Enhancing Human-AI Trust in Cyber Threat Intelligence via Interpretable Attack Phase Classification

Muhammad Saad Rashad[1], Mousa Al-Kfairy[2], Muhammad Amin[1], Hafeez Anwar[1], Waqas Ali[1], Sajid Anwar[3]

[1]National University of Computer & Emerging Sciences (NUCES-FAST), Pakistan
[2]Zayed University, Abu Dhabi, UAE
[3]Institute of Management Sciences, Peshawar, Pakistan
saad.rashad@nu.edu.pk, Mousa.Al-kfairy@zu.ac.ae, muhammad.amin@nu.edu.pk, hafeez.anwar@nu.edu.pk, Waqas.Ali@nu.edu.pk, sajid.anwar@imsciences.edu.pk

Abstract

The classification of cyber threat intelligence indicators into attack phases is essential for effective threat analysis and automated defense systems. However, such indicators are often short, sparse, and highly imbalanced, limiting the effectiveness of sequence-based deep learning approaches which is essential for establishing Human-AI trust in operational cybersecurity settings. In this work, we propose a hybrid classification framework that combines TF-IDF representations with dimensionality reduction and domain-specific binary features capturing structural properties of cyber indicators. Classical machine learning models, including Support Vector Machines, Decision Trees, and Logistic Regression, are evaluated and compared with an LSTM-based sequence models. Experimental results on a STIX-based dataset demonstrate that classical models consistently outperform the LSTM baseline, with the SVM achieving the highest macro-F1 score of 0.924 on the held-out test set. Cross-validation further confirms the robustness of the proposed approach, with only marginal variation across models. These findings highlight the effectiveness of sparse lexical representations augmented with domain knowledge for cyber threat indicator classification. Beyond predictive performance this study incorporates explainable analysis using SHAP to provide transparent insights into feature contributions across attack phases, supporting analyst trust and informed decision making. These results demonstrate that sparse lexical representations augmented with domain knowledge offer an efficient, interpretable, and trustworthy solution for attack-phase classification in operational CTI environments.

Introduction

In operational security environments, human-AI trust describes the extent to which analysts can confidently interpret, validate and rely on automated decision support systems (Baron 2025). Trustworthy collaboration between analysts and machine learning models requires more than predictive accuracy. Analysts must be able to understand which factors influence model decisions and assess their alignment with domain knowledge and investigate reasoning (Arrieta et al. 2020). Without such transparency, even high performing models may face resistance in practice, particularly in high stakes threat analysis where accountability and situational awareness are critical.

In cybersecurity an attack-phase represents a specific stage in the progression of a cyberattack, capturing the sequence of actions taken by an adversary to achieve their objectives (Hutchins, Cloppert, and Amin 2011). From an operational perspective, attack phases provide a structured lens through which adversarial behavior can be interpreted and communicated within Cyber Threat Intelligence (CTI) systems. Structured adversary behavior models such as the MITRE ATT&CK framework organize threat evidence into tactical stages that support systematic analyst reasoning about attack progression rather than isolated indicators (Roy et al. 2023).

Cyber threat Intelligence (CTI) plays a critical role in enabling organizations to detect, analyze, and respond to malicious cyber activities. Modern CTI is frequently shared in structured formats such as Structure Threat Information Expression (STIX), which standardizes the representation of indicators, attack patterns, and contextual metadata (Dimitriadis et al. 2025). Despite its syntactic consistency, a substantial portion of actionable intelligence within STIX reports embedded in unstructured textual fields, limiting their direct usability for automated analysis.

Recent studies have highlighted the growing scale and complexity of cyber threats across multiple sectors, emphasizing the need for scalable and machine-driven approaches to threat classification (Irshad and Siddiqui 2023). While deep learning and contextual language models have shown promise in processing security related text, these methods often require significant computational resources and large volumes of labeled data, which may hinder practical adoption in operational environments (Santos et al. 2025). In addition, the limited transparency of many neural network models can complicate their adoption in operational security environments, motivating the exploration of simpler and more practical classification approaches.

Unstructured threat reports often exhibit noise, sparse signals, and skewed label distributions, which degrade classifier performance and lead to unreliable pattern learning in multi-label settings. Prior research on TTP extraction confirms that data imbalance and structural variability pose significant challenges for automated CTI text analysis (Rahman

and Williams 2022).

In this work, we investigate text representation, augmented with simple domain specific indicators, can effectively support automated classification of STIX-based threat intelligence. We propose a hybrid framework that combines TF-IDF representations with binary features capturing the presence of common threat artifacts such as IP-addresses, cryptographic hashes, and email entities.

Extensive experiments on a real-world STIX dataset demonstrate that classical machine learning models, trained on proposed feature space achieve robust and competitive performance, comparable to neural baselines, while maintaining simplicity and interpret-ability. These findings suggest that domain aware feature engineering remains a viable effective approach for large-scale cyber threat text classification.

