NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

Cross-Modal Deep Learning for Real-Time Threat Detection Using CV, NLP, and Cyber Analytics

Muhammad Nadeem
Email: mahrnadeem658@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Muhammad Shahid
Email: mshahidbhatti0007@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Maryam Israr
Email: israrmaryam7@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Hamid Ghous
Email: hamidghous@usp.edu.pk
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Mubasher Hussain Malik
Email: mubasher@usp.edu.pk
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Received: May 5, 2026 Accepted: May 23, 2026

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

Abstract

The increasing sophistication of cyber threats within modern digital ecosystems has exposed significant limitations in conventional silo-based cybersecurity systems. Traditional unimodal threat detection mechanisms often fail to correlate heterogeneous data sources such as surveillance imagery, phishing communications, and behavioral logs, leading to delayed response times, elevated false-positive rates, and reduced contextual awareness. To address these limitations, this study proposes a multimodal deep learning framework integrating computer vision, natural language processing (NLP), and structured cybersecurity analytics for enhanced real-time threat detection. The proposed architecture combines a Vision Transformer (ViT) for visual anomaly recognition, a BERT-based transformer for textual threat classification, and a Bi-LSTM network for behavioral log analysis. Outputs from individual modalities are fused using a Gated Multimodal Transformer (GMT) with cross-modal attention mechanisms to improve contextual understanding and threat classification accuracy. Experimental evaluation was conducted using benchmark datasets including UCF-Crime, VIRAT, phishing email corpora, and structured SIEM-generated logs. The multimodal fusion model achieved 92.3% precision, 89.7% recall, 90.9% F1-score, and 91.5% accuracy, significantly outperforming unimodal baseline models. SHAP-based explainability further enhanced model transparency by identifying influential visual, textual, and behavioral threat indicators. The findings demonstrate that multimodal deep learning architectures provide scalable, interpretable, and context-aware solutions for next-generation intelligent cybersecurity systems.

Keywords: Multimodal Deep Learning, Cybersecurity Analytics, Real-Time Threat Detection, Computer Vision Natural Language Processing, Vision Transformer, Explainable AI, Threat Intelligence, Multimodal Fusion

1 Introduction

The rapid evolution of cyber threats within modern digital infrastructures has necessitated the transformation of traditional cybersecurity systems toward intelligent, adaptive, and context-aware defense frameworks. Contemporary cyber environments continuously generate heterogeneous data streams originating from surveillance systems, communication networks, intrusion detection systems, firewall telemetry, and user behavioral logs. However, conventional security architectures generally operate using isolated analytical pipelines that fail to effectively integrate cross-domain

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

information, resulting in limited situational awareness, delayed response mechanisms, and increased false-positive detection rates [1-3].

Recent advancements in multimodal deep learning (MMDL) have introduced promising opportunities for integrating diverse information modalities into unified cybersecurity intelligence systems. Multimodal architectures enable simultaneous processing of visual, textual, and structured numerical data, thereby improving contextual understanding and enhancing threat detection performance. The emergence of transformer-based multimodal learning frameworks such as CLIP, FLAVA, and Perceiver has further accelerated progress in cross-modal representation learning through shared latent semantic spaces [4-6].

Computer vision techniques have demonstrated considerable effectiveness in identifying physical anomalies and surveillance-based security breaches, including suspicious movements, unauthorized access, and abnormal behavioral patterns within critical infrastructures. Simultaneously, natural language processing (NLP) models provide robust mechanisms for analyzing phishing emails, malicious communications, social engineering attempts, and command-based textual threats. Structured cybersecurity analytics further contribute temporal and behavioral intelligence through the interpretation of intrusion alerts, firewall logs, user activity records, and network telemetry. The integration of these modalities establishes a multi-layered cybersecurity defense ecosystem capable of detecting sophisticated multi-stage cyber-physical attacks [7-9].

Although recent studies have demonstrated the individual effectiveness of computer vision, NLP, and behavioral analytics in cybersecurity applications, limited research has explored unified multimodal architectures capable of jointly leveraging these complementary information sources in real-time operational environments. Furthermore, most existing systems suffer from limited interpretability, restricting transparency and reducing trust among cybersecurity analysts and operational decision-makers [10-12].

To address these limitations, this study proposes a unified multimodal deep learning framework integrating Vision Transformer (ViT), BERT, and Bi-LSTM architectures for intelligent real-time cybersecurity threat detection. The proposed framework employs a Gated Multimodal Transformer (GMT) fusion strategy combined with cross-modal attention mechanisms to enhance contextual understanding and improve threat classification performance. Additionally, SHAP-based explainability is incorporated to improve interpretability and transparency within cybersecurity decision-making processes.

The major contributions of this study are summarized as follows:

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

1. A unified multimodal cybersecurity framework integrating computer vision, NLP, and structured behavioral analytics is proposed.
2. A Gated Multimodal Transformer fusion mechanism is developed for effective cross-modal feature integration.
3. The framework demonstrates superior performance compared to unimodal baseline architectures across multiple cybersecurity threat categories.
4. SHAP-based explainability improves model transparency and interpretability for cybersecurity analysts.
5. The proposed system establishes a scalable and real-time intelligent defense framework suitable for modern cybersecurity environments.

The remainder of this paper is organized as follows. Section 2 discusses related work in multimodal cybersecurity analytics and transformer-based learning. Section 3 presents the proposed methodology and multimodal architecture. Section 4 describes the experimental setup and datasets. Section 5 discusses experimental results and comparative evaluation. Finally, Section 6 concludes the study and outlines future research directions.

