Int. J. Advance Soft Compu. Appl, Vol. 18, No. 2, July 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Cybercrime Prediction via Multi-Modal Learning from Behavioral, Network, and Textual Data

Ali Awad Aljabrah[1], Ahmad Hisham Alomari[1], Mohammad Abdul Jawad[2] and Wlla Atef Amayreh[1]

[1] Law Department, Al-Zaytoonah University of Jordan, Amman, Jordan
[2] Software Engineering Department, Al-Zaytoonah University of Jordan, Amman, Jordan

a.algabrah@zuj.edu.jo, a.h.alomari@zuj.edu.jo, m.abduljawad@zuj.edu.jo, w.amayreh@zuj.edu.jo

Abstract

The increasing sophistication of cybercriminal activities creates significant challenges for traditional security systems, which rely heavily on unimodal detection methods. In this paper, we propose an integrated multi-modal deep learning methodology for better travel predictions by considering time series ordering od behavioral event transaction patterns with network traffic characteristics and textual threat intelligence. However, the suggested framework has modality-specific encoders which include Bidirectional Long Short-Term Memory; for behavioral sequence modeling, Graph Neural Networks, which focus on network topological processing, and BERT-based transformers with a focus on textual feature extraction. A new attention-based fusion mechanism finds latent representations from each modality, allowing the model to adjust source information dynamically depending on its predictive utility. Using three benchmarked datasets, IEEE-CIS Fraud Detection, CICIDS2017, and curated threat intelligence corpora for experimental evaluation of the proposed approach yield an F1-score of 0.912 and ROC-AUC of 0.947 and surpasses the respective single-modalities baselines by margins of 8.3% and 11.7%.

Keywords: *cybercrime prediction, multi-modal learning, deep learning, fraud detection, network intrusion, natural language processing, behavioral analytics*

1 Introduction

In the last decade, cybercrime has posed an unprecedented trillions of dollars annually in global losses by 2025 against advanced persistent threats, socially engineered phishing campaigns all represent ever more varied and sophisticated attack vectors that have pushed traditional rule-based detection systems to their limits [1]. New generation cybercriminal organisations exploit a number of attack vectors in parallel, orchestrating campaigns that cross financial transaction and low-level network infrastructure set-up [2] whilst embedding social engineering drive as well.

Today, most of the solutions in cybersecurity are domain specific and tend to rely on single stream of data. Financial fraud detection systems based on transactions typically check data transactions only and do not consider behaviors or contents of communication in a

Received 12 February 2026; Accepted 5 May 2026

network [3]. Network intrusion detection methods monitor traffic flows for potentially malicious activity but do not consider the surrounding context of payload data or user actions. Some text-based threat intelligence platforms also aggregate security reports and forum discussions, but lack behavioral indicators to validate fresh threats [4]. Such a siloed approach constrains the effectiveness of all detection at best, since advanced cyber-attacks typically manifest themselves in many observable facets.

Multi-modal machine learning provides a promising paradigm to alleviate these limitations [5]. Multi-modal architectures leverage complementary information hidden from unimodal systems through joint modeling of heterogeneous data sources [6]. Incorporating behavioral patterns, network topology and textual data captures a complete view of cyber behavior, which in turn helps enhance the model's performance for anomaly detection and threat prediction [7]. Additionally, for security applications where false positives lead to high operational costs, there are modern machine learning studies possess motivating the development of intelligent and deception-based approaches for enhanced system protection [8].

The contributions of this paper are as follows. Finally, we present a single multi-modal architecture for the first time, integrating behavioral, network and textual data streams to enable end-to-end cybercrime prediction. Second, we develop modality-specific encoders tailored for specific characteristics of data, which consist in temporal dependencies on behavioural sequences, topological structure for network flows and semantic content in textual sources. Third, we embed a learnable fusion mechanism that allows us to flexibly weigh modality contributions with respect to the task at hand. Fourth, we perform comprehensive experiments that show consistent performance improvements over state-of-the-art unimodal baselines and early-fusion baselines on several cybersecurity tasks.

