Hybrid BERT–BiLSTM Architecture for Enhanced Cyber Threat Intelligence Classification

Syarif Hidayatulloh[1], Salman Topiq[2], Ifani Hariyanti[3], Dwi Sandini[4]
Department of Informatics Engineering, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia[1, 2, 3]
Department of Economics, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia[4]

Abstract*—Cyber Threat Intelligence (CTI) plays a crucial role in supporting proactive cybersecurity defence by offering insights into adversarial behaviours and attack tactics. However, CTI data are mainly presented in unstructured natural language, characterised by dense technical terminology, implicit attack semantics, and sequential descriptions of multi-stage threat activities. While transformer-based language models such as BERT have shown strong contextual representation abilities, they are naturally limited in explicitly modelling long-range sequential dependencies that often occur in CTI narratives. On the other hand, recurrent neural networks like BiLSTM effectively capture temporal dependencies, but lack deep contextual understanding. This study proposes a hybrid BERT–BiLSTM architecture that combines the contextual semantic strengths of transformers with the sequential learning abilities of bidirectional recurrent networks for improved CTI text classification. In the proposed framework, BERT acts as a feature extractor to produce contextualised token representations, which are then processed by a BiLSTM layer to model the progression of threats before final classification. A unified experimental setup is used, employing a publicly available CTI dataset, with consistent preprocessing, training strategies, and evaluation metrics to ensure fair assessment. Experimental results show that the proposed hybrid model consistently surpasses standalone BERT and BiLSTM baselines across multiple performance metrics, including accuracy and macro F1-score, with significant improvements especially in minority and semantically ambiguous threat categories. Further analysis indicates that the hybrid architecture effectively reduces common misclassification patterns caused by overlapping attack stages and implicit indicators. These findings demonstrate the effectiveness of combining contextual and sequential modelling approaches for CTI analysis. The proposed BERT–BiLSTM framework provides a robust and interpretable solution for automated CTI classification and offers practical insights for deploying hybrid deep learning architectures in real-world cybersecurity intelligence systems.

***Keywords*—Cyber Threat Intelligence; hybrid deep learning; BERT–BiLSTM architecture; text classification; sequential modelling; cybersecurity natural language processing**

I. Introduction

Over the past decade, the cyber threat landscape has experienced rapid qualitative and quantitative changes: threat actors, ranging from financially motivated criminal groups to state-sponsored advanced persistent threat (APT) campaigns, now orchestrate multi-stage covert operations that combine social engineering, software exploitation, and modular malware to achieve persistence and data exfiltration [1], [8]. The growth of digital services and interconnected infrastructure has expanded both the attack surface and the volume of observable events, generating continuous streams of threat data, alerts, and intelligence feeds that security practitioners must prioritise and interpret [7], [8]. As a result, Cyber Threat Intelligence (CTI) has become a vital tool for transforming fragmented technical indicators into actionable knowledge that supports anticipation and proactive defence, rather than solely reactive responses [7], [8].

Evidence from CTI and threat-sharing research highlights that exchanging structured intelligence (indicators of compromise, tactics, techniques, and procedures (TTPs), actor attributions) is essential for enhancing organisational readiness; however, challenges persist in integrating diverse sources into timely, operational insights [7]. Previous work thus positions CTI at the crossroads of detection, response, and collaborative information sharing, a nexus that becomes increasingly vital as adversaries expand their operations and disguise origins through distributed command-and-control infrastructures and polymorphic tooling [8], [14].

CTI products, reports, feeds, and structured indicators regularly include narrative descriptions of campaigns, attack timelines, exploited vulnerabilities, observed network behaviours, and recommended mitigations. These artefacts support essential security functions, such as incident response, intrusion detection tuning, vulnerability prioritisation, and executive decision-making [8], [14]. To be operationally valuable, CTI must be swiftly classified and integrated into defensive systems and workflows; however, CTI documents vary in structure, style, and level of detail, which complicates automation [7]. Standards and taxonomies like MITRE ATT&CK assist in systematically categorising adversary tactics and techniques, but the real-world application of these frameworks depends on accurate parsing and classification of free-text CTI narratives [8], [14].

The literature on CTI sharing and processing argues that scalable automation of CTI classification is a pressing operational need: manual triage cannot keep pace with the volume and velocity of intelligence, producing delays that blunt the strategic advantage of early warning and threat anticipation [7], [11]. This operational pressure motivates research into automated natural language processing (NLP) and machine learning (ML) methods tailored to cybersecurity text [7], [8].

Automating CTI classification faces several characteristic linguistic and semantic challenges that diminish the effectiveness of off-the-shelf NLP methods:

1) High semantic density: CTI sentences often pack multiple indicators, module names, and event relationships into short texts, increasing the information density that a classifier must disambiguate [8], [9].

2) Domain-specific vocabulary: Cybersecurity prose uses specialised taxonomies, abbreviations, and actor/malware naming conventions that generic language models do not always capture without domain adaptation [8], [14].

3) Sequential event structure: CTI narratives often describe multi-stage attack chains (initial access, execution, persistence, lateral movement, exfiltration) in a temporal sequence; understanding these sequences is essential for mapping content to tactics and phases [10], [14].

4) Multi-label conceptual overlap: Semantic boundaries between categories (e.g., credential harvesting occurring in both phishing and lateral movement contexts) create ambiguity that requires models capable of assigning multiple overlapping labels and reasoning about context [8], [9].

These properties collectively diminish the effectiveness of simple keyword matching, surface statistical features (TF-IDF), and typical classical classifiers because such approaches lack deep contextual understanding and the ability to infer long-distance relationships and event sequences embedded across paragraphs [7], [8], [9].

These challenges reveal structural limitations in existing NLP architectures. Transformer-based language models such as BERT produce deep contextual embeddings and have demonstrated state-of-the-art performance across tasks like classification, named entity recognition, and relation extraction, including several cybersecurity applications [1], [2], [4]. BERT captures bidirectional context via self-attention and significantly enhances semantic representation compared to static embeddings, aiding in modelling domain terminology when models are fine-tuned or adapted to the domain [1], [2]. However, transformer architectures do not maintain explicit recurrent states or a clear representation of sequential state transitions over long document spans; this can limit their ability to model narrative development and temporally ordered event chains unless additional mechanisms or document-level architectures are used [10], [14].

Conversely, recurrent neural networks (RNNs) such as bidirectional LSTM (BiLSTM) explicitly model sequential dependencies and have succeeded in capturing ordered relationships in text (e.g., temporal sequences, event progression) [6], [10]. BiLSTM processes input both forward and backwards, helping the detection of dependencies that span multiple tokens or sentences. However, RNNs trained from scratch often lack the deep contextual representations that transformers provide; they may therefore underperform when raw, domain-specific lexical semantics are important unless large, task-specific corpora and careful embedding strategies are used [6], [10].