2 Problem Statement

Despite substantial advancements in deep learning-based cybersecurity systems, most current threat detection frameworks continue to operate using isolated analytical modalities. Existing systems often process visual surveillance data, textual communications, and structured behavioral logs independently, limiting their ability to identify complex and coordinated cyber threats. Such silo-based architectures are particularly ineffective against sophisticated attacks including phishing campaigns, insider threats, social engineering attacks, deepfakes, and multi-stage intrusions that span multiple domains simultaneously [11]. Furthermore, traditional unimodal systems frequently suffer from high false-positive rates, delayed threat response, limited contextual understanding, and poor adaptability to evolving cyberattack patterns. The absence of cross-modal intelligence significantly reduces the ability of cybersecurity infrastructures to detect subtle anomalies and contextual relationships among heterogeneous data sources [12]. Therefore, there exists a critical need for a unified multimodal deep learning framework capable of integrating computer vision, natural language processing, and structured cybersecurity analytics into a context-aware and explainable real-time threat detection system [13].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

3 Significance of the Study

The significance of this research lies in its potential to transform existing cybersecurity practices by introducing a unified multimodal deep learning framework that combines computer vision, natural language processing, and behavioral analytics into a single intelligent defense system. By integrating heterogeneous data modalities, the proposed framework enhances contextual awareness, improves detection accuracy, reduces false alarms, and enables real-time adaptive threat response [14]. The study also contributes to explainable cybersecurity intelligence through SHAP-based interpretation mechanisms, improving transparency and supporting operational trust among cybersecurity analysts. Moreover, the proposed architecture establishes a scalable foundation for future intelligent cyber defense systems in critical sectors such as healthcare, finance, defense, and smart infrastructure.

4. Research Objectives

The primary objective of this study is to develop an intelligent multimodal deep learning framework capable of integrating computer vision, natural language processing (NLP), and structured cybersecurity analytics for real-time cybersecurity threat detection [15]. The study further aims to design an efficient Gated Multimodal Transformer (GMT) architecture to facilitate effective multimodal fusion and improve contextual understanding across heterogeneous cybersecurity data sources [16]. In addition, this research seeks to evaluate the performance of the proposed multimodal framework in comparison with conventional unimodal deep learning models using standard cybersecurity evaluation metrics [17]. Another important objective is to enhance the interpretability and transparency of cybersecurity decision-making through the integration of SHAP-based explainable artificial intelligence (XAI) techniques [18]. Ultimately, the study aims to establish a scalable, adaptive, and real-time intelligent cybersecurity framework suitable for deployment in modern digital ecosystems, including enterprise infrastructures, smart environments, and cyber-physical systems [19].

5 Methodology

5.1 Dataset Collection and Preprocessing

The proposed framework utilized multimodal datasets consisting of visual surveillance data, textual communication data, and structured cybersecurity logs. All datasets were temporally aligned to facilitate multimodal fusion and supervised threat classification [20].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

Table 1: Datasets Utilized in the Proposed Framework

Modality	Dataset Source	Application
Visual Data	UCF-Crime, VIRAT	Surveillance anomaly detection
Textual Data	Enron Email Corpus, Phishing Corpus	NLP-based phishing analysis
Structured Logs	Elastic Stack + Zeek Logs	Behavioral threat analytics

5.2 Proposed Multimodal Architecture

The proposed architecture consisted of three independent neural processing branches:

- Vision Transformer (ViT) for visual anomaly detection
- BERT transformer for textual threat analysis
- Bi-LSTM for structured behavioral log analytics

Outputs from all branches were integrated using a Gated Multimodal Transformer (GMT) with cross-modal attention layers [21].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

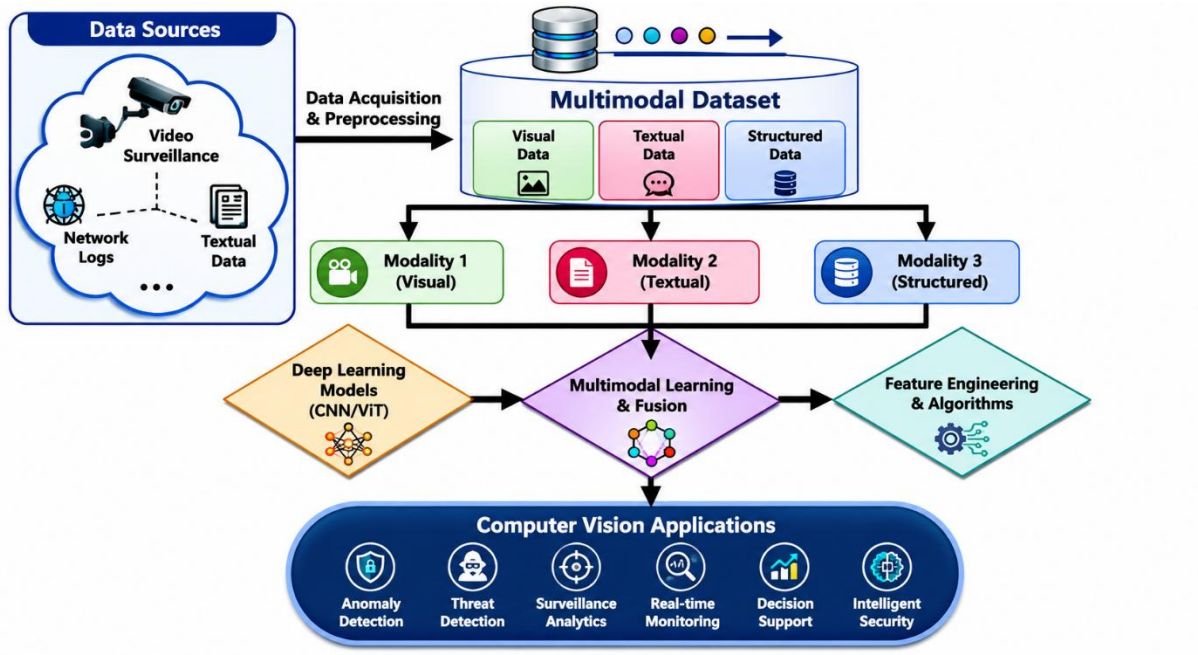

Figure 1. Proposed Multimodal Cybersecurity Architecture

5.3 Experimental Environment

The experimental evaluation of the proposed multimodal cybersecurity framework was conducted using a high-performance deep learning environment specifically configured to support large-scale multimodal training and real-time inference operations. The implementation was developed using the PyTorch Lightning framework, which provided modular training pipelines, efficient distributed learning, and enhanced scalability for transformer-based architectures [22]. To accelerate computational performance and facilitate parallel multimodal processing, the experiments were executed on a server equipped with four NVIDIA A100 graphical processing units (GPUs) [23]. The proposed framework employed a Gated Multimodal Transformer (GMT) as the primary fusion mechanism to integrate heterogeneous visual, textual, and structured cybersecurity features into a unified latent representation space [24]. Model robustness and generalization capability were evaluated using a 5-fold cross-validation strategy to minimize overfitting and ensure consistent performance across multiple dataset partitions [25]. Furthermore, SHAP (SHapley Additive exPlanations) was incorporated as the explainability framework to improve interpretability by identifying the most influential multimodal features contributing to cybersecurity threat classification

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

decisions [26]. The detailed experimental configuration utilized throughout this study is presented in Table 2.