Beyond predictive accuracy, the practical deployment of automated threat classification systems depends on transparency and analyst trust. Security analysts need visibility into which features drive model decisions before relying on automated outputs in investigation and response workflows. To support this requirement, we use SHAP to provide global explanations of feature contributions across attack phases. By anchoring these explanations in semantically meaningful CTI artifacts, the proposed approach fosters interpretable and trustworthy humman-AI collaboration in cyber threat analysis.

The proposed CTI classification framework is hybrid in both feature representation and learning strategy. It combines textual features with domain specific binary indicators and employs model specific inputs: TF-IDF plus binary features for classical classifiers , and raw tokenized sequences plus binary features for the LSTM model. This hybridization enables the joint exploitation of statistical text patterns, sequential dependencies, and interpretable cybersecurity signals.

Literature Review

Recent research has increasingly explored the use of natural language processing (NLP) techniques to extract actionable insights from Cyber Threat Intelligence (CTI), which is predominantly shared through unstructured or semi-structured textual reports (Jo, Lee, and Shin 2022). Surveys by Arazzi et al. and Albarrak et al. provide a comprehensive overview of NLP and machine early attack prediction (Arazzi et al. 2025; Albarrak, Salonitis, and Jagtap 2026). These studies highlight that while modern NLP models can effectively capture contextual semantics from threat narratives, their deployment is often challenged by noisy data, inconsistent reporting formats, limited labeled datasets, and high computational requirements. Such constraints limit the practicality of complex models in real-world, operational CTI environments (Arazzi et al. 2025; Albarrak, Salonitis, and Jagtap 2026).

Advanced deep learning approaches, particularly transformer-based architectures, have shown promising results in processing unstructured CTI text (Li, Huang, and Chen 2024). Li et al demonstrated the use of DistilBERT for automatically mapping free-text threat reports to MITRE ATT&CK tactics and techniques, achieving strong classification performance through contextual sentence-level modeling (Li, Huang, and Chen 2024). Similarly, recent large language model (LLM)-based approaches have explored fine-grained attack techniques classification using multi-stage inference and data augmentation strategies (You and Park 2024). Despite their effectiveness, these methods introduce significant computational overhead, reduced interpret-ability, and increased inference latency, making them less suitable for large-scale or resource-constrained cyber defense (Li, Huang, and Chen 2024; You and Park 2024).

In the literature while classification of threats is a major topic of discussion, several studies emphasized on the extraction of structured intelligence from CTI events (Marchiori, Conti, and Verde 2023). Finished intelligence reports were utilized for automated extraction of tactics, techniques and procedures (TTPs) by deploying TTPXHunter framework, which reduced analyst workload while improving context awareness (Rani et al. 2024). Techniques derived for CTI event extraction highlight that constructive analysis of threat reports must be contingent upon robust pre-processing pipelines, particularly when dealing with heterogeneous data sources and multilingual content (Al-Yasiri et al. 2024). These studies show that inconsistencies in reporting formats and language variation significantly impact downstream extraction performance. At the same time, many proposed solutions rely on sophisticated NLP pipelines and manually annotated datasets, which raises concerns regarding scalability and adaptibility when applied to divers or evolving CTI repositories (Rani et al. 2024; Al-Yasiri et al. 2024).

Alongside these complex neural techniques, several lightweight, feature-based methods continue to be effective for cybersecurity text classification (Hossen, Borshon, and Badrudduza 2025). Particularly, Hossen et al. proposed a modified TF-IDF representation that enhances both computational efficiency and classification performance on cyber threat text (Hossen, Borshon, and Badrudduza 2025). Other studies employing classical machine learning classifiers, such as Support Vector Machines and Random Forests, have demonstrated that engineered textual features can achieve competitive performance compared to neural models. These approaches offer advantages in terms of interpret-ability, ease of deployment , and reduced dependency on large labeled datasets, making them attractive for operational CTI systems (Li et al. 2022; Tian et al. 2025).

In addition to efficiency and performance, recent literature emphasizes that the adoption of automated CTI analysis in operational environments depends heavily on transparency and analyst trust. Analysts must be able to understand and validate model behavior before relying on automated predictions in security workflows. Consequently, explainable machine learning has emerged as key enabler fro trustworthy human-AI collaboration in cybersecurity, particularly in high-stakes threat analysis scenarios where opaque decision making can hinder practical deployment (Arrieta et al. 2020; Ofusori, Bokaba, and Mhlongo 2025).

Despite the growing body of work on NLP-Driven CTI

analysis, a gap remains in developing practical and interpretable frameworks that effectively transform unstructured STIX-based threat intelligence into machine-readable representations without relying on complex neural architectures (Li, Huang, and Chen 2024; Li et al. 2022). Most existing approaches either emphasize high-capacity models with limited operational feasibility or focus on extraction tasks without systematic evaluation of classification performance. Motivated by these limitations, this work investigates a hybrid feature representation that combines TF-IDF based textual embeddings with simple domain specific binary indicators, enabling efficient and interpretable feature based classification of STIX-based cyber threat intelligence.

