Lightweight Deep Learning Models for Edge-Based Cyber Threat Detection

[1] Noman Mazher, [2] Zunaira Rafaqat

[1] University of Gujrat, Pakistan

[2] Chenab Institute of Information Technology, Pakistan

Corresponding Email: nauman.mazhar@uog.edu.pk

Abstract:

The rapid proliferation of Internet of Things (IoT) devices and distributed networks has shifted the cybersecurity landscape toward the edge, where real-time threat detection is essential. However, deploying deep learning models on edge devices presents challenges due to limited computational power, memory constraints, and energy efficiency requirements. This paper explores the development and implementation of *lightweight deep learning models* for *edge-based cyber threat detection*, emphasizing architectural optimizations, compression techniques, and adaptive intelligence. By leveraging model pruning, quantization, and knowledge distillation, researchers have enabled high-performance models that operate effectively in resource-constrained environments. Furthermore, integrating these optimized models into edge networks facilitates faster detection of malicious activities, reduces data transmission to centralized servers, and enhances privacy by processing sensitive information locally. The paper also examines challenges such as adversarial robustness, model update synchronization, and data heterogeneity across edge nodes. Ultimately, lightweight deep learning represents a pivotal advancement in creating scalable, privacy-preserving, and energy-efficient cyber defense systems for next-generation edge infrastructures.

Keywords: Edge Computing, Cyber Threat Detection, Lightweight Deep Learning, IoT Security, Model Compression, Quantization, Pruning, Knowledge Distillation, Federated Edge AI

I. Introduction

As the digital ecosystem becomes increasingly interconnected, the growth of Internet of Things (IoT) and edge devices has redefined the cybersecurity paradigm. From smart homes and autonomous vehicles to industrial control systems, billions of devices generate and process data

at the network edge [1]. This decentralization, while enhancing efficiency and reducing latency, also exposes new vulnerabilities [2]. Traditional, cloud-centric cybersecurity solutions often fail to provide timely responses to evolving threats due to communication delays and bandwidth limitations. To address these challenges, *edge-based cyber threat detection* powered by *lightweight deep learning* has emerged as a transformative solution.

Edge-based security systems operate closer to the data source, allowing faster anomaly detection and minimizing dependency on centralized infrastructures [3]. However, deep learning models—though effective at identifying complex cyber threats—typically require high computational resources, making them unsuitable for direct deployment on constrained edge devices. Standard models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) demand significant memory and energy, creating trade-offs between performance and deployability. This limitation has driven the need for lightweight architectures that retain high detection accuracy while minimizing computational overhead. Recent advancements in *model optimization* techniques such as *pruning, quantization,* and *knowledge distillation* have made it possible to design compact yet effective deep learning frameworks [4]. Pruning removes redundant neurons or layers, quantization reduces numerical precision, and knowledge distillation transfers knowledge from a large "teacher" model to a smaller "student" model. These techniques collectively enable efficient inference on edge hardware, ensuring real-time performance without compromising accuracy. Additionally, specialized architectures like *MobileNet*, *SqueezeNet,* and *TinyML models* are increasingly adopted for threat detection in low-power environments.

The benefits of deploying lightweight deep learning models at the edge extend beyond computational efficiency. Edge-based analysis ensures *data privacy* by keeping sensitive information local, mitigating the risks associated with transmitting raw data to central servers. Furthermore, distributed edge nodes can collaboratively learn and share insights, forming a decentralized intelligence layer that strengthens overall network defense. Nevertheless, designing lightweight deep learning models for cybersecurity introduces new challenges. Edge devices are heterogeneous in capability, making it difficult to standardize model deployment. The continuous evolution of cyber threats requires frequent model updates, which may strain bandwidth and storage. Moreover, adversaries can exploit lightweight models through adversarial attacks that target reduced complexity and robustness.

This paper explores the architecture, optimization, and deployment strategies for lightweight deep learning in edge-based cyber threat detection. It emphasizes the role of edge AI in building autonomous, adaptive, and scalable defense systems that align with the growing demands of real-time security.