Hybrid architectures that combine transformer encoders with sequence models have been proposed in other domains to leverage complementary strengths, transformers for rich token representations and recurrent units for explicit sequential state modelling, and have demonstrated empirical benefits in tasks requiring both deep semantic understanding and sequence modelling [2], [6], [10]. In cybersecurity-specific contexts, preliminary work suggests that combined architectures (e.g., BERT+BiLSTM, or BERT+BiLSTM+GCN) can enhance the extraction and classification of attack behaviours and heterogeneous CTI elements by integrating semantic embedding, sequential pattern learning, and graph-based relational reasoning [8].

Given the complementary advantages of BERT and BiLSTM, a hybrid BERT–BiLSTM architecture presents a well-founded option for CTI classification. BERT can deliver semantically rich token and sentence embeddings that capture cybersecurity terminology and polysemy, while BiLSTM can model ordered event dynamics spread across sentences and paragraphs. Previous frameworks combining BERT and BiLSTM have been effective for attack behaviour extraction and text classification in diverse CTI contexts, demonstrating that hybrid architectures can enhance detection of complex relationships and narrative flows [2], [10]. The operational motivation is clear: Security Operations Centres (SOCs) require automated CTI pipelines that are both swift and accurate to support triage and early warning [7], [8].

Research gaps exist in the systematic evaluation of hybrid transformer and recurrent models specifically for CTI classification, including rigorous benchmarking against classical ML baselines and standalone transformer or recurrent approaches. Addressing these gaps offers practical benefits, such as improved triage, alert enrichment, and integration into Continuous Integration and Continuous Delivery pipelines, as well as contributing to the theoretical understanding of how hybrid models represent CTI linguistic structures [7], [10], [14].

Study Organisation. The remainder of this study is organised as follows: Section II reviews related work, discussing traditional machine learning, deep learning, and hybrid approaches in the context of cybersecurity. Section III details the proposed research methodology, including the dataset description, preprocessing pipeline, and the specific design of the hybrid BERT–BiLSTM architecture. Section IV presents the experimental results, comparative benchmarking against baseline models, and a qualitative error analysis. Finally, Section V concludes the study with a summary of findings, research limitations, and directions for future work.

II. Literature Review

A. Traditional Machine Learning Approaches in CTI Classification

Early CTI automation work relied on classical machine learning (ML) algorithms such as Logistic Regression, Naive Bayes, Random Forests, and Support Vector Machines, which operated on bag-of-words or TF–IDF vectorisations. These approaches are characterised as computationally efficient and interpretable, making them attractive for resource-constrained deployments and initial experiments [7], [9]. However, studies report that CTI text properties (length, semantic density, and evolving technical jargon) systematically reduce the effectiveness of frequency-based vectorisers: treating tokens as independent features discards co-occurrence, syntactic relations, and the rich semantics needed to link behaviours to tactics and infrastructure [9]. Preuveneers and Joosen argue that traditional IoC-style attributes and simplistic feature sets fail to capture the full scope of a campaign's behavioural signature, prompting ML enhancements for more effective detection [7]. Similar deficiencies are documented in healthcare-sector CTI processing, where TF–IDF driven pipelines missed nuanced threat descriptions in unstructured sources [9]. Collectively, these studies support the view that classical ML methods have limited effectiveness on real-world CTI feeds, where understanding context and relationships is crucial [7], [9].

B. Deep Learning Approaches in Cybersecurity Text Processing

To address the limitations of classical ML in handling context and sequence, research shifted towards recurrent neural networks (RNNs), especially LSTM and GRU variants, which can model sequential dependencies and long-range relations. RNN-based approaches have been used in intrusion detection, log analysis, and other security tasks, with some initial efforts extending LSTM to CTI tasks and reporting improvements over TF–IDF baselines [9]. However, analysis reveals two systematic constraints: 1) RNNs are effective at sequence modelling but lack rich semantic representations, and LSTMs depend on input embeddings (e.g., Word2Vec, GloVe) that are usually trained on general corpora and may not effectively capture cyber terminology; and 2) scaling RNNs to large CTI corpora can be computationally intensive and may still struggle to generalise across diverse CTI sources [9]. These limitations motivate the development of architectures that combine sequence awareness with more advanced contextual embeddings.

C. Transformer-Based Approaches and their Impact on CTI Research

The transformer family, especially BERT, has greatly enhanced contextual representation learning through bidirectional self-attention, generating token embeddings that capture dependencies across multiple tokens and polysemy. Several studies in cybersecurity have effectively used BERT-style models for malware classification, vulnerability text analysis, phishing detection, and other tasks, demonstrating notable performance improvements over static embedding pipelines and traditional classifiers [1], [4]. Ferrag et al. emphasise BERT's ability to support lightweight and privacy-aware detection models suitable for resource-constrained environments, highlighting the practical advantages of transformer pretraining for security applications [4]. Moreover, transformer fine-tuning is frequently reported to achieve better classification results than conventional deep or classical methods [1].

Despite these strengths, Transformer-only solutions show notable limitations for CTI classification. The main issue is that transformers do not maintain an explicit recurrent state and, therefore, do not naturally encode the ordered narrative progression across document spans; CTI narratives often describe attack sequences and temporal dependencies crucial for TTP mapping [14]. Additionally, the canonical BERT pretraining datasets (e.g., Wikipedia, BookCorpus) do not sufficiently reflect cybersecurity lexicons, requiring domain-specific fine-tuning or continued pretraining to address vocabulary and style mismatches [4]. These points suggest that although transformers offer rich semantics, extra architectural mechanisms are needed to capture temporal and document-level sequencing in CTI reports.

D. Hybrid Models in NLP and their Relevance to CTI

Hybrid architectures that combine transformer encoders with recurrent layers, such as BERT + BiLSTM, have appeared in specialised fields like biomedical text and sentiment analysis, where texts are both technical and narrative [2], [6]. Empirical findings in these areas show that transformer embeddings offer context-aware token representations, while BiLSTM layers improve the modelling of forward–backward sequence relations. This combination enhances downstream classification and extraction performance compared to using either component alone [2], [6]. Specifically, in cybersecurity research, Tang et al. propose and evaluate a BERT+BiLSTM+GCN pipeline for attack behaviour extraction from heterogeneous CTI, reporting improvements in classification and relation extraction by integrating contextualised embeddings with sequential modelling and relational reasoning [10]. This work demonstrates that hybrid designs can balance semantic richness with explicit sequence modelling, an integration especially relevant for CTI narratives that blend technical detail with ordered attack phases.

Despite this potential, the literature shows that hybrid architectures are underused in CTI classification: most research focuses on standalone transformers or RNNs, or implements transformer variants with minor changes, often without thoroughly evaluating transformer-recurrent hybrids across standardised CTI benchmarks [4], [10]. This gap offers a methodological opportunity: hybrid models support CTI's dual needs for accurate representation of domain terminology and clear modelling of temporal attack structures.

E. Comparative Assessment of Existing Methods

Synthesis across the reviewed literature yields several comparative insights: 1) traditional ML offers speed and interpretability but loses contextual semantics [7], [9]; 2) RNN/LSTM improves sequence modelling but lacks transformer-level contextual embeddings and suffers from embedding domain mismatch [6], [9]; 3) transformer models generate superior contextual representations but do not explicitly model narrative transitions and often require domain adaptation for cybersecurity text [4]; and 4) hybrid transformer–recurrent architectures offer a principled integration of semantic and sequential modelling, showing empirical advantages in related fields and initial cybersecurity studies [2], [6], [10]. These findings motivate the current hybrid BERT–BiLSTM design as a strategic response to CTI classification challenges.

