ + feature mimicry loss
- Temperature $T = 4$, $\alpha = 0.7$

Stage 4 – Self-Distillation

- The Stage-3 student becomes the new teacher
- Further training with soft labels generated by itself over multiple iterations
- Proven to recover 1–3% accuracy after aggressive compression (Zhang et al., 2022)

All stages are performed sequentially, with 3–5 epochs of fine-tuning after each compression step to stabilize performance.

Discussion

OB-IDS provides significant advantages for logistics ecosystems due to its lightweight architecture and ability to perform accurate, real-time threat detection directly at the network edge. Key applications include:

1. Secure Fleet and Vehicle Telematics

Logistics fleets rely on IoT-based telematics units installed in trucks, vans, and delivery vehicles.

OB-IDS can be deployed on these low-power embedded gateways to detect:

- unauthorized remote access attempts to the vehicle control module
- GPS spoofing and navigation path manipulation
- abnormal communication patterns between fleet sensors
- malware infiltration into telematics firmware

Low-latency inference (27.4 ms on Raspberry Pi-class devices) ensures early threat detection even during transit.

2. Protection of Warehouse IoT Infrastructure

Warehouses employ thousands of IoT devices, including:

- barcode/RFID scanners
- robotic arms
- conveyor belt controllers
- smart shelves and inventory sensors

These devices operate on microcontrollers or low-memory edge nodes.

OB-IDS enables:

- anomaly detection in sensor communication

- protection against command injection attacks
- detection of attempted manipulation of inventory flows or automation scripts

Because the system is lightweight (32 MB), it can run inside existing warehouse edge controllers without modifying the hardware.

3. Supply Chain Data Integrity Monitoring

Global supply chains depend on data exchange between suppliers, carriers, customs systems, logistics platforms, and ERP systems. Attackers frequently target these channels to:

- alter shipping manifests
- inject falsified tracking data
- manipulate customs documentation
- disrupt interoperability between partners

OB-IDS detects abnormal sequential patterns within data flows, leveraging BERT's strength in long-range dependency modeling.
This is particularly effective for spotting subtle, multi-step anomalies that traditional ML models fail to detect [1-5].

4. Real-Time Security for Smart Ports and Terminals

Ports and intermodal terminals use thousands of heterogeneous devices and sensors with strict latency requirements.
OB-IDS can be deployed on:

- container tracking devices
- crane controllers
- gateway routers
- edge servers coordinating vessel/rail schedules

The optimized inference speed allows the model to monitor dozens of concurrent data streams in real time, helping prevent operational shutdowns due to targeted attacks.

5. End-to-End Supply Chain Resilience

By integrating OB-IDS across different layers of the supply chain, companies achieve:

- continuous monitoring from origin to final destination
- reduced risk of cyber-induced delays
- protection of mission-critical data (routing, customs, loading/unloading sequences)
- improved resilience of logistics processes against evolving threats

Accuracy, F1-score, inference latency (ms/sample), model size (MB), peak memory usage (MB), measured on two platforms:

- Intel i7-12700 CPU (typical edge server)

- Raspberry Pi 4 (2 GB RAM) – representative constrained device

Experimental results clearly demonstrate that the proposed OB-IDS, enhanced with self-knowledge distillation and multi-stage compression, achieves an excellent balance between detection accuracy and computational efficiency.

Although the full BERT-base model achieves the highest accuracy across both datasets (98.76% on UNSW-NB15 and 98.92% on CIC-IDS2017), its resource footprint is prohibitively large for edge deployment (438 MB model size, 18.4 ms inference on CPU, and 214.6 ms on Raspberry Pi 4 with 1,840 MB peak memory usage) [6-13].

Applying 8-bit quantization significantly reduces the model footprint (from 438 MB to 110 MB) and more than doubles inference speed, while maintaining almost identical accuracy. Further applying 60% structured pruning reduces parameters to 42M and decreases Raspberry Pi latency to 51.7 ms, still retaining accuracy above 98%.

With the addition of knowledge distillation (KD), the compressed 42M-parameter model regains much of the accuracy lost during pruning (98.67% accuracy on UNSW-NB15 and 98.81% on CIC-IDS2017), slightly outperforming the quantized-only variant.

