Rohan Singh, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Review of Hybrid Deep Learning Architectures for Real-Time Network Traffic Anomaly Detection: From Signature-Based Systems to Cross-Stream Attention CNN-BiLSTM Frameworks

Rohan Singh, Assistant Professor Jitender Kumar
Department of Computer Science and Engineering, GITAM, Kablana, Jhajjar

Abstract- The proliferation of encrypted network traffic and the accelerating sophistication of cyber-attack methodologies have rendered signature-based intrusion detection fundamentally inadequate for contemporary Security Operations Centre (SOC) and Network Operations Centre (NOC) environments, a problem underscored by a global cyberattack cost exceeding eight trillion United States dollars in 2023. This paper reviews the methodological evolution of network traffic anomaly detection, tracing the progression from signature-matching platforms through statistical and behavioural baselining to deep learning architectures spanning convolutional neural networks, recurrent and attention-augmented sequence models, graph neural networks, and Transformer-based classifiers. Particular attention is given to the limitations of single-modality architectures, which capture either the spatial packet-level structure or the temporal session-level dynamics of network traffic but not both simultaneously, and to the comparative strengths and weaknesses of early, late, and attention-based multi-stream fusion paradigms. The review further examines loss-level strategies for the severe class imbalance characteristic of intrusion-detection benchmarks, and surveys the explainability mechanisms required for analyst trust and MITRE ATT&CK alignment in production SOC deployment. Based on the gaps identified, the paper discusses an emerging direction—lightweight cross-stream attention CNN-BiLSTM frameworks exemplified by TrafficGuardNet—and outlines future research priorities, including model compression for ultra-high-throughput deployment, continual learning under concept drift, and feature-engineering adaptations for TLS 1.3-encrypted environments.

Keywords— Network intrusion detection, network anomaly detection, deep learning, convolutional neural network, bidirectional LSTM, cross-stream attention, class imbalance, focal loss, explainable AI, SHAP, Security Operations Centre.

I. INTRODUCTION

Enterprise and critical-infrastructure networks now generate traffic volumes and threat diversity that have decisively outpaced the operational capabilities of first-generation, signature-based intrusion detection systems (IDS). The global economic cost of cyberattacks surpassed eight trillion United States dollars in 2023, and this figure continues to climb as adversaries adopt increasingly evasive, multi-stage attack campaigns [1]. Compounding this challenge, the widespread adoption of Transport Layer Security across enterprise application traffic has eliminated payload visibility for the large majority of network communications, closing off the primary information source on which classical deep packet inspection depended.

Security Operations Centre and Network Operations Centre teams consequently require detection frameworks that infer malicious behaviour from flow-level and behavioural metadata alone, that

operate at sub-second latency on commodity server hardware without dedicated GPU or FPGA acceleration, and that produce analyst-interpretable explanations rather than opaque categorical labels. Meeting these requirements simultaneously has proved difficult: architectures that achieve the highest published detection accuracy typically rely on computationally intensive graph construction, quadratic-complexity self-attention, or large-scale transformer pretraining, while architectures engineered for lightweight commodity deployment have historically sacrificed either spatial or temporal representational capacity.

Automated network anomaly detection has progressed through three broadly identifiable generational paradigms, summarised in Table I. The first generation relied on manually curated attack signatures, exemplified by platforms such as Snort and Suricata, and offered predictable, low-false-positive detection of previously catalogued threats but was structurally incapable of identifying zero-day exploits. The second generation introduced statistical and behavioural baselining—Shannon entropy analysis of header-field distributions, CUSUM change detection, autoregressive traffic modelling, and Principal Component Analysis—which generalised to previously unseen volumetric and scanning attacks but remained vulnerable to low-rate, slow-scan, and advanced persistent threat campaigns deliberately engineered to remain statistically indistinguishable from benign traffic at the flow-aggregate level. The third and current generation applies deep learning architectures capable of automatic, end-to-end feature discovery and multi-class attack categorisation across convolutional, recurrent, graph-based, and Transformer-based model families [1], [2].