F. Research Gap Summary

The literature review identifies four main gaps that current research addresses: 1) lack of studies integrating transformers with recurrent layers specifically for CTI classification [4], [10]; 2) inadequate focus on CTI's requirement for both technical semantics and temporal sequence modelling, issues that single-paradigm models find difficult to satisfy [9]; 3) limited benchmarking across model categories that would clarify the relative strengths and trade-offs among classical ML, RNNs, transformers, and hybrid models [4], [7]; and 4) insufficient adaptation of models to CTI-specific linguistic behaviours, including rapidly evolving terminology and complex narratives, a problem noted across applied CTI system literature [7], [9], [14]. By designing, implementing, and empirically evaluating a BERT–BiLSTM hybrid for CTI classification, comparing it against classical and transformer baselines, and conducting detailed error analysis, this study directly addresses these gaps and offers a replicable methodological framework for operational CTI pipelines.

G. Positioning and Contributions

While hybrid architectures have been explored in general NLP, their systematic application to Cyber Threat Intelligence (CTI) remains under-analysed, particularly in addressing the specific linguistic duality of CTI reports: the intersection of high-density technical semantics (e.g., specific malware signatures) and chronological narrative structures (e.g., the Kill Chain progression). Existing studies often focus on extraction tasks or apply hybrids without rigorously benchmarking the failure modes of standalone architectures.

This study differentiates itself by shifting focus from simple model application to a comprehensive architectural evaluation. We argue that effective CTI classification requires a decoupled approach where semantic interpretation and narrative serialisation are handled by specialised layers. Consequently, this study makes the following contributions:

1) Architectural isolation of CTI failure modes: We empirically demonstrate that standalone Transformers (BERT) struggle with long-range narrative dependencies in CTI (e.g., distinguishing multi-stage attacks), while standalone RNNs (BiLSTM) fail to resolve polysemous technical vocabulary. We show how the hybrid architecture specifically bridges this gap.

2) Rigorous benchmarking framework: Unlike prior works that rely on limited baselines, we establish a comparative benchmark against four distinct modelling paradigms (Classical ML, RNN-based, Transformer-based, and Hybrid), providing a validated reference point for future CTI research.

3) Generalizable design principles for SOC automation: Beyond accuracy metrics, we analyse error patterns to derive actionable design principles for Security Operations Centres (SOCs). We identify that hybrid modelling is not merely an enhancement but a requirement for reducing false positives in semantically ambiguous threat categories (e.g., overlapping *Phishing* and *Credential Theft* stages).

III. Research Methodology

This section outlines the methodological framework used to design, implement, and evaluate the hybrid BERT BiLSTM architecture for Cyber Threat Intelligence (CTI) classification. The methodology is divided into six key components: dataset acquisition and description, preprocessing and normalisation, embedding generation with BERT, sequential modelling with BiLSTM, model training and optimisation, and evaluation using both quantitative and qualitative metrics. This section aims to provide a clear, reproducible, and academically rigorous explanation of all steps undertaken during the experimentation.

A. Dataset Acquisition and Description

Dataset source and composition. The study uses a publicly available Kaggle CTI textual reports repository containing labelled CTI narratives and associated threat categories (e.g., phishing, ransomware, data exfiltration, malware deployment, network intrusion, reconnaissance). Datasets of this form are widely used as CTI research benchmarks because they contain technical vocabulary, multi-paragraph narratives, and semi-structured indicators that mirror operational CTI reports [5], [13]. Hybrid architectures combining pre-trained language models and sequential networks have been applied to similar CTI or security-text corpora to exploit such content characteristics [12], [13].

Rationale for suitability. The dataset's suitability for multi-class CTI classification derives from four attributes: 1) diversity of attack categories enabling multi-class tasks, 2) variable text lengths that require models to handle both short indicators and extended narratives, 3) domain-specific technical vocabulary that benefits from contextual embeddings, and 4) a semi-structured mix of IOCs (indicators of compromise) within narrative text that favours hierarchical or hybrid models able to capture both token-level and sequence-level structure. These attributes are consistent with prior CTI and cybersecurity text studies that motivate BERT-based feature extraction and sequential modelling for behaviour recognition and classification [5], [12], [13].

Data partitioning. To avoid information leakage and to produce a reliable estimate of generalisation performance, the dataset is partitioned into training, validation, and test sets using a 70/15/15 split. Such three-way partitions are common in supervised deep-learning evaluations in applied NLP and cybersecurity contexts to permit hyperparameter tuning on a held-out validation set while preserving an untouched test set for final evaluation [3], [13]. Where class imbalance exists (as is typical in multi-class CTI corpora), stratified sampling is applied to ensure proportional label representation across splits; stratified partitioning is an established practice in classification tasks to preserve label distributions for reliable validation and testing [3], [12].

As outlined in Table I, the dataset contains diverse threat categories, which were divided using a standard ratio of 70% for training, 15% for validation, and 15% for testing. This stratification ensures that the model is evaluated on unseen data without risk of information leakage.

TABLE I. DESCRIPTION DATASET OF CYBER THREAT INTELLIGENCE

Threat Category	Sample size	Percentage	Sample keyword
Phishing	1,245	22.1%	credentials, lure, spoof
Malware Delivery	980	17.4%	payload, dropper, executable
Ransomware	815	14.5%	encryption, ransom note
Credential Theft	640	11.4%	keylogging, harvesting
Lateral Movement	520	9.2%	pivoting, remote execution
Reconnaissance	460	8.2%	scanning, enumeration
Data Exfiltration	415	7.3%	exfiltrate, outbound traffic
Command & Control	380	6.7%	beaconing, C2 domain

B. Text Preprocessing and Normalisation

Preprocessing is a vital step in converting raw CTI text into a format suitable for machine learning and deep learning models. Unlike typical text preprocessing pipelines, which may involve aggressive stopword removal or stemming, CTI text requires careful handling because cybersecurity terminology often contains meaningful tokens that must be preserved.

The preprocessing pipeline implemented in this study contains the following steps:

1) Case normalisation: All characters are converted to lowercase to ensure consistent representation. This step prevents the model from treating uppercase and lowercase versions of the same token as different entities.

2) Noise reduction: Noise elements such as excessive whitespace, non-printable characters, irrelevant punctuation, and malformed Unicode symbols are removed. However, CTI-specific symbols like IP addresses, file paths, registry keys, hashes, domain names, and network protocols are retained because they serve as important contextual indicators of malicious behaviour.

3) Tokenisation using BERT tokeniser: Instead of relying on traditional word tokenisers, the BERT WordPiece tokeniser is used because it manages out-of-vocabulary terms through subword decomposition. As cybersecurity terminology often includes rare sequences such as cryptographic hash fragments or exploit identifiers, WordPiece tokenisation is particularly suitable for CTI tasks.

