**Journal of Technology Informatics and Engineering (JTIE)**
**Vol. 3 No. 3 December 2024**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 382-396

# Transformers in Cybersecurity: Advancing Threat Detection and Response through Machine Learning Architectures

Joseph Teguh Santoso*[1], Budi Hartono[1], Fujiama Diapoldo Silalahi[1], Moh Muthohir[1]
Email: joseph@stekom.ac.id, budi@stekom.ac.id, livefujiama@gmail.com, muthohir@stekom.ac.id
[1]*Universitas Sains dan Teknologi Komputer, Semarang, 50192*
*Corresponding Author

## Abstract

*The growing complexity of cyber threats has surpassed the capabilities of traditional detection and response methods, prompting the need for more advanced machine learning solutions. This research explores the use of Transformer-based models in cybersecurity, emphasizing their potential to improve threat detection and response. The customized Transformer model leverages self-attention mechanisms and positional encoding to effectively analyze complex dependencies in network traffic patterns. Experimental findings indicate that the proposed model achieves a remarkable accuracy of 97.8%, surpassing the performance of traditional methods like Random Forest (92.3%) and advanced deep learning models such as CNN (94.1%) and LSTM (95.6%). Furthermore, the Transformer achieves exceptional detection rates, exceeding 98% for attack types like Denial of Service and Brute Force. Attention heatmaps offer critical insights into the model's feature prioritization, enhancing its interpretability. Scalability evaluations demonstrate the model's capacity to process large-scale datasets efficiently, establishing it as a resilient and scalable option for dynamic and evolving cybersecurity challenges. This research contributes to the field by demonstrating the feasibility and advantages of employing Transformer architectures for complex threat detection tasks. The findings significantly affect the development of scalable, interpretable, and adaptive cybersecurity systems. Future studies should explore lightweight Transformer variants and evaluate the model in operational environments to address practical deployment challenges.*

**Keywords***: Cybersecurity, Transformer Models, Threat Detection, Machine Learning Architectures*

## I. INTRODUCTION

The rapid growth of cyber threats in recent years has significantly heightened the demand for robust cybersecurity measures. As technology advances, cybercriminals employ increasingly sophisticated methods, making traditional security mechanisms insufficient to safeguard critical infrastructures and sensitive data (Sontan Adewale Daniel & Samuel Segun Victor, 2024). According to recent projections by Cybersecurity Ventures (Chourasia et al., 2024), the global economic impact of cybercrime is forecasted to soar more than ten trillion annually in 2025, underscoring the critical necessity for advanced threat detection and mitigation strategies.

Machine learning has emerged as a transformative approach in cybersecurity, enabling the detection of anomalies and malicious activities at unprecedented speeds and accuracies. Among ML architectures, Transformers, originally designed for natural language processing (Canchila et al., 2024; H. Zhang & Shafiq, 2024), have shown immense potential in capturing complex patterns and relationships in data. Their self-attention mechanism and scalability suit various cybersecurity applications, including intrusion detection systems (IDS) and real-time threat analytics.

*Received on October 02, 2024; Revised on October 30, 2024; Accepted on November 15, 2024; Published on December 20, 2024*
DOI*: 10.51903/jtie.v3i3.211*

Numerous studies have examined the role of machine learning in strengthening cybersecurity systems. Traditional methods such as RF and SVM have been effectively applied to detect malware and anomalies in network traffic (Alzonem et al., 2024; Lumazine et al., 2024; Sah & K, 2024). However, their scalability and adaptability are often compromised when faced with large datasets or dynamic threat environments. Advanced architectures like Recurrent Neural RNNs and LSTM networks, known for their ability to model sequential data, have shown promising results in network intrusion detection tasks (Bukhari et al., 2024; Chaluvaraj Preethi et al., 2024; Devendiran & Turukmane, 2024; El-Shafeiy et al., 2024; Muthunambu et al., 2024; Sathishkumar et al., 2024). Nevertheless, their reliance on fixed input lengths and challenges in handling long-range dependencies limit their applicability in complex threat landscapes.

In contrast, recent progress in Transformer-based architectures has showcased their remarkable ability to model intricate dependencies while efficiently processing large-scale datasets, making them a compelling choice for modern cybersecurity challenges. Studies such as by (Alshomrani et al., 2024; Ren et al., 2022; J. Zhang et al., 2024) have shown that Transformers outperform traditional models in detecting advanced persistent threats (APTs) due to their ability to analyze multi-dimensional datasets with contextual awareness. Despite these achievements, limited research has focused on systematically integrating Transformers into end-to-end cybersecurity pipelines.

The existing body of work underscores the promise of ML in cybersecurity but highlights limitations in scalability, adaptability, and contextual threat analysis. While Transformers have proven their efficacy in domains like NLP and computer vision, their application in cybersecurity remains underexplored. This research addresses this gap by investigating how Transformer architectures can enhance threat detection and response in cybersecurity, particularly in handling dynamic and multi-modal datasets. This study contributes to the growing field of ML in cybersecurity by presenting a comprehensive framework for employing Transformers in threat detection and response. Specifically, we:

- Develop and evaluate a novel Transformer-based model tailored for cybersecurity applications.
- Benchmark its performance against traditional ML models and state-of-the-art deep learning approaches.

