SecureDeBERTa-CNN: A Hybrid IDS for Binary Threat Classification

1st Muhammad Sanaullah
Department of Electrical Engineering
University of Ulsan
Ulsan, South Korea
https://orcid.org/0009-0003-3376-5616

2nd Dong-Seop Jin*
Department of Electrical Engineering
University of Ulsan
Ulsan, South Korea
https://orcid.org/0009-0003-8710-8400

***Abstract*—The use of Internet of Things (IoT) devices has expanded rapidly across numerous sectors, yet their underlying network infrastructures remain ingenuous to ever increasing network vulnerabilities that pose serious risks to overall system safety. In this context, Intrusion Detection Systems (IDS) serve as a vital defense mechanism, enabling the detection and identification of malicious activities within IoT networks. In this work, we present SecureDeBERTa-CNN, a hybrid IDS framework that fuses transformer-based contextual representation learning with convolutional feature extraction to achieve high-precision binary threat classification in security sensitive systems. The architecture integrates a fine-tuned SecureDeBERTa transformer with a lightweight convolutional neural network (CNN), enabling an effective fusion of contextual semantics and spatial feature hierarchies from structured network telemetry. To address data imbalance and noise, the system incorporates SMOTE-ENN resampling, while training stability is ensured through progressive unfreezing and regularization. Evaluated on the UNSW-NB15 dataset, SecureDeBERTa-CNN achieves 92.1% Accuracy, 96.9% Precision, a 92.5% F1 score and a ROC-AUC of 0.9788, outperforming classical baselines Machine Learning (ML) and Deep Learning (DL) models. The results confirm its robustness, adaptability, and suitability for real-world intrusion detection in modern networked infrastructures.**

Keywords: IoT Devices, Intrusion Detection, UNSW-NB15 Dataset, SMOTE-Oversampling, ENN, CNN, BERT.

I. Introduction

The Internet of Things (IoT) has transformed modern connectivity by enabling smartphones, sensors, and embedded systems to exchange data automatically, with or without any human intervention. IoT architectures are typically organized into three layers perception, network, and application, each with distinct security vulnerabilities that can be exploited by malicious actors [1]. Common IoT attacks include Denial of Service (DoS), malicious control, and unauthorized data access, all of which pose significant threats to privacy, service continuity, and system integrity. While traditional defenses such as firewalls can filter incoming and outgoing packets to prevent unauthorized access, achieving comprehensive IoT security remains challenging due to the increasing complexity of network configurations and the sophistication of emerging threats. Attackers can exploit vulnerabilities even in permitted network traffic, rendering purely reactive defenses are insufficient for safeguarding critical IoT infrastructures. This growing risk landscape underscores the urgent need for proactive, adaptive, and intelligent security mechanisms capable of defending against both known and novel cyber threats.

Intrusion Detection Systems (IDSs) were developed to counter these limitations by continuously monitoring network traffic and identifying suspicious activities through methods such as deep packet inspection. IDSs can be deployed as hardware or software solutions and act as an additional layer of security, issuing alerts to administrators when anomalous or malicious patterns are detected. However, conventional IDS implementations face several well documented challenges. Signature-based IDSs, for example, often exhibit slow detection speeds because they rely on matching observed patterns to predefined signatures, which fails against zero-day exploits and novel attack variants. Moreover, static configurations can result in high false positive rates and poor adaptability to rapidly evolving threat environments. As cyberattacks grow more sophisticated, the limitations of conventional IDSs become more apparent, driving research toward approaches that deliver faster detection, higher adaptability, and improved accuracy while minimizing false alarms in dynamic IoT networks.

To build an effective Intrusion Detection System (IDS), Machine Learning (ML) can be used to develop models capable of distinguishing between normal and malicious network connections. ML enables systems to learn without explicit programming, thus improving the efficiency of data processing [2]. The key objective is to design models that can detect and mitigate threats in real time. Deep Learning (DL) techniques [3][4] further enhance IDS capabilities through automated pattern recognition. Unlike traditional rule based approaches, ML and DL based IDSs can learn complex decision boundaries from historical data and generalize to detect both known and unseen threats. By using features extracted from network telemetry, these models classify traffic with higher accuracy and faster decision times, reducing detection latency. ML and DL approaches improve both robustness and scalability. Moreover, integrating advanced deep learning architectures that capture semantic and structural traffic patterns enables IDS solutions to adapt to the evolving nature of cyberattacks, positioning these IDSs is a vital step towards securing IoT ecosystems against increasingly sophisticated adversaries.