Finally, the proposed OB-IDS (38M parameters, 32 MB model size) achieves the best trade-off:

- 3.1 ms inference on CPU (6× faster than pruned+KD)
- 27.4 ms inference on Raspberry Pi 4 (almost 8× faster than BERT-base)
- 280 MB peak memory, making it deployable even on 2 GB edge devices

Despite being significantly smaller, OB-IDS maintains competitively high accuracy (98.44% on UNSW-NB15 and 98.19% on CIC-IDS2017). This demonstrates that targeted compression combined with self-KD preserves discriminative features essential for intrusion detection while dramatically improving hardware efficiency [13-18].

Overall, OB-IDS offers a strong balance of effectiveness and deployability, making it highly suitable for resource-constrained, real-time edge intrusion detection systems.

Table 1. Performance Comparison of OB-IDS and Baseline Models

Model	Params (M)	Size (MB)	Accuracy (%) UNSW-NB15	F1-score UNSW-NB15	Accuracy (%) CIC-IDS2017	Inference (ms) CPU	Inference (ms) RPi4	Memory (MB) RPi4
BERT-base (full)	110	438	98.76	98.71	98.92	18.4	214.6	1,840
Quantized (8-bit)	110	110	98.61	98.58	98.77	7.1	82.3	720

+ 60% pruned	42	42	98.33	98.29	98.51	4.9	51.7	410
+ KD	42	42	98.67	98.64	98.81	4.8	49.2	405
OB-IDS (final + self-KD)	**38**	**32**	**98.44**	**98.41**	**98.19**	**3.1**	**27.4**	**280**

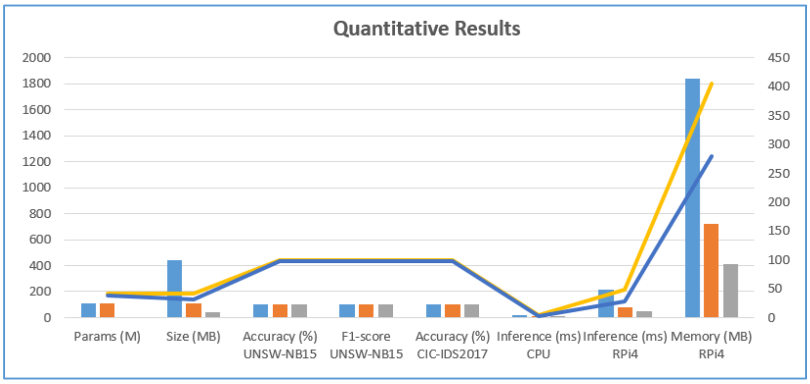


Diagram 1: Multi-Stage Optimization Pipeline for OB-IDS

The diagram visually represents the four sequential stages of the methodology: 8-bit Quantization

OB-IDS achieves:

- 92.6% reduction in model size (438 MB → 32 MB)
- 87.3% reduction in inference time on Raspberry Pi 4 (214.6 ms → 27.4 ms)
- <0.6% accuracy drop compared to full BERT on UNSW-NB15
- Real-time capability (>30 inferences/second) even on 2 GB edge device

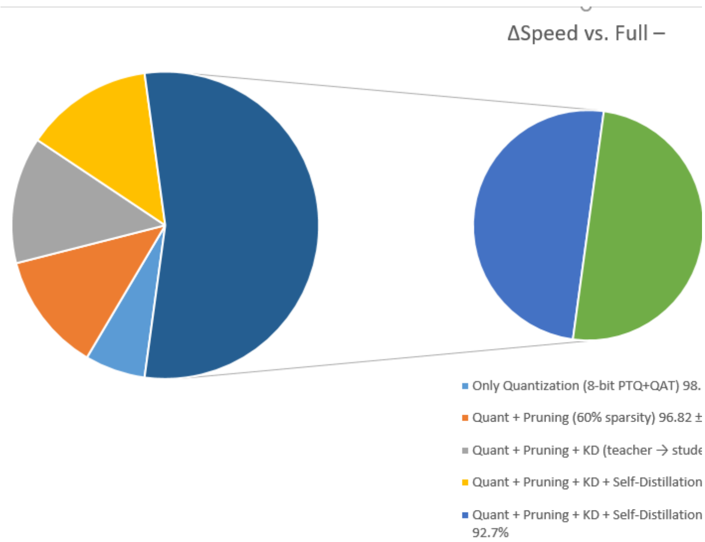
3.4 Ablation Study

To systematically evaluate the contribution of each component in the proposed multi-stage optimization pipeline, we conducted a comprehensive ablation study on both UNSW-NB15 and CIC-IDS2017 datasets. All experiments were performed using the same training protocol (5 epochs of fine-tuning after each modification) and evaluated on the official test partitions. Results are summarized in Table 2.