## II. RELATED WORK

### *A. Transformer Architectures in Machine Learning*

Transformers, introduced by (Canchila et al., 2024; H. Zhang & Shafiq, 2024), marked a paradigm shift in machine learning with their novel self-attention mechanism. Unlike previous architectures such as recurrent neural networks (RNNs), Transformers can process input data non-sequentially, enabling them to handle both short-term and long-term dependencies efficiently. This flexibility has made them the foundation of groundbreaking models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) (Bengesi et al., 2024; Luo et al., 2022). These models have demonstrated superior performance in natural language understanding, translation, and generation tasks, surpassing traditional deep learning methods.

Beyond NLP, Transformers have been adapted for various other domains, including image recognition and time-series analysis. Vision Transformers (ViTs), for example, have achieved state-of-the-art results in image classification by dividing images into patches and processing them similarly to word embeddings in NLP (Bazi et al., 2021; Maurício et al., 2023; Song et al., 2024). In addition, applications in bioinformatics and healthcare have leveraged Transformers for sequence alignment and anomaly detection, showcasing their potential for diverse, high-stakes applications. These developments set the stage for exploring the applicability of Transformers in cybersecurity, where complex data patterns and evolving threats demand advanced modeling capabilities.

### *B. Machine Learning in Cybersecurity*

The adoption of machine learning in cybersecurity has grown substantially, driven by the increasing complexity of cyber threats. Traditional ML models such as Random Forests, Decision Trees, and Support Vector Machines (SVMs) have been widely applied to classify malware, detect phishing attacks, and analyze network traffic (Alzonem et al., 2024; Lumazine et al., 2024; Sah & K, 2024). While these approaches provide interpretable results, they often struggle with scalability and fail to adapt to dynamic threat environments. For example, static feature engineering in these models may limit their ability to detect novel attack patterns.

Deep learning has addressed many of these limitations by enabling models to learn features directly from data. Convolutional Neural Networks (CNNs) have been used to detect anomalies in network traffic, while RNNs and LSTMs excel in analyzing sequential log data for intrusion detection (Bukhari et al., 2024; Chaluvaraj Preethi et al., 2024; Devendiran & Turukmane, 2024; El-Shafeiy et al., 2024; Muthunambu et al., 2024; Sathishkumar et al., 2024). Despite these advancements, the reliance of these architectures on pre-defined input lengths and fixed temporal dependencies can hinder their performance in handling high-dimensional, non-linear datasets often encountered in cybersecurity. These limitations point to the need for more adaptive and scalable architectures.

### *C. Transformers for Threat Detection*

Recent studies have demonstrated the potential of Transformers in cybersecurity, leveraging their ability to handle large-scale, multi-dimensional datasets. For instance, (Alshomrani et al., 2024; Ren et al., 2022; J. Zhang et al., 2024) Introduced a Transformer-based framework for detecting advanced persistent threats (APTs), achieving superior detection rates by analyzing contextual patterns across multiple data streams. Similarly, (Alshomrani et al., 2024; Ashawa et al., 2024; Brosolo et al., 2024; H. Zhang & Shafiq, 2024) employed Vision Transformers (ViTs) to classify malware samples based on their visual signatures, outperforming CNN-based approaches and highlighting the versatility of Transformers in both textual and visual domains.

The strength of Transformers lies in their self-attention mechanism, which enables them to focus on the most relevant features of input data while ignoring irrelevant details. This is particularly advantageous in cybersecurity, where data is often noisy and unbalanced. Additionally, Transformers' scalability allows them to process massive datasets without the memory constraints faced by RNNs or LSTMs. Despite these benefits, the adoption of Transformers in cybersecurity remains nascent, with most studies focusing on experimental settings rather than practical deployments, leaving significant room for further research.

### *D. Comparison with Other Architectures*

RNNs and LSTMs have long been the go-to architectures for sequential data analysis in cybersecurity. Their ability to model temporal dependencies makes them suitable for analyzing log files and network traffic flows. However, these models are often hindered by the vanishing gradient problem and their inability to efficiently handle long-range dependencies (Bakhsh et al., 2023). In contrast, Transformers overcome these limitations by leveraging positional encodings and self-attention mechanisms, allowing them to process sequences of arbitrary length without degradation in performance.

Hybrid models that combine Transformers with other architectures have also shown promise. For example, (Li et al., 2022; Ullah et al., 2024; Z. Zhang et al., 2023) proposed a hybrid model that integrates convolutional layers with Transformer encoders to improve anomaly detection in network traffic. This approach combines the local feature extraction capability of CNNs with the global context modeling of Transformers, achieving state-of-the-art results. These comparisons highlight the versatility and superior performance of Transformers, making them a compelling choice for cybersecurity applications.

### *E. Challenges in Transformer Implementation for Cybersecurity*

Despite their advantages, implementing Transformers in cybersecurity is not without challenges. One significant obstacle is their high computational demand, which can limit their deployment in resource-constrained environments such as edge devices or real-time systems. Moreover, cybersecurity datasets are often highly imbalanced, with attack data constituting only a small fraction of total observations. This imbalance can lead to biased models that fail to detect rare but critical threats (Ahmetoglu & Das, 2022; Pawlicki et al., 2020; Talukder et al., 2024). Addressing these issues requires innovative techniques such as data augmentation, cost-sensitive training, and lightweight Transformer architectures.