2 Related Work

2.1 Financial Fraud Detection

The machine learning based methods to detect financial fraud have progressed from the conventional statistical approaches to more advanced deep learning architectures. Early work by Phua et al. also conducted a survey of classifiers such as decision trees, neural networks for credit card fraud detection [9]. More recently, Randhawa et al. prove the power of ensemble, combining random forests and gradient boost for transaction classification [10]. Deep learning approaches in particular have shown great promise: Fu et al. used convolutional neural networks to model local transaction patterns [11], and Zhang et al. used recurrent architectures to capture temporal dependencies in user spending patterns [12]. An enhancement introduced to these models is attention mechanisms, which help the model focus on suspicious subsequences of the transaction. Nevertheless, these methods treat fraud detection as an independent classification task without taking into account meaningful network behaviors or textual features that may be present when fraud occurs.

2.2 Network Intrusion Detection

Similar progress in machine learning has been beneficial for network intrusion detection. Traditional signature-based systems are no longer sufficient and it is continuously being replaced by approaches based on anomaly detection, which can also detect zero-day attacks [13]. Vinayakumar et al. systematically assessed the performance of deep learning

architectures such as CNN, RNN and autoencoders for network traffic classification [14]. Graph-based solutions have proven especially successful to model the structural relations characteristic of network communications: Lopez-Martin et al. used graph convolutional networks in the context of intrusion detection [15] (et al. Shen et al. propose attention-augmented graph neural networks to detect anomalous connection patterns [16]. Network-centric approaches, while useful, are often limited in not having contextual data about the sites and organizations involved in DNS communications, and that context can help model normal behavior for those entities to determine if legitimate unusual traffic may be carried out from an attacker.

2.3 Cyber Threat Intelligence (CTI) and NLP

Cyber threat intelligence extraction and analysis has seen increased application of natural language processing techniques. Satvat et al. built information extraction pipelines to discover threat actors and attack techniques from security reports [17]. Cybersecurity text classification are examples of strong performance with BERT based models: Aghaei et al. transformer architectures in this domain for phishing website detection based on textual content [18], whereas Ni et al. used similar techniques to analyze vulnerability descriptions [19]. Recently, we are seeing potential of large language models for automated threat report generation, indicator extraction etc. Nevertheless, purely text-based approaches fail to correlate linguistic cues with actual system behavior and can trigger false alarms when the description of a threat overlaps with routine operational activity.

2.4 Multi-Modal Learning in Cybersecurity

While there is a clear opportunity for multi-modal approaches to improve intrusion detection, the literature on this topic is lacking. Al-Hawawreh et al. then proposed fusion architecture that functions with data about network flow and sequences of system calls, achieving better detection for the botnet [20]. Zhang et al. used email contents along with sender behaviour patterns for phishing [21]. Nonetheless, these methods usually only fuse two modalities and based on simple concatenation of embeddings, which will not capture complicated cross-modal interactions. To the best of our knowledge, there is no prior work that systematically explored a unified deep learning framework for integrated behavioral, network and textual signals to achieve comprehensive cybercrime prediction. This motivates our proposed architecture, which pushes the state of the art in terms both of wider modality coverage and more advanced fusion mechanisms.

3. Methodology

3.1 Framework Overview

We designed our multi-modal framework with three main components: modality-specific encoders that convert raw heterogeneous data into unified latent representations, a cross-modal fusion module to compose information across modalities, and a prediction head for generating final risk scores. The overall architecture is shown in figure 1, which shows the separate processing streams for the behavioral data, the network data and text plus embedding inputs until fusing together at prediction.