This paper reviews the literature underlying this trajectory, with particular emphasis on the architectural and fusion-level challenges that remain unresolved—most notably the difficulty of jointly modelling spatial packet-level and temporal session-level traffic characteristics within a single lightweight, deployable architecture, and the persistent degradation of minority-class detection performance under severe class imbalance. Section II reviews the literature across seven thematic areas. Section III presents a comparative synthesis of the reviewed architectures. Section IV consolidates the research gaps identified across this literature. Section V discusses an emerging direction—lightweight cross-stream attention CNN-BiLSTM frameworks—as a candidate response to these gaps, Section VI outlines future research priorities, and Section VII concludes the paper.

II. LITERATURE REVIEW

1. Signature-Based and Statistical Anomaly Detection: Foundations and Limits

The earliest operational intrusion detection platforms, typified by Snort and later Suricata, matched packet payloads and header fields against manually maintained libraries of textual and binary attack signatures. These systems remained computationally inexpensive, highly predictable, and reliable for known threats, properties that continue to justify their inclusion as one layer of a defence-in-depth architecture. Their central weakness is structural rather than implementational: by definition, an attack that has not previously been catalogued cannot be matched against any signature library, and the encryption of the large majority of enterprise traffic through TLS has further eroded the payload visibility on which these systems depend [2].

Statistical and behavioural anomaly detection emerged as a response, characterising the distribution of normal network behaviour and flagging departures exceeding a threshold without requiring prior knowledge of a specific attack. Khan and Mahmood, reviewing the transition from signature-based to statistical and machine-learning-based approaches in Software-Defined Networking intrusion detection, identified this shift as a necessary response to the inability of signature libraries to keep pace with attacker innovation [2]. Statistical methods generalised well to volumetric and scanning attacks that introduce clear distributional shifts, but their fundamental limitation is that adversaries aware of a deployed statistical baseline can craft attack traffic that is aggregate-statistically indistinguishable from benign traffic,

revealing its malicious character only through joint analysis across multiple temporal scales—precisely the capability that motivated the subsequent shift toward deep learning.

2. CNN-Based Spatial Feature Learning for Traffic Analysis

Convolutional architectures were among the first deep learning approaches applied to network traffic classification, motivated by the observation that traffic contains locally evident structural patterns—co-occurring protocol flags, characteristic header-field combinations, and byte-level sub-sequences—that are difficult to specify manually but readily discoverable through convolutional filters. One-dimensional CNNs treat a flow's feature vector or packet-byte sequence as a one-dimensional signal, while two-dimensional variants treat flow feature matrices as image-like spatial inputs. Doriguzzi-Corin et al. demonstrated with the LUCID architecture that a well-designed lightweight 1D-CNN could deliver strong DDoS detection performance on the CICIDS-2018 benchmark while meeting the throughput requirements of 100 Gbps network environments without dedicated hardware acceleration [3].

The persistent limitation of CNN-only architectures is the absence of temporal modelling capacity: the receptive field of a convolutional stack is bounded by filter size and network depth, and cannot directly span the long-range dependencies between flows that characterise attack campaigns unfolding gradually across an extended analysis window. Attacks whose maliciousness accumulates progressively over multiple flows or sessions therefore remain largely undetectable by CNN-only architectures regardless of network depth or filter count.

3. Recurrent and Attention-Augmented Temporal Architectures

Long Short-Term Memory networks and their bidirectional variants directly address the sequential-dependency limitation of CNN architectures by treating traffic flows as ordered sequences and learning, through gated memory, which information to retain across an analysis window. Bidirectional processing, in which forward and backward hidden states are concatenated, allows a single model to capture both the initiation dynamics of an attack campaign and its persistence dynamics. Zhou, Tang, and Wang showed that a Bi-LSTM intrusion detection model augmented with an attention mechanism could assign differential weight to the most anomalous temporal segments of an analysis window, substantially mitigating the shadowing effect of coexisting benign traffic and improving detection precision relative to LSTM-only baselines [4]. Nguyen and Tran extended the same attention-augmented Bi-LSTM paradigm to IoT network intrusion detection, confirming its benefit under the irregular communication patterns and constrained device capabilities typical of that domain [5].