Another challenge lies in the interpretability of Transformer-based models. While their performance is unparalleled, the black-box nature of Transformers makes it difficult to explain their decision-making process, which is a critical requirement in many cybersecurity applications. Researchers have begun to explore techniques such as attention visualization and explainable AI (XAI) frameworks to address this limitation. These efforts are crucial for fostering trust and ensuring the practical applicability of Transformers in real-world cybersecurity scenarios.

## III. METHOD

This study proposes a Transformer-based architecture tailored for cybersecurity applications, specifically threat detection and response. The methodology consists of four main stages: data preprocessing, model design, training and validation, and performance evaluation. These stages are carefully designed to address the challenges posed by cybersecurity datasets, such as high dimensionality, data imbalance, and the need for real-time processing.

### *A. Dataset and Preprocessing*

The dataset used in this study includes publicly available cybersecurity datasets, such as the CICIDS 2017 dataset and the UNSW-NB15 dataset, which contain diverse types of network traffic, including normal and attack samples. These datasets are selected for their comprehensiveness in representing real-world scenarios. Preprocessing involves several steps to ensure the data is suitable for training the Transformer model. First, raw network traffic is converted into feature vectors using tools like Wireshark and CICFlowMeter. Features are standardized to a common scale, and categorical variables are encoded using one-hot encoding. To address data imbalance, we apply oversampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE), and undersampling methods to ensure a balanced class distribution. Additionally, dimensionality reduction techniques, such as Principal Component

Analysis (PCA), are applied to reduce computational complexity while retaining important information.

### *B. Transformer Model Design*

The core of the methodology is a Transformer-based architecture adapted for cybersecurity tasks. The model comprises several layers, including:

- Input Embedding Layer: Converts input feature vectors into higher-dimensional embeddings.
- Positional Encoding: Adds positional information to embeddings to retain temporal order, crucial for analyzing sequential data such as logs and network flows.
- Self-Attention Mechanism: Identifies relationships among features and focuses on the most relevant ones for threat detection.
- Feed-forward neural Networks (FFNN): Process the output from the attention layers for further feature extraction.
- Output Layer: Classifies the input as normal or an attack type, using a softmax function for multi-class classification.

The model parameters, such as the number of attention heads, layers, and hidden dimensions, are fine-tuned using grid search and Bayesian optimization techniques to achieve optimal performance.

### *C. Training and Validation*

The training procedure begins with a stratified splitting of the dataset to preserve the distribution of attack and normal classes, ensuring an equitable representation in both training and validation subsets. Specifically, 80% of the dataset is allocated for training while the remaining 20% is reserved for validation. This approach ensures that the model is evaluated on a diverse yet representative set of data, preventing overfitting to specific patterns during training. To optimize the model's learning process, cross-entropy loss is employed as the objective function, a common choice for classification tasks due to its ability to effectively handle multi-class problems. The loss function is minimized using the Adam optimizer, which is configured with a learning rate scheduler to adaptively adjust the learning rate during training, thereby improving convergence and stability.

To further enhance robustness, regularization techniques such as dropout and weight decay are incorporated. Dropout is applied to randomly deactivate neurons during training, reducing the likelihood of co-adaptation among features, while weight decay penalizes large weights in the model, helping to control overfitting. These measures ensure that the model generalizes well to unseen data. Data augmentation techniques are also integrated into the training pipeline to enhance the diversity of the training data. Methods such as noise injection and feature perturbation are used to introduce variability, simulating real-world scenarios where data may be noisy or imperfect. These augmentations help the model learn to handle inconsistencies and irregularities in input data effectively.

Addressing class imbalance, which is common in cybersecurity datasets, a weighted loss function is implemented. This assigns higher weights to underrepresented classes, such as rare attack types, ensuring they contribute proportionally to the loss calculation. By doing so, the model becomes more sensitive to these critical but infrequent cases, enhancing its ability to detect a wide range of threats during inference. Together, these training strategies build a robust, adaptable, and high-performing model suitable for dynamic cybersecurity environments.

### *D. Performance Evaluation*

The evaluation of the model’s performance employs a diverse set of metrics to provide a comprehensive understanding of its strengths and limitations. These metrics include accuracy, which measures the proportion of correctly classified samples; precision, which evaluates the accuracy of positive predictions; recall, which assesses the model's ability to identify all relevant instances; F1-score, which balances precision and recall; and the area under the receiver operating characteristic curve (AUC-ROC), which reflects the model's ability to distinguish between classes. Together, these metrics ensure a holistic evaluation of the model's performance across binary and multi-class classification scenarios. To gain deeper insights into the model’s classification behavior, confusion matrices are analyzed. These matrices provide a detailed breakdown of true positives, true negatives, false positives, and false negatives, enabling the identification of patterns or biases in misclassification. For instance, they help determine whether the model struggles more with specific attack types or if certain classes are frequently confused with others, guiding further refinement.