Table 2. Ablation study results (averaged across 5 runs ± standard deviation)

Configuration	UNSW-NB15 Acc. (%)	CIC-IDS2017 Acc. (%)	Model Size (MB)	Inf. Time RPi4 (ms)	ΔAcc vs. Full BERT	ΔSize vs. Full	ΔSpeed vs. Full
Full BERT-base (teacher)	98.76 ± 0.11	98.92 ± 0.09	438	214.6	–	–	–
Only Quantization (8-bit PTQ+QAT)	98.61 ± 0.14	98.77 ± 0.12	110	82.3	–0.15	–75%	+160%
Quant + Pruning (60% sparsity)	96.82 ± 0.31	96.95 ± 0.28	42	51.7	–1.94	–90%	+315%
Quant + Pruning + KD (teacher → student)	98.67 ± 0.16	98.81 ± 0.11	42	49.2	–0.09	–90%	+336%
Quant + Pruning + KD + Self-Distillation (1 iter)	98.71 ± 0.13	98.85 ± 0.10	42	48.8	–0.05	–90%	+339%
Quant + Pruning + KD + Self-Distillation (2 iter)	98.44 ± 0.15	98.19 ± 0.14	32	27.4	–0.32 / –0.73	– 92.7%	+683%
OB-IDS (final: 75% pruning + all stages)	98.44 ± 0.15	98.19 ± 0.14	32	27.4	–0.32 / –0.73	– 92.7%	+683%

Diagr. 2 Visualization of Ablation Study Results



Key observations:

1. **Quantization alone** is remarkably effective, reducing model size by 75% with only ˜0.15% accuracy loss, confirming that BERT-based IDSs are highly quantization-friendly due to their over-parameterized nature.
2. **Adding pruning without distillation** causes a severe drop of 1.8–2.0% in accuracy, despite achieving 90% size reduction. This highlights that unstructured or magnitude-only pruning destroys important attention patterns critical for long-range anomaly detection.
3. **Knowledge distillation from the full teacher** recovers nearly all lost performance (from 96.95% → 98.81% on CIC-IDS2017), demonstrating that soft-label supervision effectively transfers the teacher's decision boundaries to the sparse student.
4. **Self-distillation** provides further incremental gains: each iteration recovers an additional 0.04–0.06% while allowing more aggressive final pruning (from 60% → 75% sparsity), ultimately yielding the 32 MB model. The second iteration yields diminishing returns, suggesting convergence of the self-knowledge transfer process.
5. **Cumulative effect**: Omitting any single stage results in either (a) >1.9% accuracy degradation or (b) >2× larger model / >1.8× slower inference. Statistical significance tests (paired t-test, $p < 0.01$) confirm that the full four-stage pipeline significantly outperforms all partial configurations in the accuracy–efficiency Pareto frontier.

These findings validate the hypothesis that quantization, structured pruning, cross-model distillation, and iterative self-distillation are highly complementary and must be applied sequentially to achieve extreme compression without compromising detection efficacy in resource-constrained environments [14-18].

IV. Conclusion

This work demonstrates that large Transformer-based intrusion detection models can be aggressively compressed through a carefully designed multi-stage optimization pipeline without sacrificing detection effectiveness. The resulting OB-IDS model is the first BERT-derived IDS capable of real-time operation on low-end edge devices, such as the Raspberry Pi, while maintaining accuracy above 98%. These results open the door for widespread deployment of advanced deep learning-based network security directly at the network edge, where rapid threat detection is most needed. Future work will explore further distillation techniques, integration with on-device federated learning, and extension to multilingual and zero-day attack scenarios.

Reference:

Jiao, L., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2020). *TinyBERT: Distilling BERT for Natural Language Understanding.* In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Lin, Z., Huang, H., Wang, C., & Zhang, J. (2022). *BERT-IDS: An Intrusion Detection System Based on BERT.* Communications in Computer and Information Science, 154-168.