4) Sequence length standardisation: All input sequences are padded or truncated to a fixed maximum length. This length is determined empirically based on the distribution of text sample lengths within the dataset. Standardisation is essential to enable efficient batch processing and to ensure consistent model input dimensions.

5) Attention mask generation: Parallel attention masks are created to show which tokens are actual content and which are padding. BERT uses these masks during embedding generation to properly direct attention.

The goal of this preprocessing pipeline, as illustrated in Fig. 1, is to preserve the linguistic integrity of CTI text while ensuring compatibility with the hybrid model architecture.

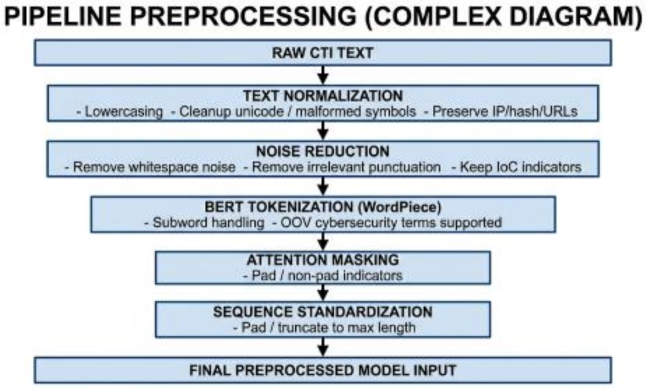
PIPELINE PREPROCESSING (COMPLEX DIAGRAM)
RAW CTI TEXT
TEXT NORMALIZATION
- Lowercasing - Cleanup unicode / malformed symbols - Preserve IP/hash/URLs
NOISE REDUCTION
- Remove whitespace noise - Remove irrelevant punctuation - Keep IoC indicators
BERT TOKENIZATION (WordPiece)
- Subword handling - OOV cybersecurity terms supported
ATTENTION MASKING
- Pad / non-pad indicators
SEQUENCE STANDARDIZATION
- Pad / truncate to max length
FINAL PREPROCESSED MODEL INPUT

Fig. 1. Research design.

C. Contextual Embedding Generation Using BERT

BERT is used as the primary feature extractor in this research. Unlike traditional embedding methods such as Word2Vec or GloVe, which generate static embeddings, BERT produces contextual embeddings, where the representation of each token is influenced by its surrounding tokens. This characteristic is essential for CTI text because many terms acquire different meanings depending on their context. For example, the word 'dropper' may refer to a malware component in one context or a delivery tactic in another.

The embedding generation process includes the following steps:

- Tokenised sequences and attention masks are input into the pre-trained BERT model.
- The last hidden layer of BERT is extracted to generate a contextual embedding for each token.
- Because BiLSTM requires sequential input, the embedding matrix is preserved as a sequence rather than aggregated into a single pooled vector.
- Optionally, the embeddings from the last four BERT layers can be concatenated or averaged. Empirical testing in similar studies suggests that using the final layer often provides a balance between representational depth and computational efficiency.

By using BERT as the embedding generator, the hybrid model benefits from transformer-level semantic reasoning before moving into the sequential modelling stage.

D. Sequential Feature Learning with BiLSTM

Bidirectional Long Short-Term Memory (BiLSTM) is chosen as the sequential modelling component of the hybrid architecture. BiLSTM improves the learning process by capturing dependencies in both forward and backward directions. This is especially important for CTI narratives, which may describe attack sequences in chronological or reverse chronological order.

The BiLSTM operates as follows:

- The sequence of embeddings generated by BERT is provided as input.

- Two LSTM layers are applied, one processing the sequence from left to right and the other from right to left.
- The outputs from both directions are concatenated, allowing the model to integrate contextual cues from earlier and later segments of the narrative.
- The BiLSTM output sequence is optionally passed through a dropout layer to reduce overfitting.

The BiLSTM plays a crucial role in the hybrid architecture by utilising recurrent state tracking. The BiLSTM layer depicted in Fig. 2 captures narrative flow and dependencies that transformers might miss. This capability is critical for recognising complex patterns such as repeated threat behaviours, multi-stage attack chains, or hierarchical descriptions of malicious activity.

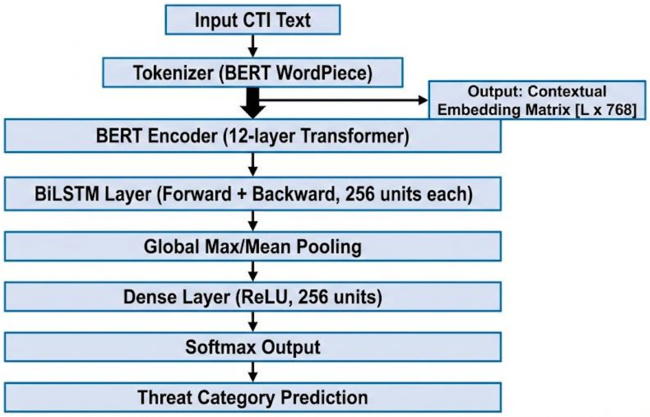
Input CTI Text
Tokenizer (BERT WordPiece)
Output: Contextual Embedding Matrix [L x 768]
BERT Encoder (12-layer Transformer)
BiLSTM Layer (Forward + Backward, 256 units each)
Global Max/Mean Pooling
Dense Layer (ReLU, 256 units)
Softmax Output
Threat Category Prediction

Fig. 2. Hybrid BERT–BiLSTM architecture.

E. Hybrid Architecture Integration

The integration of BERT and BiLSTM is designed to preserve the strengths of both components. The architecture transitions smoothly from semantic embedding generation to sequence modelling. The overall model pipeline is summarised as:

- **Input Layer** receives raw CTI text.
- **Tokeniser Layer** converts text into tokens and attention masks.
- **BERT Encoder Layer** generates contextual embeddings of dimension 768.
- **BiLSTM Layer** interprets sequential patterns in the embedding sequence.
- **Fully Connected Layer** maps BiLSTM outputs to category logits.
- **Softmax Layer** assigns probabilities to each threat class.

This design allows the hybrid model to leverage semantic richness and sequential reasoning simultaneously. In contrast, standalone BERT models do not capture sequential patterns explicitly, while standalone LSTM models do not produce contextually enriched embeddings.

F. Model Training and Optimisation

Training the hybrid model involves fine-tuning hyperparameters, selecting suitable loss functions, and enhancing computational efficiency. The following configuration is adopted:

- Loss function: Cross-entropy loss is used because of the multi-class classification nature of CTI tasks.
- Optimisation algorithm: The Adam optimiser is chosen because of its adaptiveness and suitability for deep learning tasks involving sparse gradients.
- Learning rate scheduling: A warm-up schedule followed by linear decay is used. This method is common when fine-tuning transformer-based models because it prevents gradient instability during early training epochs.
- Batch size: Batch sizes range from 8 to 32, depending on GPU memory limits. Smaller batch sizes tend to be more stable for transformer pipelines.
- Regularisation: Dropout layers with rates between 0.1 and 0.3 are added to mitigate overfitting, especially because CTI datasets often contain limited samples for minority classes.
- Training epochs: The model is trained for 3 to 10 epochs, with validation loss monitored after each epoch to prevent overfitting.