Methodology

This work proposes a hybrid framework for attack-phase classification of cyber threat indicators using both classical machine learning and deep learning approaches. The methodology follows as shown in 1 a shared preprocessing and feature extraction stage followed by two model specific pipelines: a TF-IDF based representation for classical classifiers and a sequence based representation for a Long Short Term Memory (LSTM) network. In both pipelines, domain specific binary features are incorporated to enhance semantic understanding of cyber indicators.

Dataset Description

Each dataset instance consists of a cyber threat indicator value paired with an attack-phase category. The category labels represent high-level stages of malicious activity , including `'network activity'`,`'payload delivery'`, `'payload installation'`, `'external analysis'`, `'artifacts dropped'` and `'other'`. The indicator values are heterogeneous and may include file, names, cryptographic hashes, IP address, email addresses, executables, timestamps and related artifacts observed during different phases of an attack lifecycle.

Data Preprocessing

Textual indicator values are normalized prior to feature extraction, HTML artifacts are removed, URLs are replaced with a placeholder token, and all text is converted to lowercase to ensure consistent token representation. Records containing missing values are excluded from further analysis. Although the indicators do not form natural language sentences, token-based preprocessing enables the learning of structural and contextual patterns present in cyber threat data.

Domain Specific Binary Feature Extraction

To inject structured cybersecurity knowledge, binary domain features are extracted from each indicator using regular expression. In order to leverage explicit domain cues that may not be consistently captured through textual representations, three features poised inclusion and exclusion of IP addresses, hash values (MD5, SHA1, and SHA256) and email addresses. Each feature is encoded as a binary flag to capture domain context.

Label Encoding and Data Splitting

The dataset comprises of both attack-phase categories and their values, while later are subjected to TF-IDF pipeline, categories are encoded using label encoder. Furthermore, the dataset split into training and testing sets is performed through stratified sampling to preserve class distribution.

Classical Machine Learning Pipeline

TF-IDF Vectorization Pipeline Feeding threat indicators to machine learning models needs proper handling and for that these values are subjected to Term Frequency Inverse Document Frequency (TF-IDF) to obtain features, which are constructed from both unigrams and bigrams.

Since the original values are heterogeneous in nature, this technique transforms the variable length indicators into fixed size numerical vectors suitable for subsequent classifier. TF-IDF vectorization acquired entails high dimensionality and sparsity, Truncated Singular Value Decomposition (SVD) is employed to obtain a compact latent feature space while preserving the most information component.

Feature Union and Class Balancing In this step, the binary domain specific features obtained are fused with the resultant SVD reduced TF-IDF vector space to complement the lexical semantics with structural, indicator type information. Ensuring minimal class imbalance to reduce bias exhibited in dataset, the Synthetic Minority Oversampling Technique (SMOTE) is applied to the training set after feature integration step. The complete strategy proposed in this step is vital to improve class balance without causing unintended class bias transfer into next phase that entails model training and evaluation.

Model Evaluation Three machine learning models are evaluated to answer these queries: (i) SVM is evaluated to answer, whether indicators belonging to the same attack phase lie closer together in the semantic domain feature space than to other phases. (ii) Logistic Regression is employed to verify that attack categories can be distinguished using linear combinations of lexical and domain specific cues. (iii) Decision Tree is utilized to know the identification of attack phases using small number of high impact rules. Furthermore, to asses model robustness the three classical machine learning models are evaluated using five fold cross validated.

Deep Learning Pipeline

Raw Sequence Tokens In this step, threat indicators are first tokenized. Since the nature of these values is heterogeneous the tokens are normalized using fixed sequence length padding, which ensures consistent input dimensionality for the recurrent neural network.

Token Fusion with Binary Features The obtained fixed length token sequences are mapped to a dense embedding space and processed using multilayer Long Short Term Memory (LSTM), which is used to capture sequential dependencies in the provided padded tokens. Trained sequences are fused together with binary features before given to the classification layer. This method allows the model to