In this research, we employ the UNSW-NB15 Network Intrusion Dataset [5], developed by the University of New South Wales Cyber Security Lab in 2015, which contains more than 2.5 million records with 49 features representing nine attack types Denial of Service (Dos), Shellcode, Analysis, Backdoor, Exploit, Fuzzers, Generic, Reconnaissance, and Worms with addition to Normal traffic. For our study, we focus on binary classification, grouping all attack categories into a single 'Attack' label alongside the 'Normal' class. A major challenge in this dataset is class imbalance, where attacks significantly outnumber normal samples, potentially biasing predictions. To address this, we adopt a two-stage balancing strategy first applying the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic minority samples while preserving the decision boundaries, followed by Edited Nearest Neighbors (ENN) to remove mislabeled and noisy data. Our proposed SecureDeBERTa-CNN hybrid IDS integrates SecureDeBERTa a pretrained version of DeBERTa that underwent training using cybersecurity related data (Books, Articles, Survey Papers, Security Reports, Blogs/News), and contextual semantic feature extraction is achieved with a lightweight CNN for capturing localized spatial patterns. These feature sets are fused and passed to dense layers for final classification. We evaluated performance against classical baseline performance evaluation metrics such including accuracy, precision, recall, and F1 score, demonstrating that our approach delivers superior precision and robustness even under imbalanced conditions. The remainder of this paper is organized as follows Section II reviews related work pertinent to the present study. Section III outlines the proposed methodology for developing the SecureDeBERTa-CNN framework. Section IV presents and discusses the experimental results, and Section V concludes the paper with key findings and potential directions for future research.

II. Related Work

The UNSW-NB15 dataset has become a widely used benchmark for (IDS) evaluation due to its realistic traffic patterns and diverse attack scenarios. Traditional (ML) methods, including ensemble algorithms like Random Forest, Extra Trees, AdaBoost, and XGBoost, have been applied for binary classification, achieving competitive accuracy of 86.99% through diversified decision boundaries [6]. Feature selection approaches, such as gain ratio combined with multi layer perceptron networks, have yielded lightweight IDS frameworks suitable for real-time detection [7] but only got accuracy of 76.96%. While effective, these models rely heavily on manual feature engineering and often struggle with highly complex or evolving attack patterns. To overcome such limitations, (DL) models have been explored for automated hierarchical feature extraction. CNN-LSTM hybrids, for instance, capture both spatial and temporal dependencies, while ANN based IDS solutions optimized with adaptive algorithms like Adam have reported strong classification accuracy of 87%.[8].

Recent work has further integrated attention mechanisms with LSTM networks to prioritize critical traffic features, improving sequential modeling performance [10] ,but suffers with lower recall and f1 score. Similarly, CNN-LSTM [9] architectures optimized via Bayesian methods have been used for IoT intrusion detection tasks, providing effective spatial temporal analysis of network traffic but only get accuracy of 78.47% for binary classification.[11] used an ANN with multiple optimizers, identifying Adam as best 94.83% training accuracy but omitting other metrics. Building on these developments, our approach introduces a hybrid SecureDeBERTa-CNN framework that combines the contextual modeling capabilities of the SecureDeBERTa transformer with the spatial feature extraction strengths of CNN layers. The fused architecture is designed for binary intrusion detection and is trained end-to-end using cross-entropy loss with the AdamW optimizer.

III. Methodology

The SecureDeBERTa-CNN architecture was developed after a detailed evaluation of the shortcomings found in some of the traditional intrusion detection systems and iterative experimentation with hybrid deep learning designs. To achieve high detection accuracy and robustness, the model integrates a dual-branch structure, the SecureDeBERTa transformer for contextual and semantic feature extraction, and a lightweight CNN for capturing localized structural patterns from numerical network features. The data preprocessing pipeline addresses the challenge of class imbalance in the UNSW-NB15 dataset by applying the SMOTE-ENN technique, which combines oversampling of the minority class using SMOTE with noise reduction through (ENN). This hybrid architecture leverages the transformer's contextual understanding with the CNN's pattern recognition capability, producing a fused representation that is highly discriminative for binary classification tasks. The performance evaluation against classical baselines, ML and DL models, demonstrates that the SecureDeBERTa-CNN achieves superior detection accuracy, precision, recall, and better F1 scores. The architecture of the proposed system is illustrated in Figure 1.