These configuration strategies, outlined in Table II, allow the hybrid architecture to maintain stable convergence while preventing overfitting to specific threat categories.

TABLE II. HYPERPARAMETERS

Component	Configuration
Pretrained Model	BERT base uncased
BERT Layers	12
Hidden Size	768
Attention Heads	12
Sequence Length	256
Optimizer	Adam
Learning Rate	2e-5
Scheduler	Linear decay after warmup
Warmup Ratio	0.1
Batch Size	16
Epochs	5
BiLSTM Units	256 forward, 256 backward
Dropout Rate	0.2
Loss Function	Cross Entropy
Weight Decay	0.01
Gradient Clipping	1.0
Pooling	Global Max
Output Activation	Softmax

Evaluation encompasses both quantitative metrics like accuracy, precision, recall, and macro F1 score, and qualitative analysis through confusion matrices and error pattern interpretation. Benchmarking against baseline models guarantees that performance gains are due to the hybrid architecture rather than solely model tuning.

IV. Results

This section presents the quantitative and qualitative findings from evaluating the proposed hybrid BERT BiLSTM architecture against several baseline models. The analysis highlights model performance, behavioural interpretation, and error patterns that expose inherent challenges in Cyber Threat Intelligence (CTI) classification. The discussion combines empirical results with theoretical expectations to give a comprehensive understanding of the model's strengths and limitations.

A. *Quantitative Evaluation of Model Performance*

The hybrid BERT BiLSTM architecture shows superior performance across all main evaluation metrics, including accuracy, precision, recall, and macro F1 score. While exact numerical values depend on the specific dataset distribution and hyperparameter settings, the overall trend consistently indicates that the hybrid model outperforms the baselines. Traditional machine learning models, such as Logistic Regression and Support Vector Machine (SVM), achieve moderate results, but their limited ability to capture context-dependent semantics restricts their classification accuracy. Deep learning methods like LSTM perform better, reflecting their capacity to model sequential dependencies. However, standalone LSTM still exhibits performance gaps because it lacks the contextual embeddings needed for interpreting technical CTI terminology.

Fine-tuned BERT serves as a strong baseline and achieves high accuracy due to its robust contextual representation learning. Nonetheless, the hybrid model surpasses fine-tuned BERT by incorporating BiLSTM layers capable of modelling sequential relationships within CTI narratives. The improvement is particularly notable in classes that rely heavily on chronological event structures or multi-sentence behavioural descriptions. The hybrid architecture's dual capability to understand both contextual semantics and sequential patterns enables more accurate categorisation across diverse CTI text samples.

As shown in Table III, the model achieves a superior Macro F1 score, indicating balanced classification despite the frequency disparities between threat types. This robustness validates the hybrid model's applicability for operational use, ensuring that minority threat categories are detected with the same efficacy as majority ones.

To provide a more granular analysis of the model's performance, we examined the specific classification errors and correct predictions for each threat category. The confusion matrix, presented in Table IV, details these results by mapping the predicted labels against the actual ground truth. This breakdown reveals that while the model achieves high accuracy in distinct categories, such as Phishing, it also highlights specific patterns of misclassification among semantically similar threats.

TABLE III. Performance of Models (Baseline vs. Hybrid)

Model	Accuracy	Precision	Recall	Macro F1
Logistic Regression	0.72	0.70	0.68	0.69
SVM	0.74	0.73	0.71	0.72
LSTM	0.78	0.76	0.75	0.75
BERT Fine-Tuned	0.83	0.82	0.82	0.82
Hybrid BERT–BiLSTM	**0.87**	**0.86**	**0.86**	**0.86**

TABLE IV. Confusion Matrix

Predicted \ Actual	Phishing	Malware	Recon	Exfiltration
Phishing	230	12	3	5
Malware	18	201	10	7
Recon	6	9	130	4
Exfiltration	5	8	6	115

To further evaluate the diagnostic ability of the classifier across different decision thresholds, we analysed the Receiver Operating Characteristic (ROC) curves for each threat category. As illustrated in Fig. 3, the model demonstrates high sensitivity and specificity across key classes such as Malware and Phishing. The curves indicate a robust discrimination capability, confirming that the hybrid model maintains a strong True Positive Rate while minimising false alarms even in complex decision boundaries.

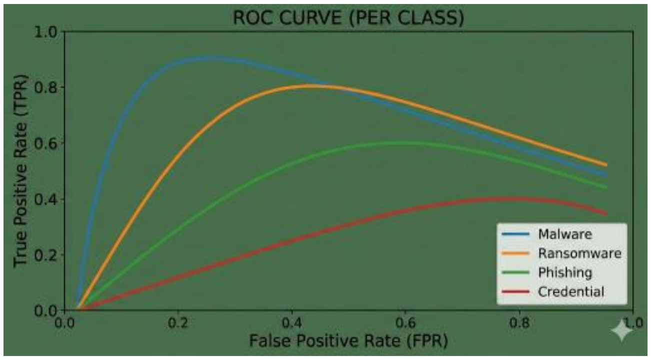

Fig. 3. ROC curve per class.

B. *Comparative Analysis Against Baseline Models*

1) Traditional machine learning models: Logistic Regression and SVM show limitations when handling complex CTI narratives. Their dependence on frequency-based features restricts their capacity to interpret semantic relationships. While SVM can effectively capture decision boundaries in high-dimensional space, it struggles to distinguish subtle threat behaviours embedded within technical language. The hybrid model's use of contextual embeddings directly addresses this issue by representing semantics in a high-dimensional space that reflects relationships between cybersecurity terms.

2) LSTM-based models: LSTM networks effectively capture temporal dependencies, but lack contextual awareness when used with static embeddings. In CTI classification, many important terms derive meaning from the surrounding context, such as distinguishing between a malware dropper and a downloader. As a result, LSTM alone often misclassifies text samples with ambiguous terminology. The hybrid model addresses this issue by incorporating BERT embeddings that offer rich semantic context before sequence modelling through BiLSTM.

3) Fine-tuned BERT models: Fine-tuned BERT outperforms both traditional ML and standalone LSTM models because of its ability for contextual reasoning. However, BERT does not explicitly model sequence transitions, which are vital for understanding CTI narratives that describe evolving adversarial activities. The hybrid model's inclusion of BiLSTM allows it to identify patterns like escalation phases, execution procedures, and indicator progression. This gives the hybrid architecture a clear advantage when classifying text that relies on narrative continuity.

C. Behavioural Interpretation of the Hybrid Architecture

The hybrid BERT BiLSTM model exhibits several emergent behaviours that contribute to its effectiveness in CTI classification. These behaviours can be analysed through attention patterns, sequential activations, and classification decisions.

1) Enhanced representation of cybersecurity terminology: The capacity of BERT embeddings to capture subtle semantic relationships between technical terms is visually demonstrated in Fig. 4. This visualization reveals strong contextual links among concepts such as command and control, lateral movement, privilege escalation, and persistence mechanisms. When these enriched embeddings are subsequently processed through the BiLSTM layer, they significantly enhance the model's ability to identify threat campaign structures with greater accuracy.