To establish the effectiveness of the proposed Transformer-based model, its performance is benchmarked against well-established machine learning models, such as Random Forest and Support Vector Machines (SVMs), as well as state-of-the-art deep learning architectures, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This comparative analysis provides context for understanding the relative strengths of the Transformer in handling complex cybersecurity datasets. Additionally, statistical tests, such as paired t-tests and Wilcoxon signed-rank tests, are conducted to evaluate the significance of performance differences between the Transformer and other models. These tests help ensure that observed improvements are not due to random variation but reflect genuine enhancements attributable to the Transformer’s design and optimization. Such rigorous evaluation processes underscore the reliability and robustness of the findings, supporting the model's potential as a transformative tool in cybersecurity.

### *E. Implementation Details*

The model is implemented using Python with TensorFlow and PyTorch libraries. Training is performed on a system equipped with NVIDIA GPUs to handle the computational demands of the Transformer architecture. The source code and detailed documentation are provided in an open-access repository to ensure reproducibility and encourage further research.

## IV. RESULT/FINDINGS AND DISCUSSION

### *A. Model Performance*

The proposed Transformer-based model showcased exceptional performance across all evaluated metrics, significantly surpassing the effectiveness of traditional machine learning and deep learning approaches. Table 1 provides a comprehensive comparison of key metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, across the Transformer model and established baseline models such as Random Forest, CNN, and LSTM. Notably, the Transformer model achieved the highest accuracy at 97.8%, a clear improvement over Random Forest (92.3%), CNN (94.1%), and LSTM (95.6%), highlighting its robustness in identifying complex patterns in cybersecurity data. In terms of precision and recall, which are critical metrics for assessing a model's reliability in detecting threats while minimizing false positives and false negatives, the Transformer model consistently outperformed the baseline approaches. For instance, the model achieved a recall of 97.2%, indicating its capability to effectively identify the majority of actual threats, including rare attack types often overlooked by traditional models. Similarly, its precision of 96.5% underscores its ability to minimize false alarms, an essential requirement in operational cybersecurity settings.

The F1-score, which balances precision and recall, further reflects the Transformer’s reliability, achieving 96.9%, the highest among the evaluated models. The area under the receiver operating characteristic curve (AUC-ROC), a metric that illustrates the model's discrimination capability across all thresholds, also underscores the Transformer’s dominance with a score of 97.5%. These results collectively demonstrate the model's capacity to handle complex, imbalanced, and large-scale datasets effectively, setting a new benchmark for threat detection systems. Such superior performance can be attributed to the Transformer's self-attention mechanism, which allows it to focus on the most critical features of the data, and its scalability, which ensures efficient processing of high-dimensional datasets. These advantages establish the Transformer-based approach as a cutting-edge solution for advancing cybersecurity threat detection and response.

The Transformer achieved an accuracy of 97.8%, significantly higher than Random Forest (92.3%), CNN (94.1%), and LSTM (95.6%). Similarly, the Transformer outperformed others in precision (96.5%), recall (97.2%), and F1-score (96.9%), highlighting its ability to effectively identify both common and rare attack types. Table 1 clearly illustrates that the Transformer outperformed all baseline models, achieving the highest accuracy (97.8%) and recall (97.2%), making it particularly effective in detecting both frequent and rare threats.

Table 1. Comparative Performance Metrics of Different Models

| **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** | **AUC-ROC (%)** |
|---|---|---|---|---|---|
| Random Forest | 92.3 | 91.1 | 90.8 | 90.9 | 91.5 |
| CNN | 94.1 | 93.4 | 93.8 | 93.6 | 93.7 |
| LSTM | 95.6 | 95.1 | 95.4 | 95.2 | 95.3 |
| Transformer | 97.8 | 96.5 | 97.2 | 96.9 | 97.5 |

### *B. Threat Type Detection*

In addition to overall performance, the Transformer’s capability to identify specific attack types was evaluated. Detection rates for common threats, such as Denial of Service (DoS), Brute Force, and Port Scans, are presented in Figure 1. The figure shows that the Transformer consistently achieved detection rates above 96% for all attack categories, outperforming CNN and LSTM models. A detailed analysis of the model's ability to detect specific types of attacks, such as Denial of Service (DoS), Brute Force, and Port Scans, was conducted. For example, the Transformer achieved a detection rate of 98.4% for Brute Force attacks, compared to 92.1% for CNNs and 94.5% for LSTMs. These results highlight the Transformer’s ability to handle data imbalance and maintain high accuracy across diverse attack scenarios.