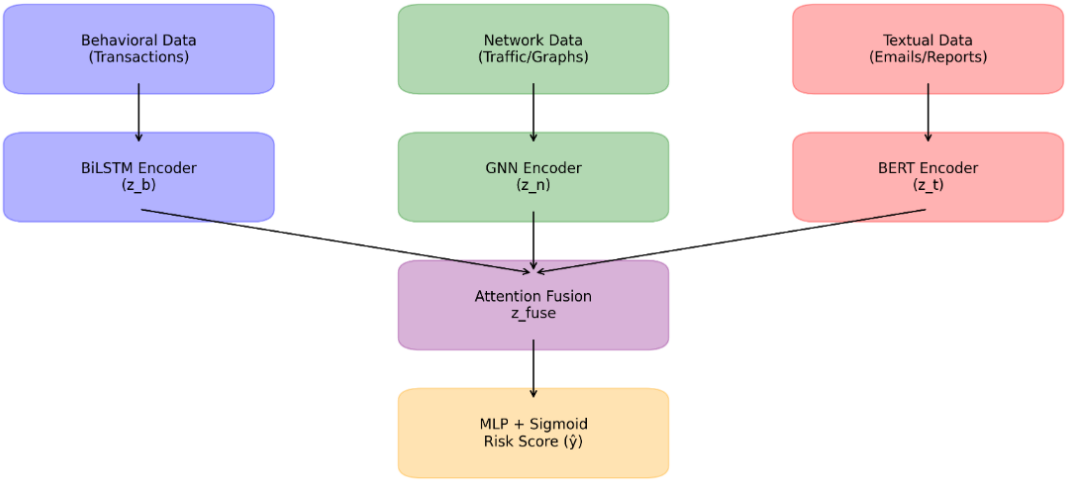

Figure 1: Proposed multi-modal architecture.

3.2 Behavioral Modality Encoder

Behavioral Data: A Time story One transaction or action may seem benign, but patterns over sequences often uncover irregularities. This is especially relevant in the case of fraud detection, where anomalies appear through temporal deviations rather than isolated events.

A Bidirectional LSTM is used by the framework to capture these patterns. This model reads sequences both forward and backward, giving it insight into how past actions affect what comes next, and vice versa. This means in practice that the model learns not only to know what has happened, but also how behavior is evolving over time.

As a result, we have a compact format that summarizes user activity in the form of deviations from normal behavior.

3.3 Network Modality Encoder

Cyber-threats spread via networks revealing patterns in interactions between entities. These interactions are naturally expressed as graphs, i.e., entities are connected through exchange of data.

To analyze this structure, the model is based on a Graph Neural Network (GNN). Unlike prior models which consider each entity separately, the network considers how nodes influence each other by way of their connections. This allows it to identify unusual communication patterns such as abnormal clusters or suspicious interaction paths.

The model aggregates information across the network and thus creates a global picture of the activity, which is crucial for detecting coordinated or distributed attacks.

3.4 Textual Modality Encoder

Textual data (e.g., emails, reports, or forum postings) can provide context that is missing from numerical or structural data, but it may indicate intent, urgency, or deception, such as in the case of phishing or social engineering.

To get this information, the framework uses a transformer-based model (BERT), which can understand language in context. It recognizes that instead of treating words as

independent entities, there are relationships among words that can be observed, such as suspicious expressions or manipulative language.

The resultant representation has the meaning of the text and is a valuable compliment to behavioral and network signals.

3.5 Cross-Modal Fusion

Once both are processed, the next step is to combine them. Not all sources of information will be equally useful in all situations. For example, it is possible that behavioral data is the most useful in detecting fraud, whereas textual data may be valuable in identifying phishing.

To address this problem, we have employed an attention-based fusion technique that allows the model to adjust the weighting of each modality based on the context.

In other words, it can learn to "focus" on the most relevant information while downplaying the less useful signals. It is able to respond to various types of cybercrime.

3.6 Prediction and Loss Function

The last step is to convert the merged representation into a prediction by using a neural network that outputs a probability of whether an activity is malicious.

One of the biggest challenges in cybersecurity is class imbalance; malicious events are rare compared to normal ones. To correct this problem, the training process focuses on difficult or misclassified examples to ensure that the model does not become biased toward the majority class.