Attention-augmented recurrent architectures were first validated in adjacent security tasks before their application to network intrusion detection. Liu, Zhang, and Zhao demonstrated that a bidirectional LSTM with attention improved detection of obfuscated malicious URLs by learning to focus on the most suspicious character subsequences within otherwise benign-looking strings [6], while Zhang and Wang applied an analogous attention-augmented Bi-LSTM to HTTP traffic anomaly detection, showing that the model learned to assign high attention weight to request fields most relevant to injection and path-traversal attempts, yielding a degree of interpretability alongside its accuracy gains [7].

The common limitation shared by pure LSTM and Bi-LSTM architectures, even with intra-stream attention, is that their input representation encodes only flow-aggregate statistics. Two traffic sessions—one malicious, one a benign low-volume application polling session—can present nearly identical flow-aggregate statistics while differing substantially in packet-level structure: flag distributions, inter-packet timing, and header-field value distributions. Because this distinguishing information resides in the spatial rather than the temporal dimension, it cannot be recovered by a temporal-only architecture, motivating the dual-stream designs discussed in Section II-E.

4. Graph Neural Network and Transformer-Based Architectures:

Capability versus Computational Cost Graph neural network approaches explicitly represent network endpoints as graph nodes and communication sessions as directed edges, propagating information through the graph structure to detect anomalous relational patterns. Mani, Rao, and Vijay combined Graph Convolutional Networks with Bi-LSTM processing of temporal behavioural sequences to detect insider threats, capturing both relational access-pattern anomalies and their temporal evolution [11]. Kim, Park, and Lee reported 96.1% accuracy for a graph-neural-network-based detector evaluated on the CAIDA backbone traffic dataset [8], while related inductive graph representation-learning approaches such as E-GraphSAGE, evaluated by Lanvin, Garcia, and Papadopoulos on CICIDS-2017, achieved an F1-score of 0.917 [10]. The principal drawback of graph-based methods is the computational cost of constructing and processing a communication graph in real time: at throughput levels exceeding 10 Gbps, graph construction, node-feature computation, and message-passing rounds can collectively exceed the inference-latency budget available per analysis window on commodity server hardware.

Transformer-based architectures replace convolutional inductive bias with multi-head self-attention over traffic-derived token sequences, enabling direct modelling of relationships between distant positions. Zhao and Patel applied a BERT-based Transformer to encrypted TLS metadata classification, reporting an F1-score of 0.914, though restricted to binary rather than multi-class classification [9]. The limitation shared by both graph-based and Transformer-based architectures is computational: quadratic self-attention complexity and graph-construction overhead are difficult to reconcile with the sub-second, GPU-free inference budgets required for Tier-1 SOC deployment on commodity hardware.

5. Multi-Stream Feature Fusion Paradigms: Early, Late, and Attention-Based Fusion

Combining complementary feature streams within a single detection architecture has consistently improved multi-class attack detection over single-stream alternatives, and the literature identifies three distinct fusion paradigms. Early fusion concatenates raw features from multiple sources into a single high-dimensional input vector prior to model processing; while implementationally simple, it forces categorical protocol fields, byte-range values, and floating-point statistical measures into a shared representational space without any architectural mechanism to reconcile their differing statistical distributions, and cannot resolve the temporal misalignment between flows of very different duration represented at the same fixed dimensionality.

Late ensemble fusion instead trains each feature stream independently and combines output probability distributions through weighted averaging, majority voting, or a trained stacking classifier. This preserves robustness to individual stream failure, a valuable property in SOC environments with fluctuating sensor availability, but cannot detect compound attack signatures in which neither modality alone reveals malicious intent but their co-occurrence does—precisely the signature structure exhibited by advanced persistent threat lateral movement and low-rate DDoS campaigns.