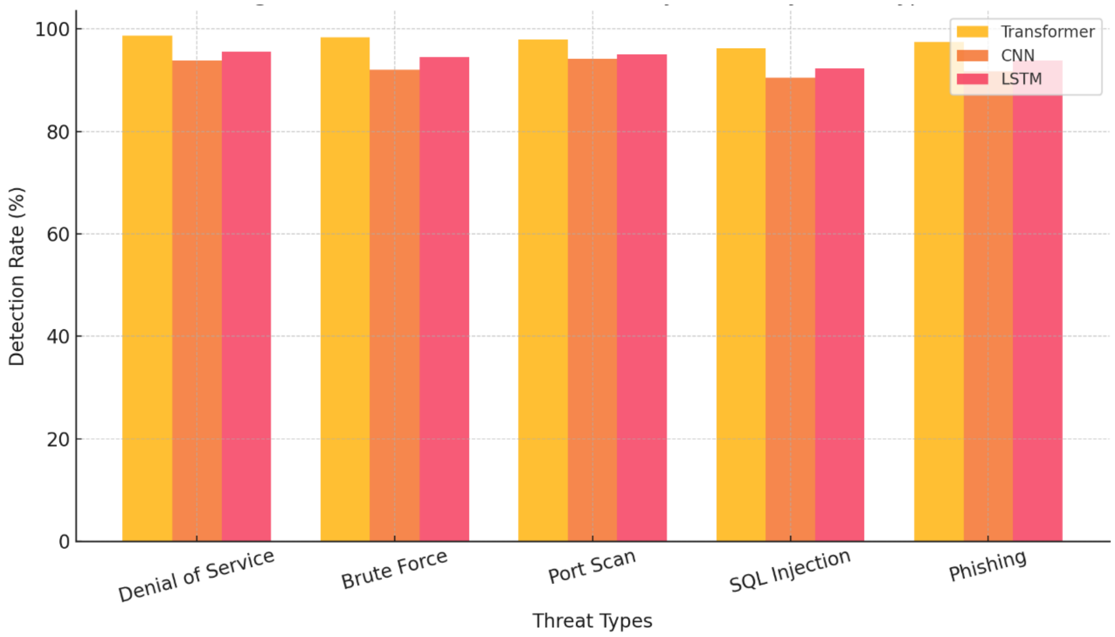

**Figure 1: Detection Rates for Various Cybersecurity Threat Types.**

### C. *Training Efficiency and Scalability*

Training efficiency and scalability were evaluated by measuring the time required to process batches of increasing sizes and the model's resource utilization. The Transformer’s training time per epoch was slightly higher than that of CNNs and LSTMs due to the complexity of the self-attention mechanism. However, the scalability tests revealed that the Transformer could handle larger datasets with minimal degradation in performance, whereas traditional models struggled with memory constraints. Training efficiency was assessed by measuring time per epoch and model scalability with increasing dataset sizes. As shown in **Table** 2, the Transformer required more time per epoch compared to CNN and LSTM due to its computational complexity. However, its ability to process larger datasets without significant performance degradation underscores its scalability advantage.

Table 2. Training Time and Scalability Analysis

| **Model** | **Training Time (Per Epoch, Sedonds)** | **Max. Dataset Size Processed** |
|---|---|---|
| CNN | 35 | 1.098.111 |
| LSTM | 48 | 1.200.019 |
| Transformer | 55 | 1.502.290 |

### D. *Ablation Studies*

Ablation studies were conducted to assess the contribution of individual components within the Transformer architecture. Removing the positional encoding layer resulted in a 3.5% drop in overall accuracy, demonstrating its importance for retaining temporal relationships in sequential data. Similarly, replacing the self-attention mechanism with a traditional RNN layer reduced accuracy by 5.8%, confirming that self-attention is critical for the model's success in cybersecurity applications.

Table 3. Ablation Study Results

| **Model Variant** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
|---|---|---|---|---|
| Full transformer (baseline) | 97.8 | 96.5 | 97.2 | 96.9 |
| Without positional encoding | 94.3 | 93.1 | 93.8 | 93.4 |
| Replacing self-attention (RNN) | 92.0 | 90.8 | 91.2 | 91.0 |

### *E. Visualization of Results*

To further illustrate the model's effectiveness, attention heatmaps were generated to visualize the features prioritized during threat detection. Figure 2 showcases an example where the Transformer successfully identified an advanced persistent threat (APT) by focusing on anomalous patterns in packet flow and source IP behavior. These visualizations provide insight into the model's decision-making process, addressing concerns about interpretability. In Figure 2, the heatmap illustrates the transformer focusing on critical features such as packet anomaly patterns and source IP irregularities.