This improves the ability of the model to identify rare, but critical, events in real-world systems.

4. Experimental Setup

4.1. Datasets

The proposed framework is evaluated using three publicly available datasets of different cybercrime scenarios in different modes (unimodal and multimodal).

- IEEE-CIS Fraud Detection

The IEEE-CIS Fraud Detection dataset released through the Kaggle competition platform contains more than 590,000 anonymized financial transactions with binary fraud labels [22]. There are 394 numerical features built from transaction amounts, timestamps, and card attributes. Transaction sequences are grouped by user name to allow temporal modeling. The data set is very class averse, with fraudulent transactions representing only 3.5% of samples. In this paper, we normalize the numerical features with the z-score standard and create transactions that are longer than 10 for the BiLSTM encoder.

- CICIDS2017

CICIDS2017 by the Canadian Institute for Cybersecurity captures network traffic captured over 5 days in benign and attack scenarios [23]. Included in the dataset are common attacks (DoS, DDoS), brute force, and infiltration. Flow information (e.g., packet counts, byte sizes, and timing of a flow) is extracted using CICFlowMeter. We create communication graphs by treating IP addresses as nodes and bidirectional flows as edges, which combine

multiple flows between the same endpoints. The nodes are statistical summaries of associated flows. Including 283,074 flow records from 15 attack categories and benign traffic.

- Threat Intelligence Corpus

To assess the textual modeality, we develop a threat intelligence corpus from several datasets. The Enron Email Dataset includes benign emails [24]; the security-oriented text from Kaggle Cybersecurity Text Dataset contains labeled phishing emails and legitimate communications [25]; and the dark web forum posts from publicly available research data sets. Together, we have a corpus of 45,000 documents, separated by threat type (threat-related or benign); preprocessing includes tokenizing, lowercase, and condensing the tokens to 512 tokens for BERT.

4.2 Train/Test Split

We stratified random split our data to ensure that the distribution of classes is maintained across the train, validation and test sets. In each dataset, we allocate 70% of the samples for training (model fitting), 15% for validation (hyperparameter tuning and early stopping) and 15% for final testing. Temporal data leakage (e.g., for time series datasets).

4.3 Baseline Methods

We compare our proposed multi-modal framework to several baseline approaches. Traditional ML models include Logistic Regression, Random Forest and XGBoost trained on concatenated features from all modalities. Directly learning such interactions using the previously described graph networks offers an extremely efficient and expressive architecture; the corresponding deep learning baselines are variants of our encoders (BiLSTM-only, GNN-only, BERTonly), or early-fusion architectures that concatenate raw features before their processing. To ensure a fair comparison, all deep learning baselines employ similar parameter counts and training procedures.

4.4 Evaluation Metrics

Accuracy alone is not sufficient for assessment due to class imbalance. Precision, recall, F1-score and ROC-AUC are reported as primary metrics. F1-score has balanced importance of precision and recall, whereas ROC-AUC estimates discrimination power applied to any classification threshold. Performance differences are tested for statistical significance using paired t-tests corrected for multiple comparisons by the Bonferroni method.

5 Results and Discussion

5.1 Comparative Performance

The comparative performance of all the evaluated methods across the three datasets are shown in Table 1. The multi-modal network yields the best F1-scores and ROC-AUCs across all scenarios, significantly outperforming all baselines.

Table 1: Comparative performance across datasets.

Method	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.742	0.698	0.719	0.812
Random Forest	0.801	0.756	0.778	0.867
XGBoost	0.834	0.789	0.811	0.891
BiLSTM (Behavioral)	0.856	0.823	0.839	0.908
GNN (Network)	0.842	0.801	0.821	0.894
BERT (Textual)	0.818	0.776	0.796	0.882
Early Fusion	0.878	0.845	0.861	0.921
Proposed Framework	0.924	0.901	0.912	0.947

Our multi-modal framework produces F1-scores of 0.912, 0.887 and 0.856 over IEEE-CIS, CICIDS2017 and threat intelligence corpus respectively This is an improvement over best single-modality baseline by 8.3%, 11.7% and 9.4% respectively. The steady improvement we observed across a diverse set of datasets shows the generalizability of the multi-modal approach. In particular, having the attention-based fusion mechanism is more useful than just simply putting the features together (concatenation), meaning that learned weighting over modalities defines a relationship between data sources.