II. Architectural Optimizations for Lightweight Threat Detection

Designing lightweight deep learning architectures for edge-based cyber threat detection involves optimizing both model structure and computation. The objective is to minimize memory and energy usage while maintaining robust detection accuracy. Several architectural innovations have enabled this balance, leading to practical deployment in resource-constrained edge environments. One of the most effective approaches is the use of *depthwise separable convolutions*, as introduced in MobileNet. This technique decomposes standard convolutions into two separate operations—depthwise and pointwise convolutions—drastically reducing the number of parameters and computations required. Similarly, *SqueezeNet* achieves comparable accuracy to AlexNet with 50× fewer parameters by using "fire modules" that compress and expand data channels strategically. These architectures have proven highly efficient for tasks such as anomaly and intrusion detection at the edge.

Pruning further contributes to model efficiency by eliminating weights or neurons that have minimal impact on predictions. Structured pruning, for example, removes entire filters or layers, simplifying computation while preserving interpretability. Combined with *quantization*, which represents model parameters with lower-bit precision (e.g., 8-bit or 4-bit), it reduces both memory footprint and inference time. Together, these methods allow deployment on low-power devices such as Raspberry Pi, Jetson Nano, and embedded IoT gateways [5].

Knowledge distillation has also become a cornerstone of lightweight deep learning. In this method, a smaller "student" model learns to mimic the behavior of a large, complex "teacher" model. This process transfers generalized knowledge while compressing the model, allowing the student network to achieve near-teacher performance with far fewer parameters. In cybersecurity applications, distilled models have been applied to detect network intrusions, phishing URLs, and malware behaviors effectively.

Beyond architectural simplifications, *edge-oriented accelerators* like Tensor Processing Units (TPUs) and Neural Processing Units (NPUs) enhance local inference efficiency. These hardware

solutions complement lightweight models by optimizing tensor operations for energy efficiency and speed. When combined with on-device learning capabilities, such setups enable real-time adaptation to evolving threats [6].

However, lightweight models face trade-offs in robustness. The simplification of network structure can make them more susceptible to adversarial perturbations and misclassifications. Hence, *adversarial training*—exposing models to manipulated examples during training—has become crucial for maintaining resilience. Additionally, hybrid architectures that combine lightweight deep learning with rule-based anomaly detection can provide layered protection, compensating for model limitations. Overall, architectural optimization serves as the foundation for efficient edge-based threat detection, balancing performance, resource efficiency, and robustness.

III. Deployment Challenges and Future Prospects in Edge Cyber Defense

While lightweight deep learning models offer promising advantages for edge-based cyber defense, their practical deployment introduces several challenges related to data distribution, model management, scalability, and trust. The edge environment is inherently decentralized, with devices differing in computational power, network connectivity, and security policies. This heterogeneity complicates model synchronization and consistent threat detection across the network [7].

A critical challenge lies in *data heterogeneity*. Edge devices observe localized data patterns influenced by their specific contexts, leading to non-identical distributions. Consequently, a model trained on one edge node may not generalize well to others [8]. To mitigate this, *federated learning* frameworks allow edge nodes to train locally and share model updates rather than raw data, facilitating collaborative learning while maintaining data privacy. Federated approaches ensure scalability and adaptability in multi-edge scenarios, fostering a unified but decentralized defense network. *Model update and version management* also pose difficulties. Cyber threats evolve rapidly, necessitating frequent model retraining. However, transmitting new models to thousands of distributed nodes is resource-intensive. Techniques such as incremental learning and compressed model updates help minimize communication overhead. Furthermore, adaptive learning mechanisms enable models to evolve autonomously by retraining on new data without human intervention [9].

Security of the models themselves is another major concern. Lightweight deep learning models

can become targets of *model poisoning*, *inference attacks*, and *adversarial perturbations*. Attackers may attempt to manipulate input data or inject malicious gradients during updates. To counter these risks, defense strategies such as *secure aggregation*, *differential privacy*, and *robust model validation* are employed. Blockchain technology has also been proposed for ensuring model integrity and transparent update verification across distributed edge nodes. From an operational standpoint, maintaining *energy efficiency* while supporting high-performance inference remains a delicate balance [10]. Edge devices must perform continuous monitoring and classification with limited battery life or thermal capacity. Emerging low-power AI chips, neuromorphic processors, and adaptive scheduling algorithms are paving the way toward sustainable, always-on threat detection systems.