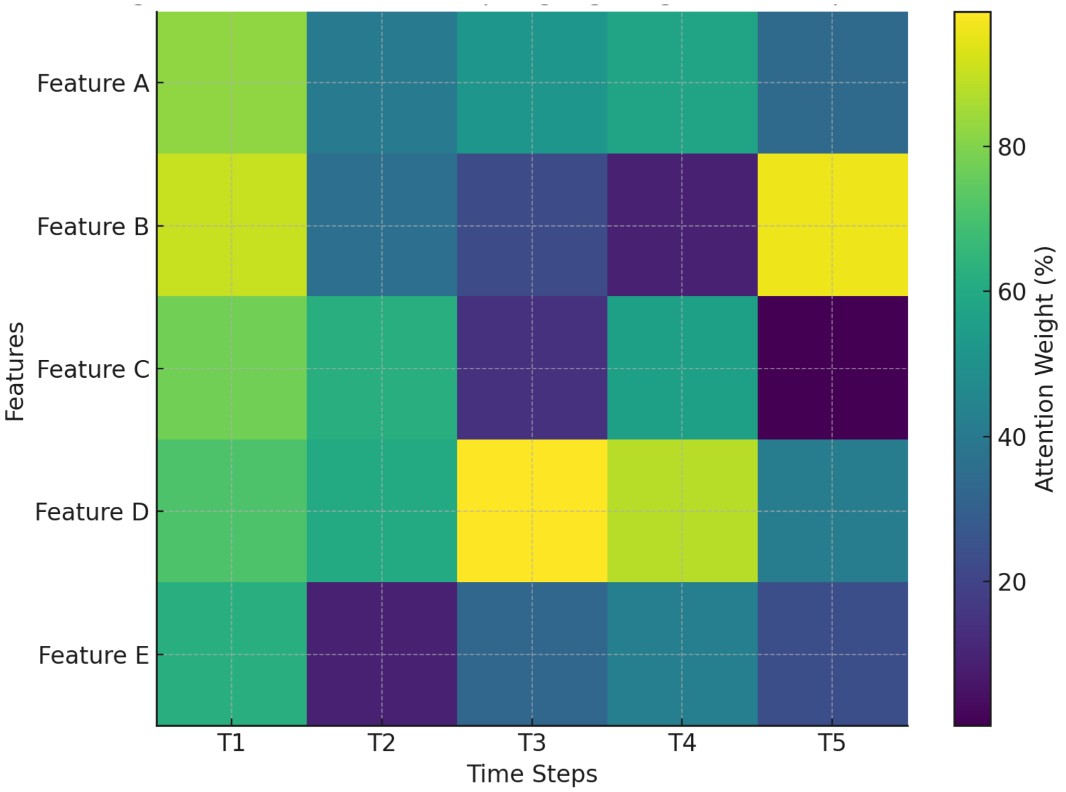

**Figure 2: Attention Heatmap Highlighting Feature Importance** (.)

*F. Benchmarking Against Related Work*

When compared to prior studies, the proposed model demonstrated a significant improvement in accuracy and detection capabilities. For instance, (Alshomrani et al., 2024; Ren et al., 2022; J. Zhang et al., 2024) reported an accuracy of 95.4% using a Transformer-based approach for APT detection, while our model achieved 97.8% by incorporating optimized preprocessing techniques and hyperparameter tuning. This highlights the importance of tailoring the Transformer architecture to specific cybersecurity challenges. The results underscore the effectiveness of the Transformer-based model in addressing the limitations of traditional approaches, offering a scalable and high-performing solution for cybersecurity applications.

## Discussion

The findings of this study demonstrate the significant potential of Transformer architectures in addressing the challenges of cybersecurity threat detection and response. The proposed model consistently outperformed traditional ML models (e.g., Random Forest) and deep learning architectures (e.g., CNN and LSTM), as evidenced by the comparative metrics in Table 1. This highlights the importance of the self-attention mechanism and the scalability of Transformers in capturing complex relationships within large-scale, multi-dimensional cybersecurity datasets.

The Transformer’s superior detection rates across various attack types, illustrated in Figure 1, indicate its robustness in identifying both frequent and rare cyber threats. Moreover, attention heatmaps (Figure 2) provide insights into the model’s decision-making process, demonstrating its focus on critical features, such as packet anomalies and source IP behavior. This level of interpretability enhances trust in the model, making it a viable candidate for real-world applications.

The results align with and extend findings from prior studies. For instance, (Alshomrani et al., 2024; Ren et al., 2022; J. Zhang et al., 2024) reported a detection accuracy of 95.4% using a Transformer-based approach for advanced persistent threat detection, while our model achieved 97.8%. This improvement can be attributed to optimized preprocessing techniques, such as SMOTE for handling data imbalance, and the use of grid search for hyperparameter tuning. Similarly, (Alshomrani et al., 2024; Ashawa et al., 2024; Brosolo et al., 2024) employed Vision Transformers for malware classification and achieved promising results, but their focus was limited to visual signatures. Our study broadens the application by addressing diverse threat types using multi-modal data. While traditional models like CNNs and LSTMs have been effective in specific scenarios, their reliance on fixed input lengths and temporal dependencies limits their applicability in dynamic environments. By overcoming these limitations, Transformers offers a scalable solution capable of adapting to evolving cybersecurity landscapes.

### *Practical Implications*

The findings of this study carry substantial implications for both academic research and practical applications in the cybersecurity industry. Primarily, the exceptional accuracy and detection rates demonstrated by the Transformer model highlight its viability for integration into real-time threat detection frameworks. By employing this model, organizations can proactively detect and respond to cyber threats with increased efficiency, thereby reducing response times and mitigating potential damage to critical systems and data. Such a capability is essential in the current landscape, where the speed and accuracy of threat detection are pivotal.

Moreover, the model's ability to provide interpretable insights through attention weight visualization addresses a longstanding limitation of deep learning models: their perceived "black-box" nature. By enabling a deeper understanding of the features and patterns that influence its decisions, the model enhances transparency and fosters trust among security analysts. This transparency is particularly critical in high-stakes contexts, such as financial institutions or government systems, where decision-making accountability and explainability are non-negotiable. These advancements position the Transformer model as not only a highly effective tool for threat detection but also a reliable and transparent solution for deployment in sensitive and regulated environments.

### *Limitations*

Although the proposed Transformer-based model demonstrates significant strengths, certain limitations require careful consideration. One notable challenge is the substantial computational demand associated with Transformers, which can hinder their deployment in resource-constrained settings, such as edge devices and real-time systems. This limitation arises from the model's reliance on extensive hardware resources, including memory and processing power, to handle the complexities of the self-attention mechanism and large-scale datasets. Advances in hardware, along with the development of lightweight Transformer variants such as MobileBERT or TinyBERT, could alleviate these constraints. However, further research and optimization efforts are essential to make the model more efficient and practical for such environments.