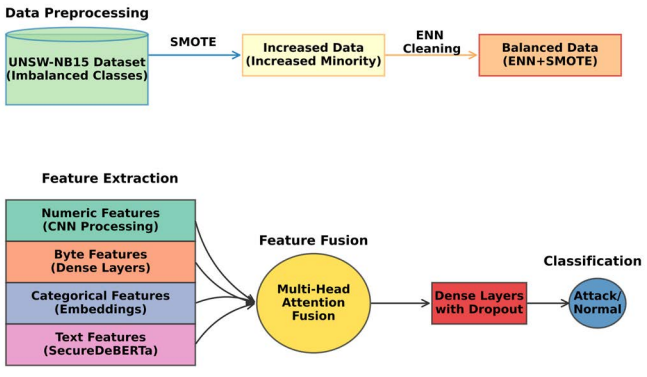

Fig. 1. SecureDeBERTa-CNN System Architecture.

A. Dataset Description

In this study, we employ the UNSW-NB15 dataset, developed in 2015 by the Cyber Security Lab at the University of New South Wales. The dataset was generated in the Cyber Range Lab using the IXIA PerfectStorm tool, which simulates a combination of contemporary benign network activities and diverse attack scenarios. The training and testing subset contains 175,341 and 82,333 records respectively, each described by 44 features, and encompasses nine distinct attack categories like Denial of Service, Exploit, Analysis, Worms, Backdoor, Shellcode, Fuzzers, Generic, and Reconnaissance, in addition to normal traffic. A notable characteristic of the UNSW-NB15 dataset is its class imbalance, where certain attack categories have substantially more samples than others. For the purposes of this research, we perform a binary classification by consolidating all attack categories into a single 'Attack' class (Label = 1) and treating normal traffic as the 'Normal' class (Label = 0). The resulting distribution in the training set consists of 119,341 samples for the attack class and 56,000 samples for the normal class, highlighting the need for effective data balancing strategies.

B. Data Balancing

Addressing the issues related to class imbalanced datasets are known as resampling which involves either reducing the size of the majority class that is called under sampling or increasing the size of the minority class that called oversampling. In comparison, oversampling methods have generally shown better results than randomly removing majority class instances, as done in under sampling methods. Popular oversampling approaches include SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling Technique). In this study, we utilized SMOTE for oversampling, as it preserves information by interpolating between existing minority class instances, avoiding the potential loss of valuable data that occurs with under sampling. However, while SMOTE is effective, it can also introduce noise and lead to overfitting if applied carelessly. To address this, we subsequently applied the ENN method to refine the dataset by removing misclassified and ambiguous instances. This technique identifies and eliminates samples that deviate from the majority of their closest neighbors, thereby improving dataset quality, reducing noise, and enhancing class separation. As a result of applying the SMOTE + ENN pipeline, the final balanced dataset contained 119,341 samples in the Attack class (Label = 1) and 106,081 samples in the Normal class (Label = 0), providing a more reliable and representative training set that ultimately contributes to improved model generalization and classification performance.

C. Dual-Branch Architecture

The architecture follows a dual path design, consisting of a Transformer Path for semantic representation learning and a CNN Path for statistical feature extraction. Both branches operate in parallel on distinct representations of the same input payload and are fused for the final classification.

1) SecureDeBERTa Feature Extraction: The text representation of network traffic (combining protocol, service, state, and packet statistics) is processed using SecureDeBERTa's pretrained tokenizer with a maximum sequence length of 128 tokens. The tokenizer automatically handles padding and truncation while preserving structural information in the network traffic descriptors. The tokenized sequences (input IDs and attention masks) are fed into the SecureDeBERTa model which outputs contextual embeddings. For classification, we extract the final hidden state of the `[CLS]` token (768-dimensional) as the semantic representation of the input. The transformer component remains frozen during the initial training phases, with gradual unfreezing of layers (2→6→12) in subsequent phases to enable fine-grained adaptation to the intrusion detection task while preventing catastrophic forgetting.

2) Hybrid Feature Processing: Numerical features including duration, packet counts, byte volumes, and TTL values are directly extracted from the structured dataset columns. These features undergo Robust Scaling (centering and scaling to interquartile range) rather than standardization. The processed features are reshaped into a temporal dimension and processed through a multi scale 1D CNN architecture with parallel convolutional branches (kernel sizes 3, 5, and 7). Each branch employs Swish activation, followed by batch normalization and dropout (rate=0.2). The architecture includes a novel attention mechanism where a sigmoid-activated convolutional layer learns feature importance weights, which are multiplied with the original features before global max pooling. The resulting compact representation preserves the most salient numerical patterns for intrusion detection.

D. SecureDeBERTa and CNN Branches

The model employs a phased fine-tuning approach with learning rate decay for the warm-up phase only the top 2 transformer layers are trainable with the (learning rate = 5e-5) while maintaining the frozen backbone, allowing initial adaptation of the classification head, later the body phase unfreezes 6 total layers (learning rate = 3e-5) for intermediate feature adaptation finally the finetuning phase unfreezes all 12 layers (learning rate = 2e-5) with gradient clipping (norm=1.0) for final representation tuning.