Attention-based fusion allows each stream to dynamically query the other at each processing step, learning through gradient descent which cross-stream associations are most diagnostically relevant without manual specification. Krishnan and Naik showed that a CNN combined with Bi-LSTM through attention-based fusion outperformed both single-stream baselines, particularly for low-intensity attacks anomalous at neither the spatial nor the temporal dimension in isolation [12]. Rahman et al. demonstrated the generality of this paradigm outside network security, in surveillance-derived conflict monitoring, where attention learned that specific spatial features in individual frames were diagnostically significant only in the context of specific temporal patterns across the frame sequence—a structural parallel to cross-stream traffic fusion in which spatial packet patterns are informative only in specific temporal session contexts [13]. A further design choice concerns

fusion granularity: attention applied at the intermediate representation level, allowing fine-grained querying of specific spatial embeddings by specific temporal states, captures compound signatures that coarser output-level fusion cannot.

6. Loss-Level Strategies for Class Imbalance in Intrusion Detection

Independent of architectural choice, the severe class imbalance characteristic of intrusion-detection benchmarks represents a persistent barrier to reliable minority-class detection. In datasets such as CICIDS-2018, benign and high-volume attack traffic (e.g., DDoS) constitute the overwhelming majority of records, while high-severity categories such as Infiltration and Remote-to-Local exploits are represented by a small fraction of samples. Under standard categorical cross-entropy training, the gradient contribution of abundant, easily classified examples dominates parameter updates, causing the model to optimise overall accuracy primarily by performing well on majority classes at the expense of the rare categories most consequential to detect.

Focal Loss, introduced by Lin et al. for dense object detection, addresses this asymmetry by applying a modulating factor that down-weights the gradient contribution of well-classified easy examples while preserving the contribution of hard, frequently misclassified minority-class examples [14]. Carvalho and Proenca, reviewing ensemble-based approaches to mitigating data imbalance in network intrusion detection, similarly emphasised that static class-weighting schemes provide only a first-order correction and that dynamically modulated loss functions are better positioned to sustain minority-class recall throughout training [15]. This comparison suggests that architectures relying solely on static inverse-frequency weighting may face an inherent ceiling on minority-class performance that focal-loss-style dynamic modulation is positioned to overcome.

7. Explainable AI and Cross-Domain Generalisation of Spatial-Temporal Fusion Architectures

Beyond raw detection performance, the operational adoption of deep-learning-based IDS in production SOC environments depends on analyst-interpretable explanations. Opaque detections impose two related costs: analysts cannot verify a detection without evidence of the reasoning behind it, and unexplained alerts cannot be readily mapped onto established frameworks such as MITRE ATT&CK, which analysts rely on to contextualise, prioritise, and escalate security events. SHAP-based (SHapley Additive exPlanations) feature-attribution analysis and attention-weight visualisation have emerged as complementary mechanisms for addressing this gap, providing both global feature-importance rankings and per-alert, analyst-facing explanations grounded directly in model internals rather than surrogate approximations.

The architectural principle of jointly modelling spatial and temporal structure through cross-stream fusion generalises beyond enterprise network traffic to a range of adjacent security domains. Nankya et al., reviewing AI and machine-learning techniques for security and privacy in e-health systems, identified LSTM- and CNN-based anomaly detection as central to identifying illegitimate access patterns and data-exfiltration attempts in healthcare networks [17]. Islam and Ahmed showed that bidirectional recurrent anomaly detection improved intrusion detection in Internet-of-Drones communication architectures characterised by irregular, delay-sensitive traffic [18], while Bui and Yang demonstrated that the same spatial-temporal fusion principle could detect adversarial interference in UAV swarm coordination protocols, where individual command messages appear syntactically normal but the temporal pattern of the command sequence exposes tampering [19]. Abbas and Farooq identified an analogous twofold representational requirement—local structural analysis paired with global temporal dispersion analysis—in secure steganography systems [20], and Yang and Nguyen reported comparable spatial-temporal demands in detecting adversarial power-injection attacks in wireless-powered communication networks [21]. Across these domains, architectures explicitly modelling both local structural characteristics and global temporal dynamics consistently outperform architectures modelling only one, corroborating the generalisability of cross-stream attention fusion as a

representational paradigm beyond enterprise network intrusion detection.