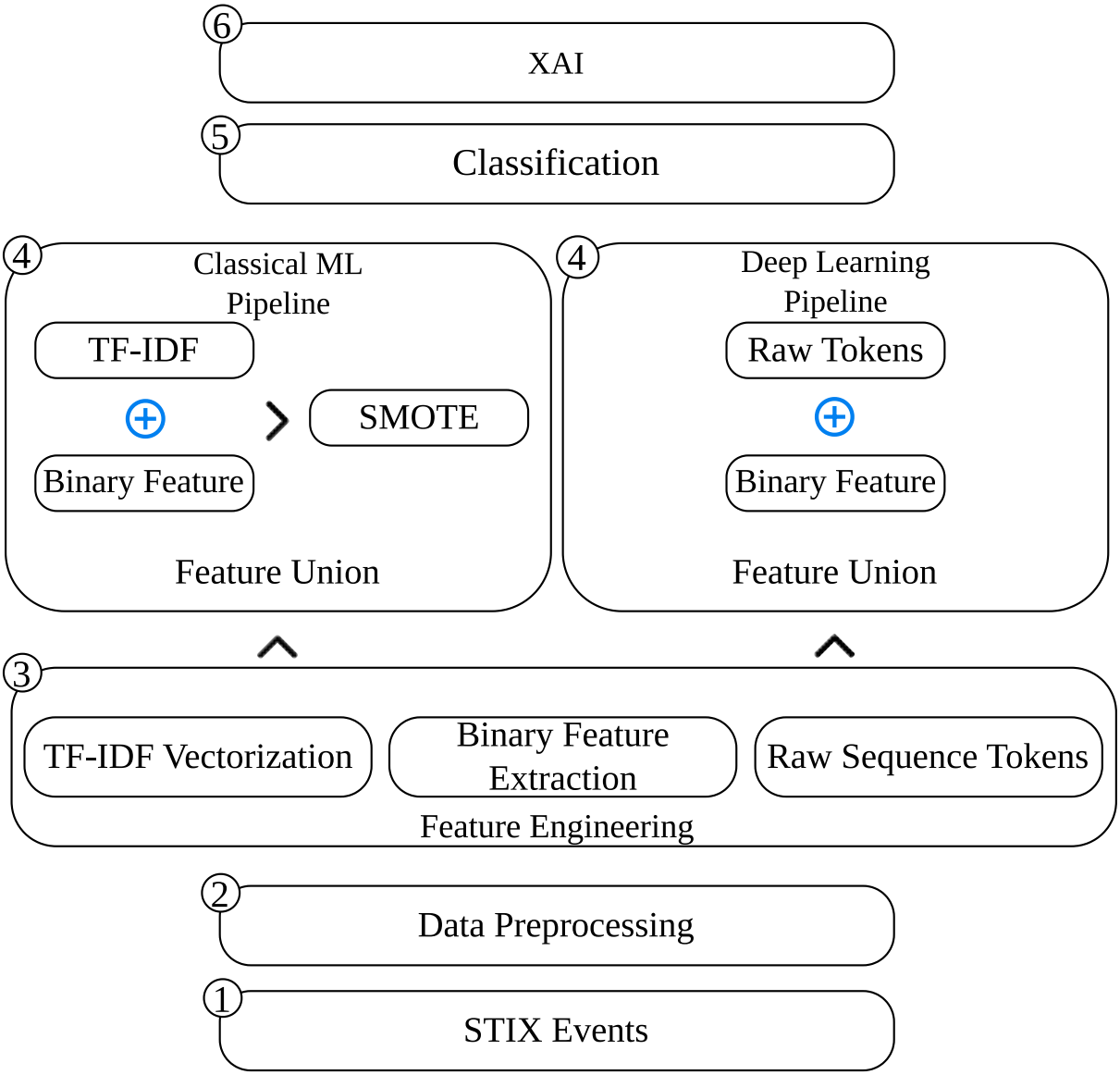

Figure 1: Hybrid Framework Architecture

learn contextual patterns from the padded vector space and absorb the domain specific signals within the unified framework.

Training and Evaluation To evaluate the process of raw tokens fused with binary features, the LSTM model for multiclass nature of the dataset is optimized using the standard Adam optimizer, whereas the loss function is calculated using sparse categorical cross entropy. To verify the effectiveness LSTM is evaluated on unseen test set using standard classification metrics (Table- 1) for consistent comparison with classical machine learning models.

Feature Representation Rationale

Different feature representations are employed based on model requirements: (i) TF-IDF vectors used for classical machine learning models, (ii) while raw token sequences are used for the LSTM to preserve temporal dependencies. In both cases, binary domain features are consistently incorporated.

To evaluate the effectiveness of the proposed hybrid framework, Classical ML models are compared with state-of-the-art LSTM network. Logistic Regression serves as interpretable baseline due to its transparency and compatibility with feature based explanations, while LSTM captures sequential patterns in textual threat data. This comparison highlights the trade-of between predictive performance and model interpretability in cyber threat intelligence classification.

Model Transparency Using SHAP

To enhance transparency in the proposed classification framework, SHAP (SHapley Additive exPlanations) is applied to the trained logistic Regression model operating on the combined TF-IDF and domain specific binary feature space. SHAP quantifies the marginal contribution of each feature to the predicted attack-phase class probabilities, enabling interpretation of how individual lexical indicators and domain features influence model decisions.

The linear and additive nature of logistic Regression, SHAP values can be computed directly from the model's learned coefficients, resulting in stable and consistent feature attributions across high-dimensional TF-IDF representations. The resulting global importance scores highlight the most influential features friving predictions and allow analysts to verify whether the model relies on meaningful adversarial indicators rather than spurious correlations.

Logistic Regression was selected for SHAP-based interpretation because its transparent structure provides clearer and ore reliable explanations compared to complex nonlinear models, thereby supporting the objective of balancing predictive performance with model transparency in cyber threat intelligence analysis.