E. Feature Fusion and Classification

The model combines features through concatenation the [CLS] tokens from SecureDeBERTa which are merged with processed numerical, byte and categorical features through channel-wise fusion which is described in Equation (1)

$$h_{\text{fused}} = [h_{\text{DeBERTa}} | h_{\text{CNN}} | h_{\text{Byte}} | h_{\text{Cat}}] \tag{1}$$

The implementation strengthens robustness by combining byte-level statistics (h_{Byte}), CNN extracted numerical patterns (h_{CNN}), and SecureDeBERTa semantic features ($h_{DeBERTa}$). This fusion enhances resilience to varied attack vectors, enabling more accurate and reliable intrusion detection in complex and dynamic network environments.

IV. Experiments

A. Implementation Environment

The hybrid IDS architecture was implemented using TensorFlow 2.16+ with Keras API, optimized for GPU-accelerated training (CUDA 11.8) on NVIDIA GeForece RTX 4090 GPU. The memory-optimized configuration employed gradient check pointing (every 2 steps), bfloat16 mixed-precision training, and XLA compilation to maximize throughput and global gradient clipping (norm=1.0) to ensure training stability. The model architecture combined a 12-layer SecureDeBERTa backbone with multi scale 1D CNN (kernel sizes 3, 5, and 7) and attention gated feature fusion, culminating in a binary classifier head.

B. Evaluation Metrics

The performance of model is evaluated using standard classification metrics. Accuracy represents the ratio of correctly classified instances to total instances.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision represents the ratio of true positives among all positive predictions.

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall represents the ratio of true positives among all actual positives.

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{4}$$

The Harmonic F1 Score represents the mean of precision and recall.

$$\textbf{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

C. Performance Evaluation

Evaluation metrics are used to measure the performance of a model, we utilized different performance metrics like accuracy, precision, recall and F1 score which are most known to access the actual ability of a model. We tested our model with multiple conditions, at first we tested our hybrid model on imbalance data, for the second condition we implemented ADASYN oversampling technique, for the next experiment we applied SMOTE oversampling and tested our model's performance, and later we applied ADASYN with ENN and SMOTE with ENN receptively and evaluated the outputs which are shown in Table I.

TABLE I
Performance Comparison of Resampling Techniques

Method	Accuracy	Precision	Recall	F1
Imbalanced Data	83.18%	76.93%	98.91%	86.55%
ADASYN	89.06%	93.37%	86.17%	89.63%
SMOTE	85.15%	79.88%	97.62%	87.86%
ADASYN + ENN	90.37%	99.61%	82.83%	90.45%
(SMOTE + ENN)	**92.06%**	**96.91%**	**88.40%**	**92.46%**

In addition to basic evaluation metrics in our experiments, we also calculated the ROC-AUC (area under the Receiver Operating Characteristic curve) and PR-AUC (area under the Precision-Recall curve). In our experiment, the method of (SMOTE + ENN) achieved ROC-AUC of 0.9788 which indicates excellent separability, and PR-AUC of 0.9847 which confirms robustness under class imbalance. The curves show strong discriminative performance.

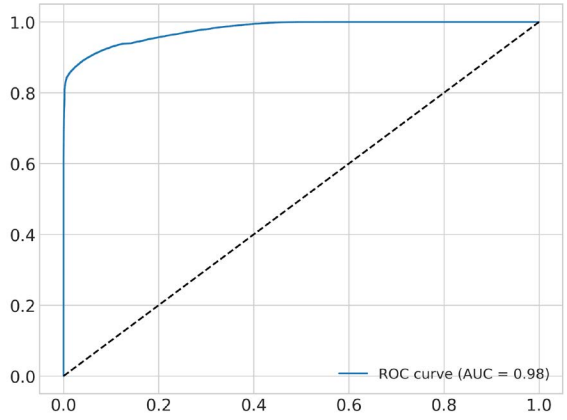

Fig. 2. ROC curve of SecureDeBERTa-CNN (AUC = 0.9788)

The binary classification results on the UNSW-NB15 dataset Table II show that the proposed SecureDeBERTa-CNN consistently outperforms both conventional and deep learning-based IDS models. It achieves the highest accuracy of 92.06%, precision 96.91%, and F1 score 92.46%, alongside strong recall 88.40%, indicating effective detection of both attack and normal traffic with minimal false positives and negatives. AT-LSTM attains similar precision 96.00% but notably lower recall 80.00%, reducing detection coverage. Random Forest delivers competitive accuracy 89.31% yet suffers from reduced recall 85.19%. ANN achieves the highest recall 93.38% but lower precision, increasing false alarms, while DNN shows high precision 95.10% but poor recall 68.40% limiting real-world applicability. CNN-LSTM variants provide balanced but generally inferior results. Overall, SecureDeBERTa-CNN offers the best balance of precision and recall, ensuring reliable intrusion detection performance.