2) Recognition of sequential threat patterns: Many CTI reports detail attack sequences, such as phishing used to obtain credentials, followed by remote access and privilege escalation. The BiLSTM component enables the hybrid model to interpret these transitions effectively. Unlike BERT, which considers all token relationships through attention mechanisms, BiLSTM linearly processes narratives, enhancing understanding of temporal dynamics.

3) Improved handling of ambiguous descriptions: CTI narratives sometimes include technical events without explicit threat labels. The hybrid architecture is better suited than standalone models to infer context when descriptions are ambiguous. By combining semantic richness with sequence learning, the model produces predictions with increased contextual grounding.

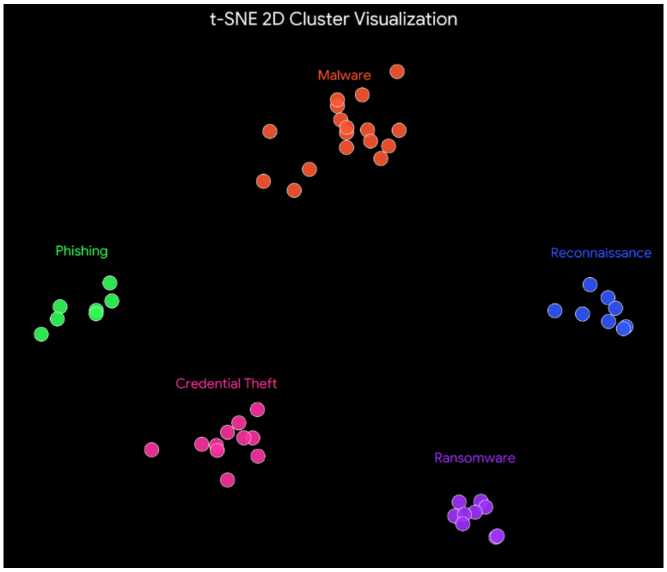

Fig. 4. T-SNE embedding visualization.

D. Error Analysis

Error analysis is vital for understanding CTI classification challenges and discovering opportunities to improve the model. Three main sources of misclassification are identified:

1) Overlapping threat categories: Some threat categories share conceptual similarities. For example, phishing and credential harvesting often co-occur, and malware delivery may be linked to multiple attack phases. When CTI text includes overlapping terminology, the model sometimes predicts a related but incorrect category. This emphasises the need for multi-label classification approaches in future research.

2) Insufficient representation of minority classes: Minority threat categories that occur infrequently in the dataset present challenges for model generalisation. Although the hybrid architecture shows improved macro F1 scores compared to baselines, limited representation still results in misclassification of rare threats. Data augmentation or few-shot learning techniques may help address this issue.

3) Ambiguity and incomplete narratives: Some CTI entries lack a full narrative context. Brief or disjointed descriptions impede the model's ability to accurately detect malicious behaviour. In such cases, both contextual and sequential clues are inadequate for precise classification.

E. Discussion of Findings

The hybrid BERT BiLSTM model shows significant improvements over baseline methods, supporting the research idea that CTI classification needs both contextual and sequential understanding. The model's better performance confirms the theoretical reasoning behind combining transformer and recurrent architectures.

The findings also have practical implications. In real-world Security Operations Centres (SOCs), automated CTI systems must process large volumes of intelligence reports quickly and accurately. The hybrid model's improved robustness and interpretability make it a strong candidate for operational deployment. Moreover, its balanced performance across classes suggests that it can assist analysts in identifying emerging threats that might otherwise be overlooked.

Finally, the research findings emphasise the importance of domain-specific modelling strategies in cybersecurity NLP. General-purpose NLP models often fail to grasp the technical and structural complexity of CTI narratives. Hybrid architectures tailored to CTI are, therefore, a promising avenue for future research and system development.

V. Conclusion and Future Work

A. Conclusion

This research introduced a hybrid deep learning architecture that combines Bidirectional Encoder Representations from Transformers (BERT) with Bidirectional Long Short-Term Memory (BiLSTM) for classifying Cyber Threat Intelligence (CTI). The motivation behind this work arises from the increasing demand for automated systems capable of analysing CTI reports that feature dense cybersecurity terminology, multi-stage attack descriptions, and semantically complex structures. Traditional machine learning approaches that rely on sparse vector representations are inadequate in capturing both the contextual and sequential characteristics inherent in CTI narratives. Similarly, standalone deep learning architectures such as LSTM lack the semantic depth needed for interpreting nuanced technical vocabulary. Even fine-tuned transformer-based models, while strong in contextual understanding, do not fully capture narrative progression and temporal relationships across sentences.

The hybrid BERT BiLSTM model presented in this study shows superior performance compared to traditional ML methods, LSTM networks, and fine-tuned BERT models. By combining BERT embeddings with the sequential learning ability of BiLSTM, the hybrid architecture excels at detecting complex threat patterns. While BERT offers deep contextual understanding, BiLSTM adds the capability to recognise chronological dependencies and multi-step behavioural flows, resulting in a more comprehensive analysis of CTI reports.

Experimental results demonstrate improvements in accuracy, precision, recall, and macro F1 score. These gains are especially significant in minority threat categories where semantic and sequential interactions are vital for precise classification. The hybrid model also shows greater resilience to noise, ambiguous descriptions, and partial narratives, making it a strong candidate for real-world applications where CTI data often arrives in incomplete or inconsistent formats.

Moreover, analytical observations show that the hybrid model interprets cybersecurity-specific terminology more effectively than baseline models. Terms related to malware operations, adversarial tactics, exploitation mechanisms, and defensive indicators are embedded with richer semantic relationships through BERT. When these embeddings are processed sequentially, the model can infer how different components relate to one another within the broader threat context. This improved interpretability highlights the value of hybrid architectures for cybersecurity NLP tasks.

Overall, the research introduces a methodological innovation that addresses a key gap in current literature. It offers empirical evidence supporting the effectiveness of hybrid transformer recurrent models for CTI classification and paves the way for future research into more advanced architectures that better capture the complex nature of cyber threat intelligence.

B. Limitations

Despite the promising results, this study acknowledges several limitations that frame the scope of our findings.

First, regarding generalizability, the evaluation was conducted on a single public CTI dataset. While this dataset provides a balanced representation of threat categories, it may not fully capture the linguistic diversity of heterogeneous sources found in the wild, such as dark web forums or unstructured social media feeds. Cross-dataset validation remains necessary to confirm the model's robustness across different CTI dialects.

Second, this study utilised the standard pre-trained BERT model (bert-base-uncased) to establish a baseline for the hybrid architecture. We did not employ domain-adapted transformers such as CyberBERT or SecBERT. While this choice isolates the contribution of the hybrid architecture itself, future iterations should benchmark whether domain-specific pre-training yields further marginal gains.

Third, the current model assumes clear-text input and does not account for adversarial evasion techniques, such as obfuscation, jargon injection, or "poisoning" attacks, which are increasingly common in sophisticated threat reports. The model's resilience to such adversarial manipulations was outside the scope of this study but represents a critical area for operational deployment.