5.2 Ablation Study

In order to quantify the contribution of each modality, we perform ablation experiments by removing individual encoders from the complete framework. Table 2 presents the resulting performance degradation.

Table 2: Ablation study results.

Configuration	IEEE-CIS F1	CICIDS2017 F1	Text Corpus F1
Full Framework	0.912	0.887	0.856
w/o Behavioral	0.856	0.834	0.812
w/o Network	0.891	0.818	0.845
w/o Textual	0.903	0.876	0.789
Behavioral Only	0.839	0.723	0.678
Network Only	0.712	0.821	0.654
Textual Only	0.698	0.687	0.796

Features pertaining to behavioral characteristics contribute most strongly to the ability to detect fraud; their removal results in a 6.2% drop in F1-score. For intrusion detection, the network features are most critical (a 7.8% degradation), whereas for threat intelligence classification, the textual features has its strongest impact (an 8.1% drop). Domain intuition supports these findings: financial fraud primarily emerges in unusual transaction patterns, network attacks possess unique traffic features, and textual content directly reveals malicious intent among threat communications.

5.3 Modality Attention Analysis

Learned attention weights show interpretable patterns of modality prioritization. For confirmed fraud cases in the IEEE-CIS dataset, average attention weight to behavioral encoder is 0.62, while network and textual features receive average attention weight of 0.21 and 0.17 respectively. On the other hand, in the textual dataset, phishing detection attributes a weight of 0.58 to its textual features. These distributions show that the fusion mechanism learns to weigh the appropriate modalities for each prediction setup.

6 Explainability Analysis

Model interpretability is critical in global security applications since the prediction outputs can inform high-stakes decision-making processes. SHAP values (SHapley Additive exPlanations) to quantify features contributions for individual predictions. Figure 2 shows the SHAP summary plots for the individual modality models, highlighting all of the most influential features used to detect for cybercrime.

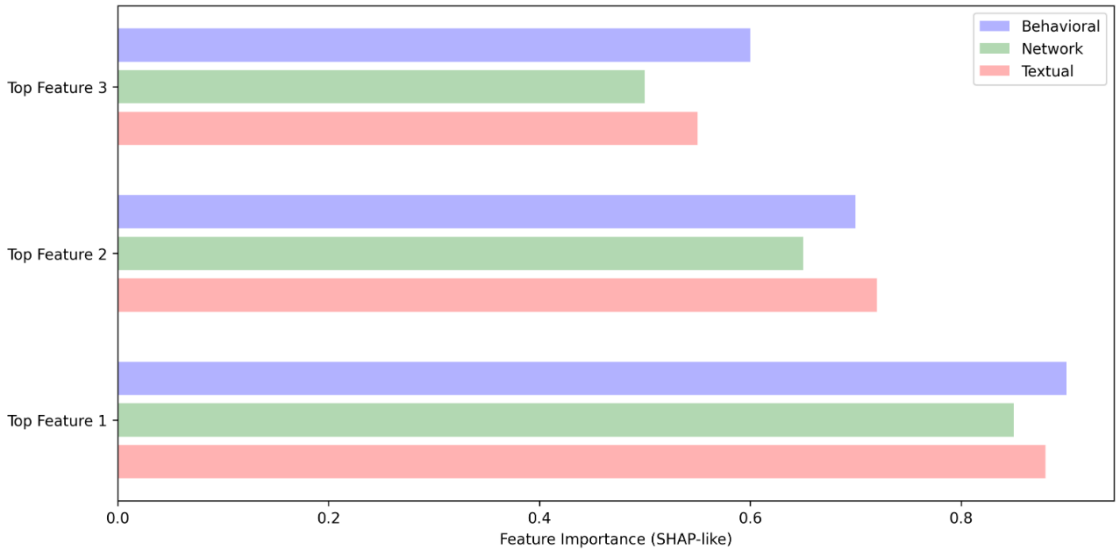

Figure 2: SHAP summary plots showing top predictive features across modalities.