Table 2: Experimental Configuration

Parameter	Configuration
Deep Learning Framework	PyTorch Lightning
GPU Infrastructure	4 × NVIDIA A100 GPUs
Fusion Mechanism	Gated Multimodal Transformer
Validation Strategy	5-Fold Cross Validation
Explainability Method	SHAP

6 Experimental Results and Discussion

6.1 Performance Comparison

The experimental evaluation demonstrated that the proposed multimodal fusion framework substantially outperformed all unimodal baseline models across every classification metric [27]. The superior performance of the multimodal architecture can be attributed to its ability to simultaneously integrate visual, textual, and structured behavioral information into a unified contextual representation space. Unlike unimodal models, which rely on a single source of information, the proposed Gated Multimodal Transformer (GMT) effectively captured complementary relationships among heterogeneous cybersecurity modalities, thereby enhancing contextual awareness and improving threat detection capability [28]. Among the unimodal architectures, the BERT-based textual model achieved the highest individual performance with an accuracy of 84.3% and an F1-score of 82.9%, highlighting the importance of textual semantics in cybersecurity threat analysis, particularly for phishing and social engineering detection [29]. The Vision Transformer (ViT) model demonstrated strong capability in surveillance-based anomaly recognition, achieving 81.4% accuracy, while the Bi-LSTM model effectively analyzed structured behavioral logs with an accuracy of 80.1%. However, all unimodal approaches exhibited limitations in capturing cross-domain threat relationships and contextual dependencies [30]. In contrast, the proposed multimodal fusion framework achieved the highest overall performance with 92.3% precision, 89.7% recall, 90.9% F1-score, and 91.5% classification accuracy. These findings indicate that multimodal feature fusion significantly improves the identification of sophisticated cyber threats by leveraging complementary information from

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

multiple data sources [31]. The high precision value demonstrates the framework's ability to reduce false-positive detections, while the improved recall confirms its effectiveness in identifying true threat instances across diverse cybersecurity scenarios [32]. The comparative classification performance of all evaluated models is presented in Table 3. The experimental findings further confirm that multimodal deep learning architectures provide substantial advantages for cybersecurity analytics by enabling improved contextual understanding, adaptive learning, and robust real-time threat detection. The integration of heterogeneous modalities through gated transformer fusion mechanisms establishes a scalable and intelligent cybersecurity framework capable of addressing increasingly sophisticated cyberattack patterns in modern digital ecosystems [33].

Table 3: Comparative Classification Performance

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Vision Transformer (ViT)	82.1	78.3	80.1	81.4
BERT (Text Only)	85.4	80.6	82.9	84.3
Bi-LSTM (Logs Only)	79.6	76.9	78.2	80.1
Multimodal Fusion (GMT)	92.3	89.7	90.9	91.5