Evaluation Metrics

All models are evaluated using accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. Macro averaging is used to account for class imbalance and ensure equal importance across attack-phase categories.

Results

This section presents the experimental results of the proposed classification framework. Models are evaluated using accuracy, precision, recall, and macro-averaged F1 score on a held-out test set. Cross-validation results and additional analyses are discussed subsequently.

Table 1 compares the performances of classical machine learning models and an LSTM baseline on the test set. Among all the models, SVM with an RBF kernel achieves the highes macro-F1 score (0.9243), indication superior overall classification performance. Decision Tree and Logistic Regression exhibit comparabale results with slightly lower F1 scores, while the LSTM underperformed relative to TF-IDF baseline approaches.

The relatively strong performance of classical models suggests that sparse lexical representation combined with domain-specific features are well suited for cyber threat indicator classification.

Model	Accuracy	Precision	Recall	F1
SVM (RBF)	0.9121	0.9419	0.9241	0.9243
Decision Tree	0.9091	0.9353	0.9235	0.9222
Logistic Regression	0.9092	0.9368	0.9219	0.9200
LSTM	0.8878	0.9152	0.9086	0.9096

Table 1: Performance comparison of different models

An ablation study was performed across all models to analyze the contribution of individual feature sets 2. Across models, the combination of TF-IDF and binary features consistently outperforms individual feature sets, underscoring the complementary benefits of integrating textual representations with domain specific indicators.

To assess robustness, classical models were further evaluated using five-fold cross validation. Figure 5 shows the cross validation F1 scores of each model. Although the Decision Tree attains the highest average F1 score across folds, its improvement over SVM is marginal ($\approx 0.2\%$), suggesting similar robustness across models.

Model	Feature Set	Macro F1
SVM (RBF)	TF-IDF only	0.819
	Binary only	0.475
	TF-IDF + Binary	**0.924**
Decision Tree	TF-IDF only	0.883
	Binary only	0.475
	TF-IDF + Binary	**0.923**
Logistic Regression	TF-IDF only	0.815
	Binary only	0.475
	TF-IDF + Binary	**0.920**

Table 2: Ablation Study Results (Macro F1 Score)