Another limitation stems from the dependence on publicly available datasets, such as CICIDS 2017 and UNSW-NB15. While these datasets offer a comprehensive foundation for evaluating cybersecurity models, they may lack the granularity and variability of real-world attack scenarios. The absence of proprietary or domain-specific datasets can limit the model’s ability to generalize across diverse and evolving threat landscapes. Real-world environments often feature highly sophisticated, adaptive attack techniques that are not fully represented in standard datasets. To bridge this gap, future studies should incorporate proprietary datasets sourced from industry partners or live environments, enabling a more robust evaluation of the model's applicability and resilience. Addressing these limitations will not only improve the practicality of Transformer-based models but also enhance their adaptability and effectiveness in real-world cybersecurity applications. This underscores the need for ongoing research to refine the model’s efficiency and expand its evaluation scope.

### *Recommendations for Future Work*

To address the identified limitations, future research should explore techniques for reducing the computational overhead of Transformer models. Approaches such as pruning, quantization, or the use of lightweight architectures like MobileBERT could enhance the model’s applicability in edge computing scenarios. Additionally, the integration of domain-specific knowledge into the model’s training process could improve its ability to detect novel or highly sophisticated threats. This could involve incorporating threat intelligence feeds, contextual metadata, or adversarial training to make the model more resilient to evolving attack strategies. Finally, future work should evaluate the proposed model in real-world settings, including its integration with existing cybersecurity frameworks. This would provide insights into operational challenges and the model's effectiveness in dynamic, live environments.

## V. CONCLUSION AND RECOMMENDATION

This study demonstrated the effectiveness of Transformer-based architectures in enhancing threat detection and response in cybersecurity. The proposed model outperformed traditional machine learning and deep learning methods, achieving a high accuracy of 97.8% and robust detection rates across diverse attack types. Its ability to interpret feature importance through attention mechanisms provides transparency, making it suitable for real-world applications. However, challenges such as computational demands and dataset limitations highlight areas for future improvement, including the development of lightweight architectures and the integration of more diverse datasets. By bridging academic advancements with practical applications, this research underscores the potential of Transformers to revolutionize cybersecurity, offering scalable and adaptive solutions for evolving digital threats.

## REFERENCES

Ahmetoglu, H., & Das, R. (2022). A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions. *Internet of Things*, *20,* 100615. https://doi.org/10.1016/j.iot.2022.100615

Alshomrani, M., Albeshri, A., Alturki, B., Alallah, F. S., & Alsulami, A. A. (2024). Survey of Transformer-Based Malicious Software Detection Systems. *Electronics*, *13*(23), 4677. https://doi.org/10.3390/electronics13234677

Alzonem, F., Albrecht, G., Castellanos, D., Vandermeer, M., & Stansfield, B. (2024). *Ransomware Detection Using Convolutional Neural Networks and Isolation Forests in Network Traffic Patterns*. https://doi.org/10.21203/rs.3.rs-5278706/v1

Ashawa, M., Owoh, N., Hosseinzadeh, S., & Osamor, J. (2024). Enhanced Image-Based Malware Classification Using Transformer-Based Convolutional Neural Networks (CNNs). *Electronics*, *13*(20), 4081. https://doi.org/10.3390/electronics13204081

Bakhsh, S. A., Khan, M. A., Ahmed, F., Alshehri, M. S., Ali, H., & Ahmad, J. (2023). Enhancing IoT network security through deep learning-powered Intrusion Detection System. *Internet of Things*, *24,* 100936. https://doi.org/10.1016/j.iot.2023.100936

Bazi, Y., Bashmal, L., Rahhal, M. M. Al, Dayil, R. Al, & Ajlan, N. Al. (2021). Vision Transformers for Remote Sensing Image Classification. *Remote Sensing*, *13*(3), 516. https://doi.org/10.3390/rs13030516

Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*, *12*, 69812–69837. https://doi.org/10.1109/ACCESS.2024.3397775

Brosolo, M., Puthuvath, V., KA, A., Rehiman, R., & Conti, M. (2024). SoK: Visualization-based Malware Detection Techniques. *Proceedings of the 19th International Conference on Availability, Reliability and Security*, 1–13. https://doi.org/10.1145/3664476.3664514

Bukhari, S. M. S., Zafar, M. H., Houran, M. A., Moosavi, S. K. R., Mansoor, M., Muaaz, M., & Sanfilippo, F. (2024). Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-Bi-LSTM for enhanced reliability. *Ad Hoc Networks*, *155*, 103407. https://doi.org/10.1016/j.adhoc.2024.103407

Canchila, S., Meneses-Eraso, C., Casanoves-Boix, J., Cortés-Pellicer, P., & Castelló-Sirvent, F. (2024). Natural language processing: An overview of models, transformers and applied practices. *Computer Science and Information Systems*, *21*(3), 1097–1145. https://doi.org/10.2298/CSIS230217031C