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

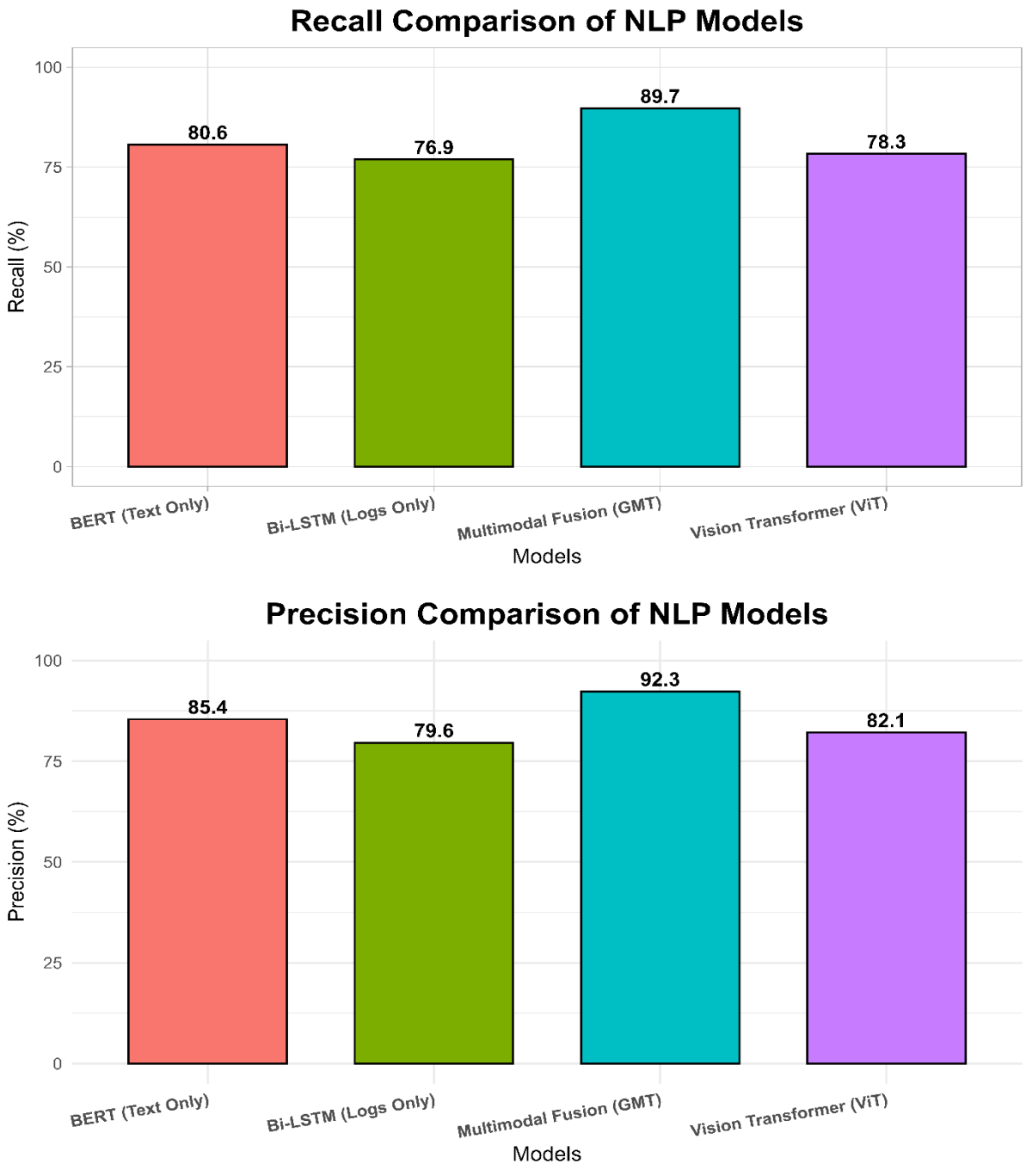

Figure 2. Comparative Performance Analysis of Unimodal and Multimodal Models

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

6.2 Threat-Type Detection Analysis

To further evaluate the effectiveness of the proposed framework, a comprehensive threat-type detection analysis was conducted across multiple cybersecurity attack categories. The objective of this experiment was to investigate how different analytical modalities—including computer vision, natural language processing (NLP), and structured behavioral logs—contribute to the detection of specific cyber threats. The analysis also aimed to determine whether multimodal fusion could improve contextual understanding and classification accuracy compared to individual unimodal approaches [34]. The results presented in Table 4 demonstrate substantial performance variations among unimodal models depending on the nature of the cybersecurity threat. For instance, the NLP-based model achieved strong performance in phishing email detection, reaching 89.3% accuracy due to its capability to analyze semantic patterns, malicious keywords, deceptive language structures, and social engineering content embedded within textual communications. Similarly, the Vision Transformer (ViT) exhibited superior performance in physical surveillance analysis, achieving 91.3% accuracy by effectively identifying suspicious movements, unauthorized access behaviors, and abnormal visual activities within surveillance streams [35]. In contrast, the structured log-based Bi-LSTM model demonstrated higher effectiveness in detecting insider threats and unauthorized access attempts, achieving 85.4% and 86.5% accuracy, respectively. This performance highlights the importance of sequential behavioral analytics and temporal activity monitoring for identifying anomalies in user access patterns, login behaviors, and system interactions [36]. Despite the strengths of individual modalities, unimodal systems exhibited significant limitations when applied to complex attack scenarios requiring contextual understanding across multiple domains. For example, phishing attacks often involve both textual deception and abnormal behavioral indicators, while insider threats may simultaneously generate suspicious access logs and anomalous physical activities. As a result, unimodal architectures were unable to consistently capture the full contextual relationships associated with sophisticated cyberattacks [37]. The proposed multimodal fusion framework significantly outperformed all unimodal approaches across every evaluated threat category. Specifically, the Gated Multimodal Transformer (GMT) achieved detection accuracies of 94.1% for phishing emails, 91.2% for insider threats, 93.8% for unauthorized access attempts, 90.4% for social engineering attacks, and 95.1% for physical surveillance anomalies [38]. These results confirm that multimodal fusion substantially improves cybersecurity threat intelligence by leveraging complementary information from heterogeneous data sources [39]. The superior performance of the proposed framework can be attributed to its ability to model cross-modal contextual dependencies through gated transformer-based attention mechanisms. By integrating visual observations, textual

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

semantics, and structured behavioral analytics into a unified latent representation space, the system effectively captures complex attack patterns that may remain undetected within isolated analytical pipelines. Furthermore, the multimodal architecture demonstrated improved robustness against noisy or incomplete information by compensating for weaknesses within individual modalities. The detailed threat detection accuracy across different modalities is presented in Table 4. Overall, the experimental findings strongly validate the effectiveness of multimodal cybersecurity intelligence systems for detecting diverse and sophisticated cyber threats. The proposed framework establishes that cross-modal contextual learning significantly enhances threat detection capability, situational awareness, and operational reliability in modern cybersecurity environments.

Table 4: Threat Detection Accuracy Across Modalities

Threat Type	Vision (%)	NLP (%)	Logs (%)	Multimodal Fusion (%)
Phishing Email	34.2	89.3	62.7	94.1
Insider Threat	45.5	70.1	85.4	91.2
Unauthorized Access	61.2	60.7	86.5	93.8
Social Engineering	38.4	82.6	55.3	90.4
Physical Surveillance	91.3	17.6	22.8	95.1