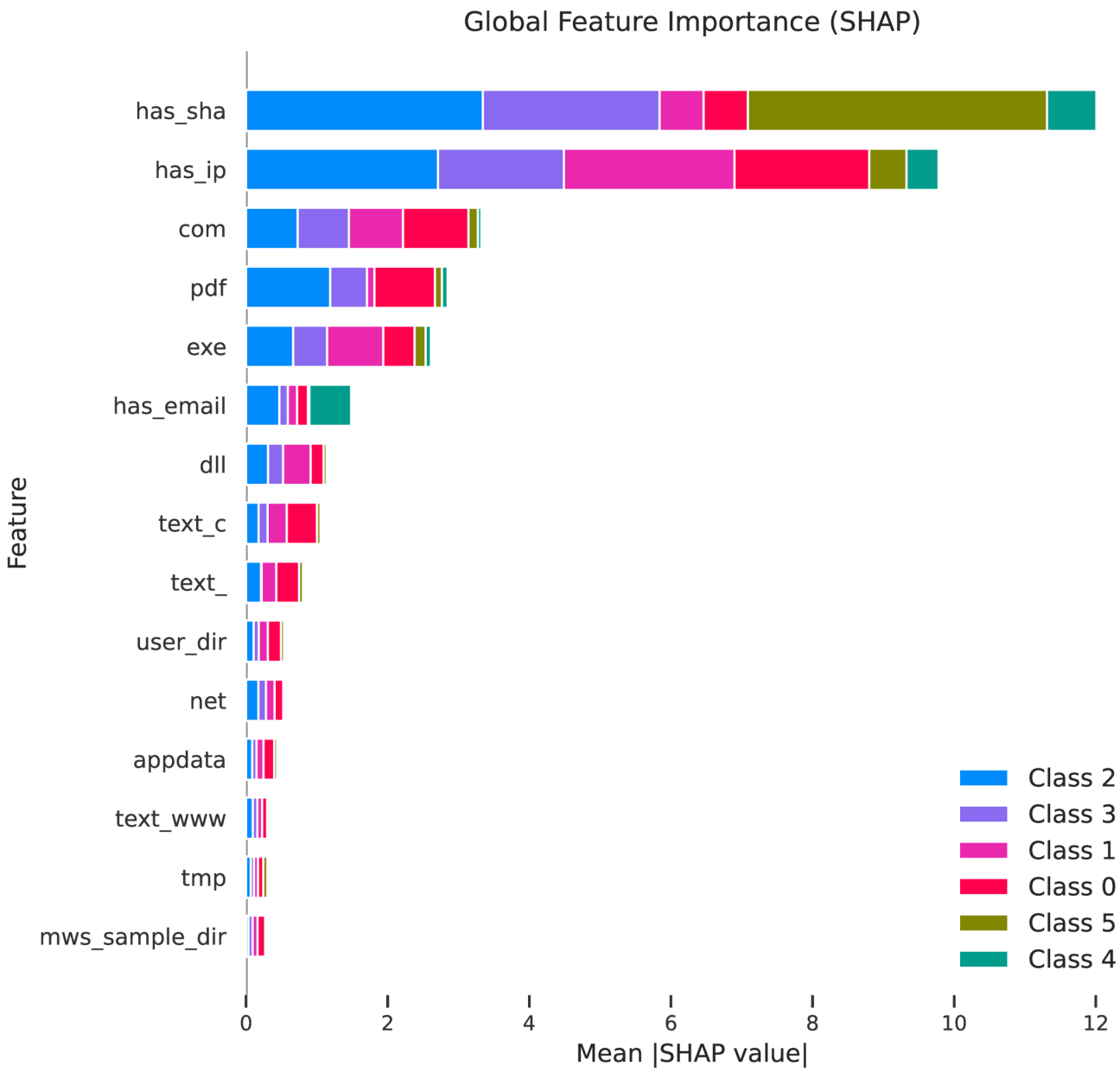

Figure 2: Global Feature Importance

To analyze per class performance, confusion matrices for all models are presented in Figures 3, 4. The matrices reveal that while SVM achieves the highest overall F1 score, Decision Tree correctly classifies several instances in minority classes that SVM occasionally mislabels. Logistic Regression shows balanced performance across classes, whereas LSTM demonstrates competitive results but struggles with specific categories likely due to limited sequence length and data sparsity.

Figure 2 presents the global SHAP feature importance for the Logistic Regression model. The results show that domain-specific indicators and executable artifacts dominate the decision making process of the model across attack phases. These features correspond to semantically meaningful CTI evidence commonly used by human analyst.

This alignment between model reasoning and analyst-recognizable indicators support trust in the automated predictions and facilitates effective human-AI collaboration in cyber threat intelligence workflows.

Within the CTI setting, SHAP based attribution enables direct mapping between discriminative lexical and domain-specific features and the predicted attack phases. This facilitates verification of whether model decisions are driven by semantically relevant threat indicators aligned with ad-