III. COMPARATIVE SYNTHESIS OF REVIEWED ARCHITECTURES

The architectures reviewed in Section II span a wide range of reported accuracy, computational complexity, and hardware requirements, summarised comparatively in Table II. A broadly observable, though not strict, association emerges between reported detection performance and architectural complexity: the highest-F1 entries—graph-neural-network and BERT-Transformer approaches—generally require GPU acceleration and, in the case of graph construction, inference latencies exceeding one second per analysis window at realistic throughput, while lightweight single-modality CNN and attention-LSTM architectures achieve CPU-deployable latency at a measurable cost in detection performance.

Second, hardware requirement and dataset generalisability are reported far less consistently than raw accuracy or F1-score, and several high-performing entries in Table II provide no CPU-latency figure at all (denoted N/A). This asymmetry constitutes a methodological gap in its own right: sub-second CPU deployability, despite being repeatedly identified as an operational precondition for SOC adoption, is not yet treated as a primary evaluation criterion alongside accuracy.

Third, among the architectures reviewed, no single published system combines CNN spatial modelling, Bi-LSTM temporal modelling, and cross-stream attention fusion while remaining deployable on commodity CPU hardware at sub-second latency—a combination discussed further in Section IV and addressed directly by the framework reviewed in Section V.

Table 2: Comparative Summary of Reviewed Network Intrusion Detection Architectures

System	Architecture	Dataset	F1-Score	Latency	Hardware
Kim et al. [8]	Graph Neural Network	CAIDA	0.941	N/A	GPU
Zhao & Patel [9]	BERT Transformer	TLS metadata	0.914	High	GPU
LUCID [3]	1D-CNN	CICIDS-2018	0.832	0.21 s	CPU
E-GraphSAGE [10]	Inductive Graph GNN	CICIDS-2017	0.917	>1 s	GPU
Ahmed et al. [22]	Federated CNN	Multi-NOC	~0.870	N/A	GPU
Sundaram & Lee [23]	Attention LSTM	NSL-KDD	0.889	N/A	GPU
Zhou et al. [4]	Bi-LSTM + Attention	Multi-dataset	0.901	N/A	GPU
Krishnan & Naik [12]	CNN + Bi-LSTM + Attention	DDoS dataset	0.911	N/A	GPU
TrafficGuardNet [25]	CNN + Bi-LSTM + Cross-Stream Attention	CICIDS-2018 / NSL-KDD	0.92	0.41 s	CPU

IV. RESEARCH GAPS

Synthesising the literature reviewed in Section II, four interrelated gaps can be identified that collectively constrain the operational readiness of current deep-learning-based network intrusion detection for SOC and NOC deployment.

First, no published architecture combines spatial packet-level and temporal session-level modelling through cross-stream attention fusion while remaining deployable at sub-second inference latency on commodity CPU hardware without GPU or FPGA acceleration. The systems reviewed fall into two categories: those that perform well through

computationally intensive multi-modal or graph-based designs difficult to accelerate in mid-tier SOCs, and those that are easy to deploy on commodity hardware precisely because their single-modality design sacrifices either spatial or temporal representational capacity. No reviewed system occupies the intersection of both requirements.

Second, published architectures achieving strong benchmark performance frequently rely on preprocessing pipelines requiring deep packet inspection hardware or proprietary network-probe appliances, which are costly and operationally complex to deploy consistently across the heterogeneous network segments typical of production SOC environments. A hardware-independent pipeline capable of computing diagnostic behavioural descriptors directly from standard NetFlow v9 or PCAP inputs, without payload inspection, remains comparatively underexplored.

Third, the operational interpretability of deep-learning-based IDS remains limited. Few reviewed architectures provide feature- or pattern-level insight into individual classification decisions, which both reduces analyst confidence in automated detections and prevents efficient mapping of detections onto frameworks such as MITRE ATT&CK that analysts use for triage and escalation.

Fourth, minority-class detection under severe class imbalance remains inconsistently addressed. High-severity, low-frequency attack categories such as Infiltration and Remote-to-Local exploits are repeatedly reported with markedly lower recall than high-volume categories such as DDoS and port scanning, a consequence of standard cross-entropy training allowing abundant majority-class gradients to dominate parameter updates.