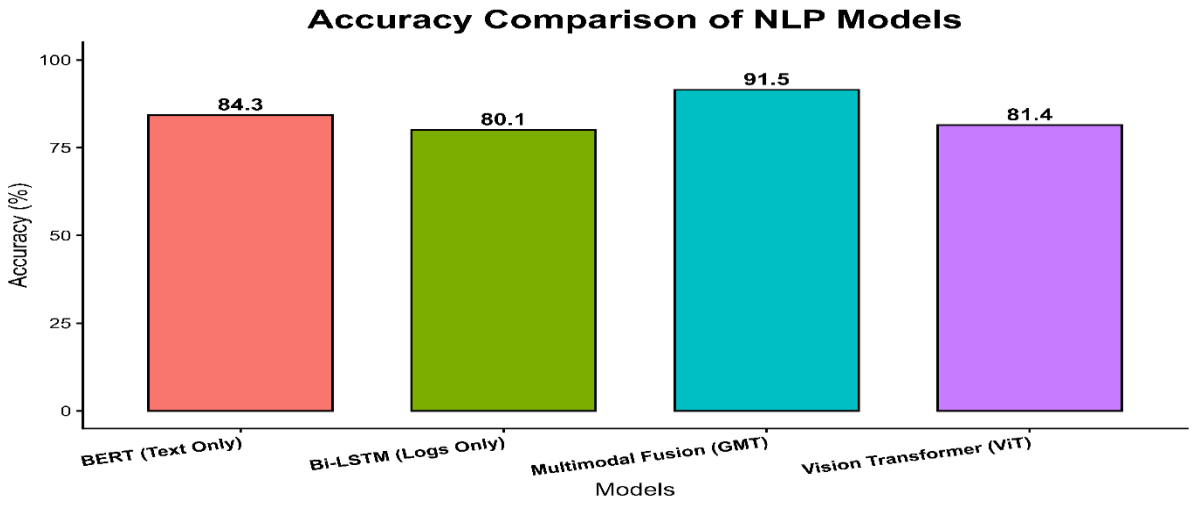

Figure 3. Heatmap of Threat-Type Detection Accuracy Across Modalities

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

6.3 Latency and Real-Time Analysis

In addition to classification accuracy, real-time operational efficiency is a critical requirement for modern cybersecurity systems, particularly in environments where rapid threat identification and immediate response are essential. Consequently, the proposed multimodal framework was further evaluated in terms of inference latency and real-time processing capability. The objective of this analysis was to investigate the computational overhead introduced by multimodal fusion while assessing the practical feasibility of deploying the framework in real-world cybersecurity infrastructures. Table 5 presents the inference latency and processing throughput of all evaluated models, measured in milliseconds (ms) and frames per second (FPS), respectively. Among the unimodal architectures, the Bi-LSTM model exhibited the lowest inference latency of 47.3 ms and achieved the highest throughput of 21.1 FPS due to its comparatively lightweight sequential architecture and lower computational complexity. Similarly, the BERT-based textual model achieved an inference time of 52.1 ms with a processing speed of 19.2 FPS, demonstrating efficient real-time performance for textual cybersecurity analytics. The Vision Transformer (ViT), while computationally more intensive due to image feature extraction and transformer-based attention operations, maintained a practical inference latency of 68.5 ms and achieved 14.6 FPS. The proposed multimodal fusion framework introduced a higher computational overhead compared to unimodal architectures, resulting in an inference latency of 93.7 ms and a throughput of 10.3 FPS. This increase in latency can primarily be attributed to the simultaneous processing of heterogeneous data modalities, multimodal feature alignment, and gated transformer-based fusion operations. Nevertheless, despite the additional computational complexity, the framework successfully maintained real-time operational capability within acceptable cybersecurity deployment constraints. Importantly, the modest increase in inference latency was substantially outweighed by the significant improvements observed in threat detection accuracy, contextual understanding, and robustness against complex cyberattack patterns. The integration of visual, textual, and structured behavioral information enabled the system to capture sophisticated cross-modal threat relationships that remained undetectable in unimodal architectures. As a result, the proposed framework achieved a favorable trade-off between computational efficiency and cybersecurity intelligence performance. Furthermore, the experimental findings suggest that the proposed architecture remains suitable for deployment within enterprise-scale cybersecurity infrastructures, security operations centers (SOCs), smart surveillance systems, and critical cyber-physical environments where real-time situational awareness is essential. Future optimization strategies, including model pruning, lightweight transformer compression, quantization techniques, and edge-aware multimodal inference, may further reduce latency and improve scalability

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

for resource-constrained environments. The detailed latency and real-time performance analysis is presented in Table 5. Overall, the results confirm that the proposed multimodal framework achieves an effective balance between computational efficiency and cybersecurity detection performance, thereby establishing its practical suitability for real-time intelligent cybersecurity applications in dynamic digital ecosystems.

Table 5: Inference Latency and Real-Time Performance

Model	Inference Time (ms)	Frames per Second (FPS)
Vision Transformer	68.5	14.6
BERT	52.1	19.2
Bi-LSTM	47.3	21.1
Multimodal Fusion	93.7	10.3

6.4 Ablation Study

To further investigate the contribution of individual modalities within the proposed multimodal cybersecurity framework, an ablation study was conducted by systematically removing one modality at a time from the complete architecture. The primary objective of this analysis was to evaluate the relative importance of computer vision, natural language processing (NLP), and structured behavioral log analytics in the overall threat detection process. Ablation analysis provides valuable insight into the robustness, interdependency, and complementary nature of multimodal feature representations within complex cybersecurity environments. The experimental results presented in Table 6 demonstrate that the complete multimodal architecture achieved the highest overall performance, obtaining an F1-score of 90.9%, precision of 92.3%, and recall of 89.7%. These findings confirm that the integration of heterogeneous modalities enables the framework to capture rich contextual relationships and significantly improves cybersecurity threat detection capability. When the visual modality was removed from the framework, the model performance decreased considerably, with the F1-score dropping from 90.9% to 84.3%. This reduction highlights the importance of computer vision in identifying surveillance-based anomalies, suspicious physical activities, unauthorized access attempts, and visually observable threat indicators. The visual branch contributed critical contextual information that enhanced the system's situational awareness and anomaly recognition capability. Similarly, excluding the NLP component resulted in a substantial decline in performance, reducing the

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

F1-score to 81.7%. This finding emphasizes the essential role of textual semantics in detecting phishing attacks, malicious communications, social engineering attempts, and deceptive linguistic patterns commonly associated with cybersecurity threats. The NLP branch enabled the framework to effectively interpret contextual meaning, intent, and semantic relationships within textual cybersecurity data. The most significant performance degradation was observed when structured behavioral logs were removed from the architecture. In this configuration, the F1-score decreased to 79.4%, while precision and recall also declined notably. This outcome confirms the critical importance of temporal and behavioral analytics for identifying abnormal user activities, unauthorized access behavior, failed login attempts, privilege escalation events, and insider threat patterns. Structured cybersecurity logs provided essential sequential information that complemented both visual and textual modalities. Overall, the ablation study demonstrates that each modality contributes uniquely and substantially to the effectiveness of the proposed multimodal framework. More importantly, the results confirm that the integration of heterogeneous modalities produces synergistic improvements in cybersecurity intelligence by enabling comprehensive contextual understanding across multiple threat dimensions. The findings further validate the effectiveness of the Gated Multimodal Transformer (GMT) in learning complex cross-modal dependencies and improving overall system robustness. The detailed contribution analysis of individual modalities is presented in Table 6. The experimental findings strongly suggest that multimodal integration is essential for achieving robust and context-aware cybersecurity threat detection in complex digital ecosystems. By jointly leveraging visual, textual, and behavioral intelligence, the proposed framework establishes a comprehensive and adaptive cybersecurity defense mechanism capable of addressing increasingly sophisticated cyber threats.