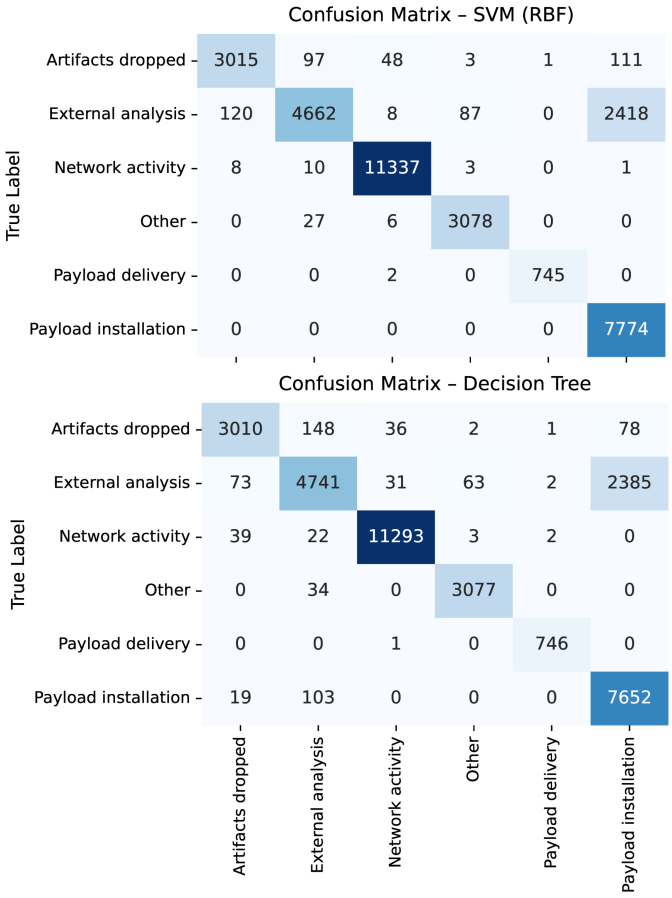

Figure 3: Confusion matrices of all evaluated models

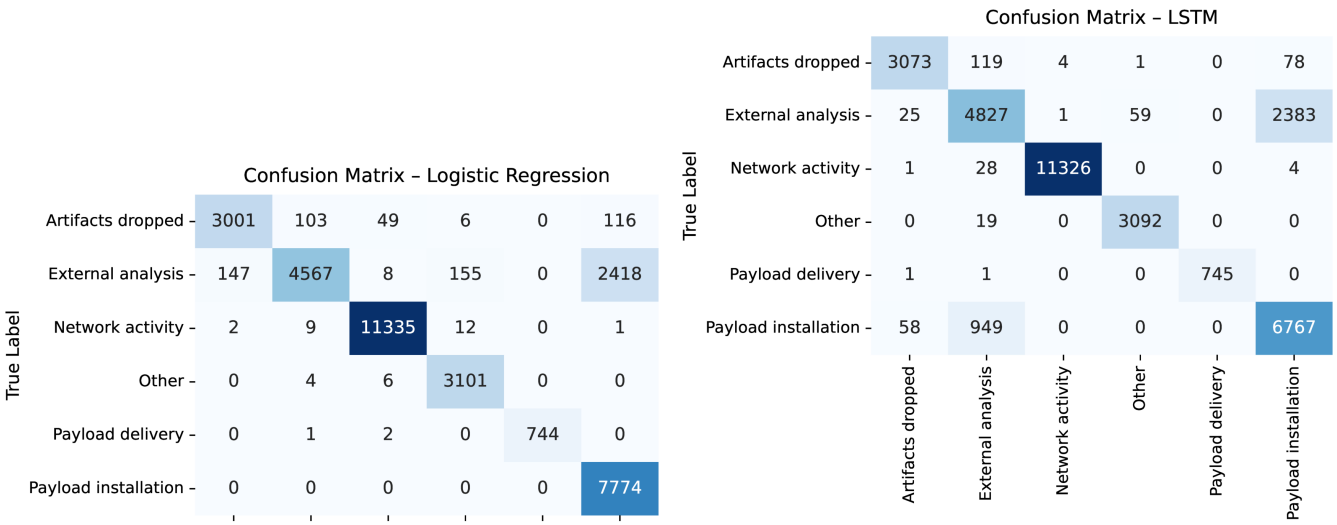

Figure 4: Confusion matrices of all evaluated models

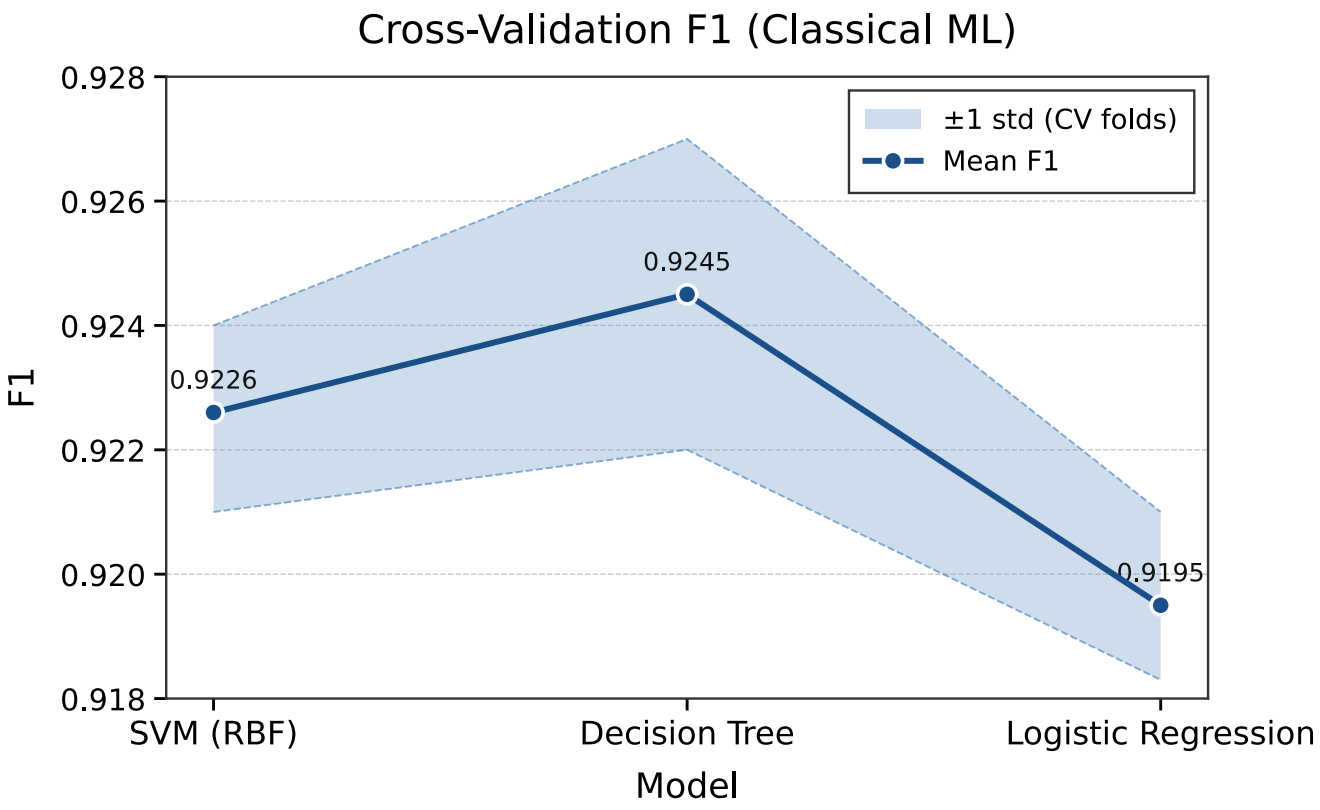

Figure 5: Mean Cross-Validation (5-Fold) of ML Models

versarial behavior patterns. By exposing class specific feature contributions, the framework supports analytical validation of attack-phase assignments and mitigates reliance on opaque prediction signals in operational threat analysis.