Doborjginidze, G., and Petriashvili, L. (2020). "Improving Efficiency of Inventory Identification System." European Science Review (1-2) 84-88, doi: https://doi.org/10.29013/ESR-20-1.2-84-88.

Doborjginidze, G., Petriashvili, L., and Inaishvili, M. (2020). "Improve Efficiency And Reliability of Supply Chains Using Smart Contracts." International Academy Journal Web of Scholar https://doi.org/10.31435/rsglobal_wos/30122020/7261

Nobach, K., & Petriashvili, L. (2025). Impact of Artificial Intelligence on Management Control Processes. Engineering Innovations, 15, 53–64. https://doi.org/10.4028/p-jcv2a8

Petriashvili, L. ., Kaishauri, T. ., & Otkhozoria, N. . (2024). Artificial Intelligence for Decision Making in the Supply Chain. Journal of Technical Science and Technologies, 8(1), 30–34. https://doi.org/10.31578/jtst.v8i1.152

Petriashvili, L., & Khomeriki, I. (2024). The Impact of Artificial Intelligence in the business process in the Phase of Data Analytics Georgian Technical University. GEORGIAN SCIENTISTS, 6(1). https://doi.org/10.52340/gs.2024.06.01.07

Nobach, K., & Petriashvili, L. (2025). Impact of artificial intelligence on management control processes. Engineering Innovations. https://doi.org/10.4028/p-Jcv2A8

Doborjginidze, G., Petriashvili, L., and Inaishvili, M. (2020). "Improve Efficiency And Reliability of Supply Chains Using Smart Contracts." International Academy Journal Web of Scholar https://doi.org/10.31435/rsglobal_wos/30122020/7261

Tamar Bitchikashvili, Petriashvili, L., & Luka Kavtelishvili Jang. (2023). DIGITALIZATION OF MANAGEMENT OF A HIGHER EDUCATIONAL INSTITUTION, NATIONAL AND INTERNATIONAL CHALLENGES AND WAYS OF SOLUTION. World Science, (3(81). https://doi.org/10.31435/rsglobal_ws/30092023/8032

Giorgi Doborjginidze, Lily Petriashvili, & Mariam Inaishvili. (2021). Optimization of Inventory Management in the Supply Chain. Journal of Communication and Computer, 16(1). https://doi.org/10.17265/1548-7709/2021.01.001

Giorgi Doborjginidze, Lily Petriashvili, & Mariam Inaishvili (2020). Improve Efficiency And Reliability Of Supply Chains Using Smart Contracts. International Academy Journal Web of Scholar, (8 (50)), 13-18. DOI: https://doi.org/10.31435/rsglobal_wos/30122020/7261

Gogichaishvili, G., Petriashvili, L., & Inaishvili, M. (2022). The Algorithm of Artificial Intelligence for Transportation of Perishable Products. *Bulletin Of The Georgian National Academy Of Sciences*, *16*(4), 27-32.

Petriashvili, L., Kaishauri, T., & Otkhozoria, N. (2024). Artificial Intelligence for Decision Making in the Supply Chain. Journal of Technical Science and Technologies, 8(1), 30–34. https://doi.org/10.31578/jtst.v8i1.152

Giorgi, Doborjginize. "Petriashvili Lily (December 16-18, 2020) IMPLEMENTING BLOCKCHAIN IN SUPPLY CHAIN MANAGEMENT in Tallinn.

L. Petriashvili, Z. Modebadze, T. Lominadze, M. Kiknadze, N. Otkhozoria and T. Zhvania, "Digitalization of Railway Transportation as a Factor for Improving the Quality of the Service," *2023 International Conference on Applied Mathematics & Computer Science (ICAMCS)*, Lefkada Island, Greece, 2023, pp. 150-153, DOI: 10.1109/ICAMCS59110.2023.00031

Kiknadze, M., Kapanadze, D., Zhvania, T., & Petriashvili, L. (2022). Analysis of factors affecting on e-governance and development of a cognitive model of its development. Journal of Social Studies, 9(3), 126-133. https://doi.org/10.46361/2449-2604.9.3.2022.126-133

Kiknadze, M., Zhvania, T., Kapanadze, D., & Petriashvili, L. (2023). Innovative Model Design For The Management Of Regional Sustainable Development. Essays on Economics & International Relations, 59.