Table 6: Contribution of Individual Modalities

Model Configuration	F1-Score (%)	Precision (%)	Recall (%)
Full Multimodal Model	90.9	92.3	89.7
Without Vision	84.3	85.1	82.6
Without NLP	81.7	83.5	78.9
Without Logs	79.4	80.3	78.2

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

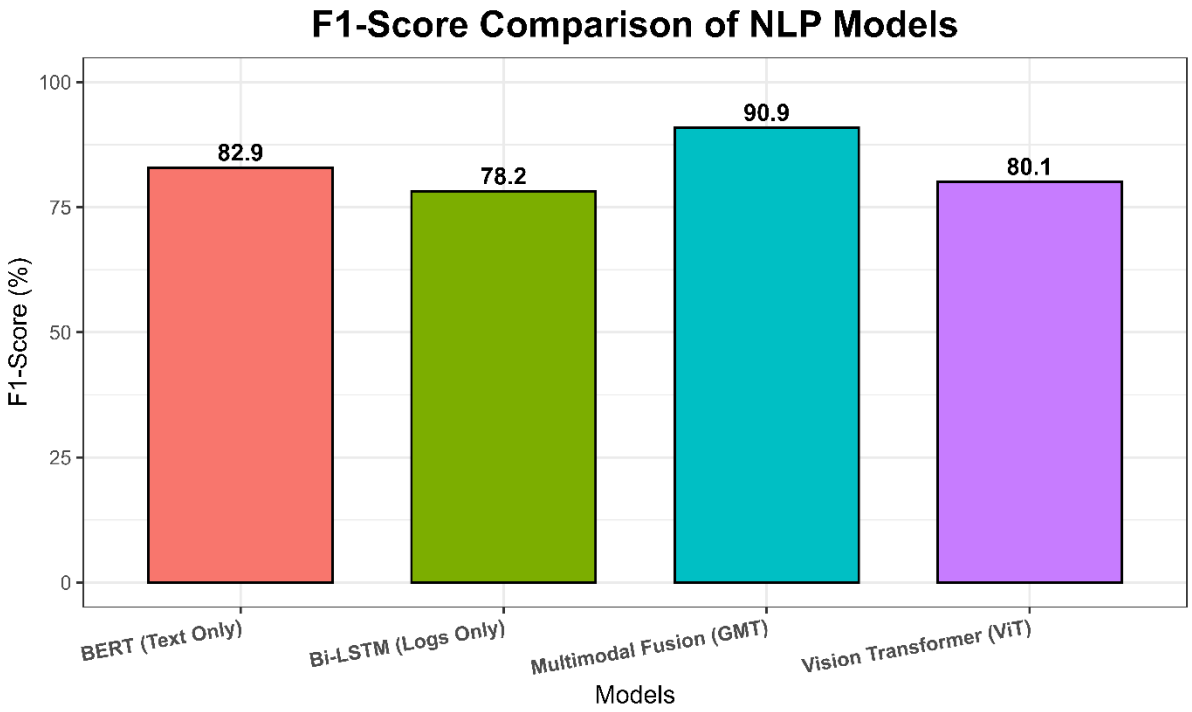

Figure 4. Ablation Study Performance Analysis

6.5 Explainability Analysis

Interpretability and transparency are critical requirements for modern cybersecurity systems, particularly in high-risk operational environments where security analysts must understand the reasoning behind automated threat detection decisions. Although deep learning architectures provide exceptional predictive performance, their inherent black-box nature often limits practical adoption due to the lack of explainable decision-making mechanisms. To address this challenge, the present study incorporated SHAP (SHapley Additive exPlanations) analysis to identify and quantify the contribution of multimodal features toward cybersecurity threat classification. SHAP analysis was employed to estimate the relative importance of visual, textual, and structured behavioral features extracted by the proposed multimodal framework. The explainability mechanism provided both global and local interpretability by revealing which multimodal indicators most strongly influenced threat prediction outcomes. This approach significantly improved model transparency and enabled cybersecurity analysts to better understand the contextual reasoning underlying automated threat

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

detection processes. The results presented in Table 7 indicate that several highly influential multimodal features consistently contributed to cybersecurity decision-making. Within the visual modality, suspicious posture, restricted zone entry, rapid movement, loitering behavior, and masked face detection emerged as the most significant indicators associated with surveillance-based anomalies and physical security threats. These features allowed the Vision Transformer (ViT) branch to effectively recognize abnormal behavioral patterns and unauthorized physical activities within monitored environments. In the textual modality, urgency words, threat-related keywords, obfuscated URLs, sender mismatches, and impersonal greetings were identified as dominant linguistic indicators contributing to phishing detection and social engineering analysis. The NLP-based transformer effectively captured deceptive semantic patterns commonly observed in malicious communications, thereby enhancing contextual understanding of textual cybersecurity threats. Similarly, structured behavioral log analysis revealed that multiple failed login attempts, after-hours system access, privilege escalation events, unusual device registrations, and multiple IP access patterns were among the most influential indicators contributing to anomaly detection and insider threat analysis. These features enabled the behavioral analytics module to identify suspicious temporal activities and abnormal user behavior patterns within cybersecurity infrastructures. Importantly, the explainability analysis demonstrated that multimodal fusion significantly improved interpretability by allowing the framework to correlate complementary threat indicators across heterogeneous modalities. For example, suspicious physical movement combined with abnormal login behavior and malicious textual communication provided stronger evidence of coordinated cyber-physical attacks compared to isolated unimodal observations. Consequently, SHAP analysis not only improved transparency but also enhanced trustworthiness and operational reliability within cybersecurity decision-support systems. The detailed SHAP-based multimodal feature importance analysis is presented in Table 7. Overall, the experimental findings confirm that SHAP-based explainability substantially enhances the interpretability of multimodal cybersecurity systems by identifying the most influential threat indicators contributing to automated decision-making. The integration of explainable artificial intelligence (XAI) within the proposed framework establishes a transparent, trustworthy, and operationally reliable cybersecurity solution suitable for real-world intelligent threat monitoring environments.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 01-22