Looking ahead, the integration of *self-learning edge AI* and *autonomous coordination* among edge nodes will define the future of cyber defense [11]. Through cross-node collaboration, these systems will collectively learn from new attack patterns, forming a distributed intelligence layer that mimics biological immune systems [12]. Moreover, *explainable AI (XAI)* approaches will enhance trust and interpretability, allowing analysts to understand model decisions and refine defense strategies in real time. In summary, lightweight deep learning models are not only technological innovations but also strategic enablers for decentralized cybersecurity. They empower edge devices to act as intelligent sentinels, detecting and responding to cyber threats locally while contributing to a global, adaptive defense ecosystem.

IV. Conclusion:

Lightweight deep learning models have become vital to advancing edge-based cyber threat detection, bridging the gap between high-performance AI and resource-constrained environments. Through innovations in model compression, pruning, and federated learning, edge devices can now detect threats efficiently and autonomously while preserving privacy and reducing latency. Although challenges remain in ensuring robustness, scalability, and energy efficiency, ongoing research in adaptive intelligence and privacy-preserving collaboration continues to strengthen the viability of edge-based cybersecurity. The future of digital defense

will increasingly rely on these compact, intelligent, and self-evolving models that bring AI-driven protection to the network's very edge.

REFERENCES:

[1] I. Ikram and Z. Huma, "An Explainable AI Approach to Intrusion Detection Using Interpretable Machine Learning Models," *Euro Vantage journals of Artificial intelligence,* vol. 1, no. 2, pp. 57-66, 2024.

[2] R. V. Rayala, C. R. Borra, P. K. Pareek, and S. Cheekati, "Enhancing Cybersecurity in Modern Networks: A Low-Complexity NIDS Framework using Lightweight SRNN Model Tuned with Coot and Lion Swarm Algorithms," in *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 2024: IEEE, pp. 1-8.

[3] H. Allam, J. Dempere, V. Akre, D. Parakash, N. Mazher, and J. Ahamed, "Artificial intelligence in education: an argument of Chat-GPT use in education," in *2023 9th International Conference on Information Technology Trends (ITT)*, 2023: IEEE, pp. 151-156.

[4] R. V. Rayala, C. R. Borra, P. K. Pareek, and S. Cheekati, "Fortifying Smart City IoT Networks: A Deep Learning-Based Attack Detection Framework with Optimized Feature Selection Using MGS-ROA," in *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 2024: IEEE, pp. 1-8.

[5] R. V. Rayala, C. R. Borra, P. K. Pareek, and S. Cheekati, "Hybrid Optimized Intrusion Detection System Using Auto-Encoder and Extreme Learning Machine for Enhanced Network Security," in *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 2024: IEEE, pp. 1-7.

[6] A. Mustafa and Z. Huma, "AI and Deep Learning in Cybersecurity: Efficacy, Challenges, and Future Prospects," *Euro Vantage journals of Artificial intelligence,* vol. 1, no. 1, pp. 8-15, 2024.

[7] A. Siddique, A. Jan, F. Majeed, A. I. Qahmash, N. N. Quadri, and M. O. A. Wahab, "Predicting academic performance using an efficient model based on fusion of classifiers," *Applied Sciences,* vol. 11, no. 24, p. 11845, 2021.

[8] R. V. Rayala, C. R. Borra, P. K. Pareek, and S. Cheekati, "Mitigating Cyber Threats in WSNs: An Enhanced DBN-Based Approach with Data Balancing via SMOTE-Tomek and Sparrow Search Optimization," in *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 2024: IEEE, pp. 1-8.

[9] F. Majeed, U. Shafique, M. Safran, S. Alfarhood, and I. Ashraf, "Detection of drowsiness among drivers using novel deep convolutional neural network model," *Sensors,* vol. 23, no. 21, p. 8741, 2023.

[10] B. Namatherdhala, N. Mazher, and G. K. Sriram, "Uses of artificial intelligence in autonomous driving and V2X communication," *International Research Journal of Modernization in Engineering Technology and Science,* vol. 4, no. 7, pp. 1932-1936, 2022.

[11] R. V. Rayala, C. R. Borra, P. K. Pareek, and S. Cheekati, "Securing IoT Environments from Botnets: An Advanced Intrusion Detection Framework Using TJO-Based Feature Selection an d Tree Growth Algorithm-Enhanced LSTM," in *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 2024: IEEE, pp. 1-8.

[12] M. A. Hassan, U. Habiba, F. Majeed, and M. Shoaib, "Adaptive gamification in e-learning based on students' learning styles," *Interactive Learning Environments,* vol. 29, no. 4, pp. 545-565, 2021.