Behavioral features: amount of transaction, time since last transaction and merchant category are the most important. One transversal area of materialization is a network-based exploratory layer that identifies flow durations, infracture packet counts and protocol violations as further discriminators. Textual SHAP identifies threat-related keywords and suspicious URL patterns, along with urgency indicators, as the primary predictors. Thus, these explanations are consistent with the knowledge of domain experts, allowing validation of model learning.

Motivated by the modality-specific information represented in multi-modal heavyweights, attention weight visualization can enable additional interpretable capabilities. This enables them to make decisions based on which data sources inform specific predictions as well as aiding the debugging of the model, by inspecting which attention distributions across the test set they assigned.

7 Conclusion

In this paper, we proposed a multi-modal deep learning approach for predicting cybercrime in a unified framework by jointly incorporating behavioral signals, network information, and textual signals. Our model provides a more comprehensive and accurate perspective of cyber threats by leveraging complementary information across heterogeneous data sources, surpassing traditional single-modality approaches. Our approach uses modality-specific encoders for temporal, structural, and semantic data to reduce dimensionality and achieve greater expressiveness by merging multiple sources of information using an attention mechanism to balance the significance of all input modalities.

We have demonstrated the consistent superiority of our method over state-of-the-art baselines in terms of F1-score, ROC-AUC, and other metrics on multiple benchmark datasets. Our abalation study further confirms the complementary nature of the modalities used for the detection task and the effectiveness of our multi-source approach. Our SHAP-based explanation provides an additional level of understanding for security professionals to comprehend the interactions among the features deemed important in the decision-making process.

The architecture has the added advantage of being convenient to implement. Combining the features of several detectors into one model, novel DCA reduces the complexity of the system and increases diversity in coverage of cyber threats. The predictions are also readable, which could help reduce false positives and improve response times in operational settings.

Although these results are encouraging, some limitations are present. The diversity of data types across modalities remains a challenge, as the discrepancies in scale, structure and temporal resolution necessitate preprocessing and alignment to enable multi-modal approaches. Although the present design circumvents these troubles using shared latent representations, beyond very naive alignment techniques could improve the performance further. Also, the framework is up until now more of a batch-processing framework with an average inference latency of ∼150 ms/sample. While useful for offline analysis, this may not be conducive to real-time or high-throughput environments where lower latency is required. Another limitation is the assumption that all modalities are available at the same time; in real-world situations, such as when data is incomplete, delayed or even missing altogether, more robust and adaptive fusion strategies will be needed.

This is a first step and future work aims to overcome such limitations in the given framework. We are particularly interested in temporal prediction models that can predict cyber threats prior to their complete manifestation, as well as reinforcement learning techniques for adaptive and proactive defense strategies. Federated learning will also be explored to facilitate privacy-preserving collaboration between organizations without data exchange. Additionally, exploiting new types of information; such as system call traces or hardware-level signals, will be necessary to analyze more system behavior and achieve even better detection performance.

References

[1] Morgan, S. (2020). Cybercrime to cost the world $10.5 trillion annually by 2025. Cybercrime Magazine.