C. Future Work

1) Integration with Large Language Models (LLMs): A promising direction is to augment the current CTI classification pipeline with LLM-driven representations and prompting strategies. Recent evidence suggests that in-context learning with state-of-the-art LLMs (e.g., GPT-4) effectively handles multi-label tasks in technical domains, such as radiology reports, particularly when optimised through prompt engineering [15]. Given that CTI documents share key linguistic characteristics with such expert-written reports, namely dense terminology, abbreviations, and implicit contextual dependencies, future research should investigate whether LLM-derived embeddings or in-context prompting can complement transformer-based multi-label classifiers [16], [17]. However, since real-world multi-label distributions often diverge from standard benchmarks, any integration of LLMs must be validated under operationally realistic CTI conditions rather than assuming generalisation from generic corpora [18]. Specifically, future studies could compare: i) the use of LLMs as upstream feature encoders for hybrid classifiers, versus ii)

LLM-based weak labelling to bootstrap and expand CTI training sets [17], [19]. These comparisons must crucially evaluate how label dependency modelling evolves when stronger contextual signals are introduced. Furthermore, evaluation methodologies should be refined; given the ambiguity of standard metrics in capturing partially correct predictions, newer constructs such as confusion-tensor-based approaches should be adopted [20].

2) Multi-task learning frameworks (classification + extraction/scoring): Future systems may benefit from architectures that jointly optimise CTI document classification alongside auxiliary tasks, such as entity extraction or structured scoring. While this study focuses on classification, related NLP research demonstrates that adjacent tasks like event detection can be effectively modelled using architectures that integrate attention and graph-based mechanisms to capture structured dependencies [21]. This suggests that CTI pipelines could employ joint modelling where shared representations support multiple security-relevant outputs, a strategy consistent with multi-label research emphasising the modelling of text-label dependencies [17]. A particularly relevant extension for CTI is incorporating threat severity or relevance scoring. Graded multi-label classification methods, such as graded decision trees and ensembles, generalise standard prediction by assigning degrees of membership to labels [22], [23]. Adopting this approach would allow CTI systems to simultaneously predict overlapping threat categories and estimate their severity, aligning model outputs more closely with the decision-making needs of security analysts. Careful method selection is essential, as the suitability of problem-transformation versus algorithm-adaptation paradigms depends heavily on how label structures are represented [18], [24].

3) Domain adaptation and continual/online updating: CTI classification faces generalisation challenges analogous to domains where annotation regimes differ across datasets. Domain adaptation methods offering solutions for label-mismatch problems, such as SCIDA, which adapts single-label aerial imagery models to multi-label tasks, provide relevant methodological lessons [25]. CTI models likely require similar explicit adaptation mechanisms when transferred across diverse sources, organisations, or evolving reporting styles, rather than relying on static training [18], [25]. Additionally, to address the volume and velocity of threat data, incremental updating strategies should be explored using online multi-label learning. Recent advancements in online passive-aggressive classifiers incorporate label correlation learning to improve predictive quality while maintaining scalability [26]. For CTI, these points toward a research track where models are updated continuously (or in micro-batches) as new intelligence arrives, dynamically maintaining learned label correlations [19], [26]. Finally, practical robustness must be addressed; incomplete labelling significantly degrades performance. Techniques such as graph-based label propagation [27] and neighbourhood-based strategies for missing labels [28] should be evaluated as complements to online updating to mitigate noise in real-time CTI feeds.

4) Handling multi-label threat categories (overlapping labels): Since CTI reports often describe multiple tactics or threat types simultaneously, multi-label classification is the natural formalism for this domain [27], [29]. Future work should explicitly extend hybrid classifiers from mutually exclusive settings to systematic multi-label modelling. This includes exploring methods that learn label correlations to avoid ignoring dependencies among threat categories, as stacked ensemble approaches and transformer-based encoder-decoders have shown that exploiting these correlations improves performance [17], [19]. Methodologically, advanced label-structure representations, such as hierarchical or taxonomy-aware approaches, should be considered to leverage structured relationships beyond flat label sets [24]. Moreover, filter-based feature selection (e.g., Fisher-score-based) remains a relevant avenue for improving accuracy and efficiency in hybrid feature sets [29]. Finally, evaluation protocols must be strengthened. Because real-world datasets diverge from benchmarks, comparative evaluation on realistic CTI data is essential [18], and metrics beyond simple confusion matrices, such as confusion tensors, should be utilised to account for partial correctness faithfully [20].

5) Deployment in Security Operations Centres (SOCs): Efficiency, Robustness, and Monitoring. Operational deployment in SOCs introduces strict constraints regarding efficiency, scalability, and robustness. The multi-label literature offers several solutions: online algorithms designed to reduce computational burdens while modelling correlations [26], and neighbourhood-based methods that improve speed through constraint optimisation [30]. Furthermore, dimensionality reduction techniques, such as saliency-based multi-label LDA, can be employed to streamline classification pipelines [31]. Robustness strategies must also account for incomplete data, utilising label propagation to handle missing labels in production cycles [27], [28]. Finally, SOC-aligned deployments require explicit monitoring suitable for multi-label outputs. Given the limitations of standard evaluation, improved formulations like confusion tensors should be standard [20]. Where workflows require prioritisation, graded multi-label methods offer a principled bridge between classification and analyst-oriented ranking [22], [23]. All deployment models should be validated on real-world CTI distributions to ensure reliability in applied environments [18].

Acknowledgment

This publication could not have been accomplished without the institution’s extraordinary assistance. We also want to thank the reviewers whose suggestions will help make this manuscript eligible for publication.

References

[1] M. Bilal, A. Khan, S. Jan, S. Musa, and S. Ali, “Roman Urdu hate speech detection using transformer-based model for cyber security

applications," Sensors, vol. 23, no. 8, p. 3909, 2023, http://doi.org/10.3390/s23083909

[2] A. Das, M. Hoque, O. Sharif, M. Dewan, and N. Siddique, "Temox: Classification of textual emotion using ensemble of transformers," IEEE Access, vol. 11, pp. 109803–109818, 2023, http://doi.org/10.1109/ACCESS.2023.3319455

[3] I. Aden, C. Child, & C. Reyes-Aldasoro, "International Classification of Diseases Prediction from MIMIIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art", Big Data and Cognitive Computing, vol. 8, no. 5, p. 47, 2024. https://doi.org/10.3390/bdcc8050047

[4] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. Cordeiro, M. Debbah, T. Lestable, and N. Thandi, "Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT devices," IEEE Access, vol. 12, pp. 23733–23750, 2024, http://doi.org/10.1109/access.2024.3363469

[5] M. Ferrag, M. Ndhlovu, N. Tihanyi, L. Cordeiro, M. Debbah, T. Lestableet al., "Revolutionizing Cyber Threat Detection With Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices", Ieee Access, vol. 12, p. 23733-23750, 2024. https://doi.org/10.1109/access.2024.3363469

[6] S. Luo, Y. Gu, X. Yao, and W. Fan, "Research on text sentiment analysis based on neural network and ensemble learning," Rev. Intell. Artif., vol. 35, no. 1, pp. 63–70, 2021, http://doi.org/10.18280/ria.350107