TABLE II
Binary Classification Results on UNSW-NB15

Method	Accuracy	Precision	Recall	F1
Random Forest[6]	89.31%	92.96%	85.19 %	91.56%
SVM	82.89%	89.78%	75.79%	85.77%
ANN [7]	86.40%	86.74%	93.38%	89.94%
DNN [4]	76.10%	95.10%	68.40%	79.60%
AT-LSTM [10]	92.00%	96.00%	80.12%	87.01%
CNN-LSTM [8]	87.10%	85.21%	88.01%	86.12%
Op CNN-LSTM [11]	78.46%	69.69%	79.69%	43.60%
(SecureDeBERTa-CNN)	**92.06%**	**96.91%**	**88.40%**	**92.46%**

V. Conclusion

In this study, we proposed a hybrid intrusion detection system (SecureDeBERTa-CNN) that effectively combines the contextual understanding of SecureDeBERTa with the spatial pattern recognition capabilities of a lightweight CNN. By integrating a SMOTE-ENN preprocessing pipeline and employing progressive layer unfreezing during training, our approach addresses both class imbalance and optimization stability two key challenges in intrusion detection for complex network environments. Evaluated on the UNSW-NB15 dataset, the proposed model achieves strong performance across multiple metrics, including a 92.06% Accuracy, 96.91% precision, 88.40% recall and 92.46% f1 score, outperforming traditional machine learning baselines such as Random Forest and SVM. These results validate the effectiveness of hybrid transformer-CNN architectures in binary threat classification tasks. Future work will explore multi class extension, real-time deployment on edge devices, and generalization to other IoT and cyber physical datasets.

References

[1] H. HaddadPajouh, A. Dehghantanha, R. M. Parizi, M. Aledhari, and H. Karimipour, "A survey on internet of things security: Requirements, challenges, and solutions," Internet of Things (Netherlands), vol. 14, Jun., 2021

[2] Mahesh, B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386., 2020.

[3] Kim, T. and Pak, W. Deep learning-based network intrusion detection using multiple image transformers. Appl. Sci. 13(5). https://doi.org/10.3390/app13052754 (2023).

[4] Vinayakumar, R., Alazab, M., Member, S. , Soman, K. P. Deep learning approach for intelligent intrusion detection system. IEEE Access 7, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334 (2019).

[5] Moustafa, N., Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). in Mil. Commun. Inf. Syst. Conf. MilCIS 2015 - Proc. (2015). https://doi.org/10.1109/MilCIS.2015.7348942.

[6] Sharma, N., Yadav, N. S. , Sharma, S. Classification of UNSW-NB15 dataset using exploratory data analysis using ensemble learning. EAI Endorsed Trans. Ind. Networks Intell. Syst. 8, 1–10. https://doi.org/10.4108/EAI.13-10-2021.171319 (2021).

[7] Mebawondu, J. O., Alowolodu, O. D., Mebawondu, J. O. Adetunmbi, A. O. Network intrusion detection system using supervised learning paradigm. Sci. Afr. 9, 11–21. https://doi.org/10.1016/j.sciaf.2020.e00497 (2020).

[8] Chen, Y. Framework design of network intrusion detection based on convolutional neural networks. Procedia Comput. Sci. 261, 1356–1362. https://doi.org/10.1016/j.procs.2025.05.013 (2025).

[9] Thaljaoui, A. Intelligent network intrusion detection system using optimized deep CNN-LSTM with UNSW-NB15. Int. J. Inf. Technol. https://doi.org/10.1007/s41870-025-02416-0 (2025).

[10] Alsharaiah, M. A. et al. An innovative network intrusion detection system (NIDS): Hierarchical deep learning model based on Unsw-Nb15 dataset. Int. J. Data Netw. Sci. 8 (2), 709–722. https://doi.org/10.5267/j.ijdns.2024.1.007 (2024).

[11] Arun, C. B. et al. Enhancing Network Intrusion Detection using Artificial Neural Networks: An Analysis of the UNSW-NB15 dataset. in 2nd IEEE Int. Conf. Integr. Intell. Commun. Syst. ICIICS 2024 (2024). https://doi.org/10.1109/ICIICS63763.2024.10859396.