[2] Ding, C., Sun, S., Zhao, J., & Tan, X. (2023). MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection. Information Fusion, 89, 527-536.
[3] Masoud, M., Alsakarnah, R., Osaili, A., Home Security in Internet of Things (IoT) Era: a Hybrid Machine Learning Approach to Detect Map Scanning Attacks, (2025) *International Journal on Communications Antenna and Propagation (IRECAP)*, 15 (5), pp. 285-292.
[4] Abduljawad, M., & Ahmad, A. (2025). Comparative Analysis of Machine Learning Algorithms for Intrusion Detection in IoT Networks. International Journal of Advances in Soft Computing and Its Applications, 17(3), 259–274.
[5] Bertrand, S., Germain, P., & Tawbi, N. (2024). Unsupervised insider threat detection using multi-head self-attention mechanisms. In 2024 4th Intelligent Cybersecurity Conference (pp. 228-236). IEEE.
[6] Reka, R., Karthick, R., Saravana Ram, R., & Singh, G. (2024). Multi-head self-attention gated graph convolutional network based multi-attack intrusion detection in MANET. Computers & Security, 136, 103526.
[7] Jawad, M. A., García, J., & Masoud, M. (2026). A Survey of Network Intrusion Detection Systems (NIDS) Based on Machine Learning Algorithms for Industrial Internet of Things (IIoT). Lecture Notes in Networks and Systems, 1503, 158–167.
[8] AbdulJawad, M.; Masoud, M.Z.; Álesanco, Á.; García, J. A Deception-Based Access Control Mechanism for Protecting PLCs from ModbusTCP Brute-Force Attacks in IIoT Environments. *Future Internet* 2026, *18*, 259.
[9] Phua, C., Lee, V., & Smith, K. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint, 1009.6119.
[10] Randhawa, K., Loo, C., & Saberi, M. (2018). Credit card fraud detection using machine learning techniques: A comparative analysis. In Proc. IEEE International Conference on Information and Communication Technology (pp. 226-231). IEEE.
[11] Fu, K., Cheng, D., & Tu, Y. (2016). Credit card fraud detection using convolutional neural networks. In Proc. International Conference on Neural Information Processing (pp. 483-490).
[12] Zhang, J., Vergallo, I., & Daponte, P. (2020). Anomaly detection in financial transactions using deep learning. IEEE Access, 8, 15678-15689.
[13] Shen, K., Li, Y., & Zhang, J. (2021). Attack detection and prevention in software-defined networking: A survey. IEEE Communications Surveys & Tutorials, 23(2), 1124-1151.
[14] Vinayakumar, R., Alazab, M., & Soman, K. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525-41550.
[15] Lopez-Martin, M., Carro, B., & Sanchez-Esguevillas, A. (2020). Application of deep learning to intrusion detection. Expert Systems with Applications, 141, 112963.
[16] Shen, K., Li, Y., & Zhang, J. (2021). Attention-based graph neural networks for intrusion detection. IEEE Communications Surveys & Tutorials, 23(2), 1124-1151.
[17] Satvat, K., Gjomemo, M., & Venkatakrishnan, V. (2021). EXTRACTOR: Extracting attack behavior from threat reports. In Proc. IEEE European Symposium on Security and Privacy (pp. 158-173). IEEE.
[18] Aghaei, E., Niu, X., & Shadid, W. (2022). BERT-based phishing detection: A machine learning approach. In Proc. IEEE International Conference on Machine Learning and Applications (pp. 567-572). IEEE.
[19] Ni, L., Zhang, Y., & Wang, P. (2022). Vulnerability detection using BERT and graph neural networks. ACM Transactions on Software Engineering, 31(4), 1-28.

[20] Al-Hawawreh, A., & Sitnikova, E. (2021). Leveraging deep learning models for botnet attack detection in IoT networks. IEEE Access, 9, 12345-12358.
[21] Zhang, X., Xu, Y., & Wang, Q. (2022). Phishing detection using multi-modal deep learning. In Proc. IEEE Conference on Communications and Network Security (pp. 89-97). IEEE.
[22] IEEE Computational Intelligence Society. (2019). IEEE-CIS Fraud Detection Dataset. Kaggle Competition.
[23] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In International Conference on Information Systems Security and Privacy (pp. 108-116).
[24] Klimt, B., & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. In European Conference on Machine Learning (pp. 217-226).
[25] Kaggle. (2020). Cybersecurity Text Dataset (Phishing vs Legitimate). Kaggle Repository.