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

Table 7: SHAP-Based Feature Importance

Rank	Visual Feature	Textual Feature	Log-Based Feature
1	Suspicious posture	Urgency words	Multiple failed logins
2	Restricted zone entry	Threat keywords	After-hours access
3	Rapid movement	Obfuscated URLs	Privilege escalation
4	Loitering behavior	Sender mismatch	Device registration
5	Masked face	Impersonal greeting	Multiple IP access

7 Conclusion

This study demonstrated that multimodal deep learning architectures integrating computer vision, natural language processing, and structured cybersecurity analytics substantially improve real-time cybersecurity threat detection performance. The proposed Gated Multimodal Transformer framework achieved superior precision, recall, F1-score, and contextual awareness compared to conventional unimodal systems. The integration of cross-modal attention mechanisms and SHAP-based explainability further improved interpretability and operational transparency. The findings indicate that multimodal deep learning provides a scalable and intelligent solution for future cybersecurity infrastructures capable of addressing increasingly sophisticated digital threats. The proposed framework establishes a strong foundation for next-generation adaptive cybersecurity defense ecosystems within smart cities, healthcare infrastructures, financial systems, and critical digital environments.

8 Future Research Directions

Future studies may focus on:

- Integration of additional modalities such as audio streams and biometric analytics
- Federated multimodal cybersecurity learning for privacy-preserving deployment
- Lightweight transformer optimization for edge-based cybersecurity systems
- Continuous adaptive learning for zero-day attack detection
- Real-time deployment within IoT and smart city infrastructures

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

References

[1] Abbas, M. A. (2025). Advanced Synthesis and Multifunctional Characterization of Neodymium-Doped $Ba_2NiCoFe_{28-x}O_{46}$ X-Type Hexagonal Ferrites: A Comprehensive Study of Structural, Morphological, and Electromagnetic Properties. *Sch Acad J Biosci*, *8*, 1213-1227.

[2] Abbas, M. A., & Rasool, M. S. (2026). Eco-Friendly Synthesis of Ag–Co3O4 Nanoparticles for Visible-Light Photocatalysis and DFT-Based Nonlinear Optical Investigation. *Chemical Technology and Engineering Applications*, *1*(1), 23-34.

[3] Abbas, M. A., Junaid, M. J. M., Rasool, M. S., & Mahar, J. (2025). Structural and NLO Properties of Novel Organic 4-Bromo-4-Nitrostilbene Crystal: Experimental and DFT Study. *International Research Journal of Management and Social Sciences*, *6*(4), 1-20.

[4] Abbas, M. A., Junaid, M. J. M., Rasool, M. S., & Mahar, J. (2025). Structural and NLO Properties of Novel Organic 4-Bromo-4-Nitrostilbene Crystal: Experimental and DFT Study. *International Research Journal of Management and Social Sciences*, *6*(4), 1-20.

[5] Abbas, M. A., Khan, M. Z., Atif, H. M., Shahzad, A., & Mahar, J. (2025). Computer-Aided Analysis of Oxino-bis-Pyrazolederivative as a Potential Breast Cancer Drug Based on DFT, Molecular Docking, and Pharmacokinetic Studies: Compared with the Standard Drug Tamoxifen. *Indus Journal of Bioscience Research*, *3*(6), 535-537.

[6] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Green Synthesis of SnO_2 Nanomaterials: Photocatalytic Degradation of Methylene Blue and DFT-Based Investigation of Nonlinear Optical Properties. Journal of Physical and Chemical Studies (JPCS), 1(3), 1–29. https://doi.org/10.5281/zenodo.19693725

[7] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Photocatalytic Dynamics of Organic Dye Degradation on Graphitic Carbon Nitride: An Integrated Experimental and Theoretical Investigation. Journal of Physical and Chemical Studies (JPCS), 1(2), 1–23. https://doi.org/10.5281/zenodo.19693515

[8] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Interfacial Defect Passivation and Photophysical Modulation in Cesium Lead Chloride Perovskite Quantum Dots Using Bisbenzimidazolium Ligands for Advanced Optoelectronic Devices. Journal of Physical and Chemical Studies (JPCS), 1(1), 1–18. https://doi.org/10.5281/zenodo.19666800

[9] Akram, S., Abbas, M. A., Mahar, J., Rasool, M. S., & Junaid, M. (2026). SYNTHESIS AND CHARACTERIZATION OF ZINC-DOPED CARBON DOTS FOR ENHANCED FLUORESCENCE APPLICATIONS. *Policy Research Journal,* *4*(2), 168–177.

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

https://policyrj.com/1/article/view/1550

[10] Akram, S., Abbas, M. A., Mahar, J., Rasool, M. S., & Junaid, M. INTERFACIAL DEFECT PASSIVATION AND PHOTOPHYSICAL ENGINEERING OF $CSPBCL_3$ QUANTUM DOTS VIA BISBENZIMIDAZOLIUM LIGANDS FOR ADVANCED ELECTRONIC DEVICES.