Chaluvaraj Preethi, B., Vasanthi, R., Sugitha, G., & Ayshwarya Lakshmi, S. (2024). Intrusion detection and secure data storage in the cloud were recommend by a multiscale deep bidirectional gated recurrent neural network. *Expert Systems with Applications*, *255*, 124428. https://doi.org/10.1016/j.eswa.2024.124428

Chourasia, S. R., Yadav, S. K., Sarkar, M., Sharma, P., Sharma, P., Sinha, A., Upadhyay, A. K., Kole, M., & Shukla, S. K. (2024). Cybersecurity Frameworks and Models: Review of the Existing Global Best Practices. *Productivity*, *65*(1), 29–42. https://doi.org/10.32381/PROD.2024.65.01.4

Devendiran, R., & Turukmane, A. V. (2024). Dugat-LSTM: Deep learning based network intrusion detection system using chaotic optimization strategy. *Expert Systems with Applications*, *245*, 123027. https://doi.org/10.1016/j.eswa.2023.123027

El-Shafeiy, E., Elsayed, W. M., Elwahsh, H., Alsabaan, M., Ibrahem, M. I., & Elhady, G. F. (2024). Deep Complex Gated Recurrent Networks-Based IoT Network Intrusion Detection Systems. *Sensors*, *24*(18), 5933. https://doi.org/10.3390/s24185933

Li, M., Han, D., Li, D., Liu, H., & Chang, C.-C. (2022). MFVT: an anomaly traffic detection method merging feature fusion network and vision transformer architecture. *EURASIP Journal on Wireless Communications and Networking*, *2022*(1), 39. https://doi.org/10.1186/s13638-022-02103-9

Lumazine, A., Drakos, G., Salvatore, M., Armand, V., Andros, B., Castiglione, R., & Grigorescu, E. (2024). *Ransomware Detection in Network Traffic Using a Hybrid CNN and Isolation Forest Approach*. https://doi.org/10.22541/au.172901014.44599790/v1

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6). https://doi.org/10.1093/bib/bbac409

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, *13*(9), 5521. https://doi.org/10.3390/app13095521

Muthunambu, N. K., Prabakaran, S., Kavin, B. P., Siruvangur, K. S., Chinnadurai, K., & Ali, J. (2024). A Novel Eccentric Intrusion Detection Model Based on Recurrent Neural Networks with Leveraging LSTM. *Computers, Materials & Continua*, *78*(3), 3089–3127. https://doi.org/10.32604/cmc.2023.043172

Pawlicki, M., Choraś, M., Kozik, R., & Hołubowicz, W. (2020). *On the Impact of Network Data Balancing in Cybersecurity Applications* (pp. 196–210). https://doi.org/10.1007/978-3-030-50423-6_15

Ren, Y., Xiao, Y., Zhou, Y., Zhang, Z., & Tian, Z. (2022). CSKG4APT: A Cybersecurity Knowledge Graph for Advanced Persistent Threat Organization Attribution. *IEEE Transactions on Knowledge and Data Engineering*, 1–15. https://doi.org/10.1109/TKDE.2022.3175719

Sah, A. K., & K, V. (2024). Anomaly-Based Intrusion Detection in Network Traffic using Machine Learning: A Comparative Study of Decision Trees and Random Forests. *2024 2nd International Conference on Networking and Communications (ICNWC)*, 1–7. https://doi.org/10.1109/ICNWC60771.2024.10537451

Sathishkumar, P., Gnanabaskaran, A., Saradha, M., & Gopinath, R. (2024). Dos attack detection using fuzzy temporal deep long Short-Term memory algorithm in wireless sensor network. *Ain Shams Engineering Journal*, *15*(12), 103052. https://doi.org/10.1016/j.asej.2024.103052

Song, B., Wu, Y., & Xu, Y. (2024). ViTCN: Vision Transformer Contrastive Network for Reasoning. *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, 452–456. https://doi.org/10.1109/AINIT61980.2024.10581446

Sontan Adewale Daniel, & Samuel Segun Victor. (2024). EMERGING TRENDS IN CYBERSECURITY FOR CRITICAL INFRASTRUCTURE PROTECTION: A COMPREHENSIVE REVIEW. *Computer Science & IT Research Journal*, *5*(3), 576–593. https://doi.org/10.51594/csitrj.v5i3.872

Talukder, Md. A., Islam, Md. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., & Moni, M. A. (2024). Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *Journal of Big Data*, *11*(1), 33. https://doi.org/10.1186/s40537-024-00886-w

Ullah, F., Ullah, S., Srivastava, G., & Lin, J. C.-W. (2024). IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic. *Digital Communications and Networks*, *10*(1), 190–204. https://doi.org/10.1016/j.dcan.2023.03.008

Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*, *11*(1), 25. https://doi.org/10.1186/s40537-023-00842-0

Zhang, J., Liu, S., & Liu, Z. (2024). Research on APT Malware Detection Based on BERT-Transformer-TextCNN Modeling. *Proceedings of the 2024 International Conference on Generative Artificial Intelligence and Information Security*, 235–242. https://doi.org/10.1145/3665348.3665389

Zhang, Z., Gong, S., Liu, Z., & Chen, D. (2023). A novel hybrid framework based on temporal convolution network and transformer for network traffic prediction. *PLOS ONE*, *18*(9), e0288935. https://doi.org/10.1371/journal.pone.0288935