Conclusion

This study investigated the problem of attack phase classification for cyber threat intelligence indicators using both classical machine learning models and deep learning approaches. A hybrid framework combining TF-IDF representations with domain specific binary features was evaluated on STIX derived dataset. Experimental results demonstrate that classical models achieve strong performance and competitive robustness. In addition, the use of explainable analysis enables transparent inspection of feature contributions, helping bridge model predictions with analyst understanding and trust. These findings demonstrate that classical machine learning models, when augmented with domain knowledge, remain effective and practical for attack phase classification in operational CTI settings. Future work will explore richer contextual representations and more fine-grained explainability techniques to further support analyst-driven cyber threat investigation.

References

Al-Yasiri, J. H.; Zolkipli, M. F. B.; Farid, N. F. N. M.; Alsamman, M.; and Mohammed, Z. A. 2024. A Threat Intelligence Event Extraction Conceptual Model for Cyber Threat Intelligence Feeds. In *2024 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS)*, 1–8. IEEE.

Albarrak, M.; Salonitis, K.; and Jagtap, S. 2026. Natural language processing (NLP)-based frameworks for cyber threat intelligence and early prediction of cyberattacks in Industry 4.0: a systematic literature review. *Applied Sciences*, 16(2): 619.

Arazzi, M.; Arikkat, D. R.; Nicolazzo, S.; Nocera, A.; KA, R. R.; and Conti, M. 2025. NLP-based techniques for cyber threat intelligence. *Computer Science Review*, 58: 100765.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Baron, S. 2025. Trust, explainability and AI. *Philosophy & Technology*, 38(1): 4.

Dimitriadis, A.; Papoutsis, A.; Kavalieros, D.; Tsikrika, T.; Vrochidis, S.; and Kompatsiaris, I. 2025. EVACTI: Evalu-

ating the actionability of cyber threat intelligence. *International Journal of Information Security*, 24(3): 123.

Hossen, M. S.; Borshon, M. Z. I.; and Badrudduza, A. 2025. An Efficient Classification Model for Cyber Text. arXiv:2511.03107.

Hutchins, E. M.; Cloppert, M. J.; and Amin, R. M. 2011. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, 1(1): 80.

Irshad, E.; and Siddiqui, A. B. 2023. Cyber threat attribution using unstructured reports in cyber threat intelligence. *Egyptian Informatics Journal*, 24(1): 43–59.

Jo, H.; Lee, Y.; and Shin, S. 2022. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. *Computers & Security*, 120: 102763.

Li, L.; Huang, C.; and Chen, J. 2024. Automated discovery and mapping ATT&CK tactics and techniques for unstructured cyber threat intelligence. *Computers & Security*, 140: 103815.

Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P. S.; and He, L. 2022. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology*, 13(2): 1–41.

Marchiori, F.; Conti, M.; and Verde, N. V. 2023. StixNet: A novel and modular solution for extracting all STIX objects in CTI reports. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 1–11.

Ofusori, L.; Bokaba, T.; and Mhlongo, S. 2025. Explainability and interpretability of artificial intelligence use in cybersecurity. *Discover Computing*, 28(1): 1–23.

Rahman, M. R.; and Williams, L. 2022. From threat reports to continuous threat intelligence: a comparison of attack technique extraction methods from textual artifacts. arXiv:2210.02601.

Rani, N.; Saha, B.; Maurya, V.; and Shukla, S. K. 2024. TTPXHunter: Actionable threat intelligence extraction as TTPs from finished cyber threat reports. *Digital Threats: Research and Practice*, 5(4): 1–19.

Roy, S.; Panaousis, E.; Noakes, C.; Laszka, A.; Panda, S.; and Loukas, G. 2023. SoK: The MITRE ATT&CK Framework in Research and Practice. arXiv:2304.07411.

Santos, P.; Abreu, R.; Reis, M. J.; Serôdio, C.; and Branco, F. 2025. A systematic review of cyber threat intelligence: the effectiveness of technologies, strategies, and collaborations in combating modern threats. *Sensors*, 25(14): 4272.

Tian, Y.; Xu, S.; Cao, Y.; Wang, Z.; and Wei, Z. 2025. An Empirical Comparison of Machine Learning and Deep Learning Models for Automated Fake News Detection. *Mathematics*, 13(13): 2086.

You, W.; and Park, Y. 2024. Cyber-attack technique classification using two-stage trained large language models. arXiv:2411.18755.