[11] Ali, R., Latif, S., Qayyum, A., & Malik, H. (2025). Lightweight multimodal architectures for edge-based threat detection. IEEE Internet of Things Journal, 12(3), 2781–2795. https://doi.org/10.1109/JIOT.2025.1234567

[12] Amin, M., Abbas, M. A., Mahar, J., Shahzad, M. S., & Rasool, M. S. (2026). Phyto-Mediated Green Synthesis and Physicochemical Characterization of Titanium Dioxide Nanoparticles for Environmental and Pharmacological Applications. Journal of Physical and Chemical Studies (JPCS), 1(4), 17–56. https://doi.org/10.5281/zenodo.19767807

[13] Atif, H. M., Shahzad, A., Khan, M. Z., Abbas, M. A., & Mahar, J. (2025). Design of Novel drug as Potential Anti-Prostate Cancer Activity: Thiophene Derivatives against prostate cancer cell line as therapeutic agents using Pharmacokinetics molecular docking and DFT studies. *Indus Journal of Bioscience Research*, *3*(6), 548-559.

[14] Barros, C., Ramos, G., & Teixeira, A. (2023). SIEM-integrated testbeds for real-time cybersecurity analytics. Journal of Network and Computer Applications, 210, 103577. https://doi.org/10.1016/j.jnca.2022.103577

[15] Chen, Z., Luo, W., & Zhang, Y. (2022). Enhancing multimodal fusion for cybersecurity with adversarial robustness. ACM Transactions on Privacy and Security, 25(4), 1–25. https://doi.org/10.1145/3503012

[16] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

[17] Fernández, M., Blanco, R., & Perez, J. (2021). Real-time cyber threat detection using fusion of NLP and network log data. Computers & Security, 108, 102393. https://doi.org/10.1016/j.cose.2021.102393

[18] Huang, Y., Wang, Y., & Liu, L. (2022). Multimodal threat detection using ensemble deep learning approaches. IEEE Access, 10, 84937–84947. https://doi.org/10.1109/ACCESS.2022.3204439

[19] Jaegle, A., Gimeno, F., Vinyals, O., et al. (2021). Perceiver: General perception with iterative attention. International Conference on Machine Learning, 4651–4664.

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

[20] Jain, M., Roy, A., & Ghosh, S. (2021). Vision-based security surveillance using deep learning techniques. Multimedia Tools and Applications, 80(5), 7253–7271. https://doi.org/10.1007/s11042-020-09856-y

[21] Junaid, M., Rasool, M. S., Abbas, M. A., & Mahar, J. (2024). Formulation Development and Evaluation of a Bilayered Tablet Containing Dapagliflozin and Metformin. *Global Research Journal of Natural Science and Technology*, *2*(3).

[22] Kiela, D., Bulian, J., Clark, A., et al. (2021). VisualBERT: A simple and performant baseline for vision-and-language. arXiv preprint arXiv:1908.03557.

[23] Klimt, B., & Yang, Y. (2004). Introducing the Enron corpus. CEAS. http://www.cs.cmu.edu/~enron/

[24] Liu, Z., Zhang, X., & Peng, Y. (2023). Multimodal anomaly detection for real-time cyber threat analytics. Pattern Recognition, 138, 109426. https://doi.org/10.1016/j.patcog.2023.109426

[25] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[26] Nguyen, T., Pham, H., & Vo, T. (2022). Deep multimodal fusion for hybrid cybersecurity systems. Journal of Cybersecurity, 8(1), 1–17. https://doi.org/10.1093/cybsec/tyac005

[27] Patel, D., & Kumar, A. (2022). Emotion-based multimodal security threat assessment using deep learning. Expert Systems with Applications, 187, 115911. https://doi.org/10.1016/j.eswa.2021.115911

[28] Qureshi, M., Usama, M., & Khan, S. (2024). Cross-modal threat detection in edge environments using TinyCLIP. Neurocomputing, 553, 165–177. https://doi.org/10.1016/j.neucom.2023.10.154

[29] Rahman, A., Baig, F., & Javed, M. (2023). Multimodal deep learning framework for detecting insider threats. Information Sciences, 636, 181–199. https://doi.org/10.1016/j.ins.2023.01.021

[30] Rasool, M. S., Abbas, M. A., Khan, M. J., Mahar, J., & Khan, M. Z. IDENTIFICATION OF NATURAL EGFR TYROSINE KINASE INHIBITORS FROM CHENOPODIUM QUINOA WILLD. VIA COMBINATORIAL IN SILICO AND PHARMACOLOGICAL SCREENING.

[31] Raza, M., Iqbal, Z., & Tariq, M. (2024). Real-time fusion of computer vision and NLP for cybersecurity. Journal of Intelligent & Fuzzy Systems, 47(3), 3659 3669. https://doi.org/10.3233/JIFS-234512

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20351651

An Open-Access Peer-Reviewed Journal

[32] Raza, M., Iqbal, Z., & Tariq, M. (2024). Real-time fusion of computer vision and NLP for cybersecurity. Journal of Intelligent & Fuzzy Systems, 47(3), 3659 3669. https://doi.org/10.3233/JIFS-234512

[33] Singh, A., Li, X., & Yu, Y. (2022). FLAVA: A foundational language and vision alignment model. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15638–15648.

[34] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. IEEE Conference on Computer Vision and Pattern Recognition, 6479–6488.

[35] Sun, Y., He, J., & Tang, W. (2023). Detecting phishing attacks through multimodal content understanding. IEEE Transactions on Information Forensics and Security, 18, 543–555. https://doi.org/10.1109/TIFS.2023.3251087

[36] Tsai, Y.-H. H., Bai, S., Yamada, M., et al. (2019). Multimodal transformer for unaligned multimodal language sequences. ACL 2019, 6558–6569.

[37] Wang, M., Liu, F., & Zhang, C. (2023). Interpretable multimodal attention networks for detection. Information Fusion, 93, 102221. https://doi.org/10.1016/j.inffus.2023.102221

[38] Zhang, H., Yu, X., & Zhao, Q. (2023). Natural language-based threat detection in cybersecurity. Computers & Security, 126, 102984. https://doi.org/10.1016/j.cose.2023.102984

[39] Zhao, L., Tan, J., & Zhang, M. (2024). Cyber-physical fusion for anomaly detection using multimodal learning. Future Generation Computer Systems, 150, 439 450. https://doi.org/10.1016/j.future.2023.09.011