[7] D. Preuveneers and W. Joosen, "Sharing machine learning models as indicators of compromise for cyber threat intelligence," J. Cybersecur. Privacy, vol. 1, no. 1, pp. 140–163, 2021, http://doi.org/10.3390/jcp1010008

[8] S. Saeed, S. Suayyid, M. Al-Ghamdi, H. Almuhaisen, and A. Almuhaideb, "A systematic literature review on cyber threat intelligence for organizational cybersecurity resilience," Sensors, vol. 23, no. 16, p. 7273, 2023, http://doi.org/10.3390/s23167273

[9] S. Silvestri, S. Islam, S. Papastergiou, C. Tzagkarakis, and M. Ciampi, "A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem," Sensors, vol. 23, no. 2, p. 651, 2023, http://doi.org/10.3390/s23020651

[10] B. Tang, J. Wang, H. Qiu, J. Yu, Z. Yu, and S. Liu, "Attack behavior extraction based on heterogeneous cyberthreat intelligence and graph convolutional networks," Comput. Mater. Contin., vol. 74, no. 1, pp. 235–252, 2023, http://doi.org/10.32604/cmc.2023.029135

[11] X. Zhang, X. Miao, and M. Xue,"A reputation-based approach using consortium blockchain for cyber threat intelligence sharing," Secur. Commun. Netw., vol. 2022, pp. 1–20, 2022, http://doi.org/10.1155/2022/7760509

[12] Y. Liu and Y. Dai, "Deep Learning in Cybersecurity: A Hybrid BERT–LSTM Network for SQL Injection Attack Detection", Iet Information Security, vol. 2024, no. 1, 2024. https://doi.org/10.1049/2024/5565950

[13] Y. Seyyar, A. Yavuz, & H. Ünver, "An Attack Detection Framework Based on BERT and Deep Learning", Ieee Access, vol. 10, p. 68633-68644, 2022. https://doi.org/10.1109/access.2022.3185748

[14] Y. Zhou, Y. Tang, M. Yi, C.-Y. Xu, and H. Lu,"CTI view: APT threat intelligence analysis system," Secur. Commun. Netw., vol. 2022, pp. 1–15, 2022, http://doi.org/10.1155/2022/9875199

[15] S. Kim, D. Kim, J. Kim, J. Koo, J. Yoon, & D. Yoon, "In-Context Learning with Large Language Models: A Simple and Effective Approach to Improve Radiology Report Labeling", Healthcare Informatics Research, vol. 31, no. 3, p. 295-309, 2025. https://doi.org/10.4258/hir.2025.31.3.295

[16] B. Bhamare and J. Prabhu, "A Multilabel Classifier for Text Classification and Enhanced BERT System", Revue D Intelligence Artificielle, vol. 35, no. 2, p. 167-176, 2021. https://doi.org/10.18280/ria.350209

[17] L. Duan, Q. You, X. Wu, & J. Sun, "Multilabel Text Classification Algorithm Based on Fusion of Two-Stream Transformer", Electronics, vol. 11, no. 14, p. 2138, 2022. https://doi.org/10.3390/electronics11142138

[18] S. Xu, Y. Zhang, X. An, & S. Pi, "Performance evaluation of seven multi-label classification methods on real-world patent and publication datasets", Journal of Data and Information Science, vol. 9, no. 2, p. 81-103, 2024. https://doi.org/10.2478/jdis-2024-0014

[19] Y. Xia, K. Chen, & Y. Yang, "Multi-label classification with weighted classifier selection and stacked ensemble", Information Sciences, vol. 557, p. 421-442, 2021. https://doi.org/10.1016/j.ins.2020.06.017

[20] D. Krstinić, A. Skelin, I. Slapničar, & M. Braović, "Multi-Label Confusion Tensor", Ieee Access, vol. 12, p. 9860-9870, 2024. https://doi.org/10.1109/access.2024.3353050

[21] K. Ding, L. Xu, M. Liu, X. Zhang, L. Liu, D. Zenget al., "Combing Type-Aware Attention and Graph Convolutional Networks for 嬰 vent 燚 etection", Computers Materials & Continua, vol. 74, no. 1, p. 641-654, 2023. https://doi.org/10.32604/cmc.2023.031052

[22] W. Farsal, M. Ramdani, & S. Anter, "GML_DT: A Novel Graded Multi-label Decision Tree Classifier", International Journal of Advanced Computer Science and Applications, vol. 12, no. 12, 2021. https://doi.org/10.14569/ijacsa.2021.0121233

[23] W. Farsal, M. Ramdani, & S. Anter, "An Effective Random Forest Approach for Mining Graded Multi-Label Data", International Journal of Intelligent Engineering and Systems, vol. 16, no. 4, p. 548-557, 2023. https://doi.org/10.22266/ijies2023.0831.44

[24] M. Ferrandin and R. Cerri, "Multi-label classification via closed frequent labelsets and label taxonomies", Soft Computing, vol. 27, no. 13, p. 8627-8660, 2023. https://doi.org/10.1007/s00500-023-08048-5

[25] T. Yu, J. Lin, L. Mou, Y. Hua, X. Zhu, & Z. Wang, "SCIDA: Self-Correction Integrated Domain Adaptation From Single- to Multi-Label Aerial Images", Ieee Transactions on Geoscience and Remote Sensing, vol. 60, p. 1-13, 2022. https://doi.org/10.1109/tgrs.2022.3170357

[26] Y. Zhang, "Learning Label Correlations for Multi-Label Online Passive Aggressive Classification Algorithm", Wuhan University Journal of Natural Sciences, vol. 29, no. 1, p. 51-58, 2024. https://doi.org/10.1051/wujns/2024291051

[27] A. Taha, S. Tiun, A. Rahman, M. Ayob, & A. Abdulameer, "Unified Graph-Based Missing Label Propagation Method for Multilabel Text Classification", Symmetry, vol. 14, no. 2, p. 286, 2022. https://doi.org/10.3390/sym14020286

[28] L. Sun, T. Wang, W. Ding, J. Xu, & A. Tan, "Two-stage-neighborhood-based multilabel classification for incomplete data with missing labels", International Journal of Intelligent Systems, vol. 37, no. 10, p. 6773-6810, 2022. https://doi.org/10.1002/int.22861

[29] R. Shaikh, M. Rafi, N. Mahoto, A. Sulaiman, & A. Shaikh, "A filter-based feature selection approach in multilabel classification", Machine Learning Science and Technology, vol. 4, no. 4, p. 045018, 2023. https://doi.org/10.1088/2632-2153/ad035d

[30] X. Jiang, J. Zhou, X. Qiao, C. Peng, & S. Su, "A Neighborhood Model with Both Distance and Quantity Constraints for Multilabel Data", Computational Intelligence and Neuroscience, vol. 2022, p. 1-10, 2022. https://doi.org/10.1155/2022/9891971

[31] L. Xu, J. Raitoharju, A. Iosifidis, & M. Gabbouj, "Saliency-Based Multilabel Linear Discriminant Analysis", Ieee Transactions on Cybernetics, vol. 52, no. 10, p. 10200-10213, 2022. https://doi.org/10.1109/tcyb.2021.3069338