V. TOWARD LIGHTWEIGHT CROSS-STREAM ATTENTION CNN-BiLSTM FRAMEWORKS

The gaps identified above point toward hybrid architectures that couple CNN spatial feature extraction and Bi-LSTM temporal modelling through learned cross-stream attention, while remaining computationally light enough for commodity CPU deployment. TrafficGuardNet, a recently reported framework of this kind, illustrates how these requirements can be reconciled within a single end-to-end trainable model [25].

The architecture processes traffic through a four-stage pipeline: traffic ingestion and window segmentation, which splits incoming PCAP or NetFlow v9 traffic into thirty-second analysis windows advanced every ten seconds; parallel dual-stream feature extraction, in which a CNN spatial stream and a Bi-LSTM temporal stream operate independently on the same traffic window; cross-stream attention fusion, which dynamically combines both representations based on a learned attention mechanism; and multi-class classification, in which the fused representation is passed to a fully connected head producing both a categorical attack label and a continuous severity score.

The CNN spatial stream ingests a 256×32 flow feature matrix—256 parallel flows observed across 32 time-slices, each described by 32 engineered features drawn from the CICIDS-2018 feature set (flag counts, byte and packet statistics, inter-arrival timing, and header-length measures) selected via mutual-information pre-screening from the full 86-feature set. Four sequential convolutional blocks (64, 128, 256, and 512 filters) extract progressively higher-level spatial structure, followed by an intra-stream self-attention layer and global average pooling to yield a 512-dimensional spatial context vector.

The Bi-LSTM temporal stream is paired with a Real-Time Flow Statistics Pipeline (RTFP), a stateless, per-window preprocessing module that computes seven behavioural descriptors—inter-arrival time variance, packet length standard deviation, flow duration, SYN/ACK ratio, flag entropy, forward/backward byte ratio, and connection frequency—directly from standard NetFlow v9 or PCAP inputs without payload inspection, guaranteeing operation under TLS-encrypted traffic. A 256-unit-per-direction Bi-LSTM processes these descriptors alongside a CNN pre-encoder to capture session-level temporal dynamics.

The cross-stream attention fusion module allows each stream to query the other at the intermediate representation level, learning which cross-stream associations are diagnostically relevant for the observed traffic context, addressing directly the compound-signature limitation of late ensemble fusion discussed in Section II-E. Classification is trained with focal loss rather than standard cross-entropy:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} \log(p_t)$$

where pt is the predicted probability for the true class, γ = 2.0 is the focusing parameter, and αt is a class-specific weight (0.25 for the Normal class and 0.75 for all attack classes), directly implementing the dynamically modulated imbalance correction identified as a research priority in Section II-F. The complete model comprises approximately 3.77 million trainable parameters—an order of magnitude fewer than typical transformer-based alternatives—distributed across the CNN spatial stream (34.7%), the temporal-stream CNN pre-encoder (23.6%), the Bi-LSTM (27.8%), and the cross-stream attention and classification head (13.9%).

An ablation study evaluated on the CICIDS-2018 held-out test partition isolates the contribution of each architectural component, summarised in Table III. Removing the Bi-LSTM temporal stream entirely reduces accuracy by 11.2 percentage points, confirming that temporal session-level modelling is essential for attacks whose maliciousness accumulates progressively, such as slow scans and low-rate DDoS. Replacing cross-stream attention with late ensemble fusion reduces accuracy by 5.6 percentage points despite using comparable computation, directly validating attention-based fusion over late ensemble alternatives for this task. Replacing focal loss with standard cross-entropy reduces accuracy by 3.5 percentage points, confirming the value of dynamic loss-level imbalance correction for minority attack categories [25].

Table 3: Ablation Study Results on the Cicids-2018 Held-Out Test Partition [25]

Configuration	Accuracy (%)	F1-Score	AUC-ROC (%)	Latency (s/win)
Full TrafficGuardNet	94.3	0.92	97.1	0.41
w/o cross-stream attention (late ensemble)	88.7	0.85	93.4	0.38
w/o Bi-LSTM stream (CNN only)	83.1	0.79	89.2	0.27
w/o focal loss (standard cross-entropy)	90.8	0.87	94.6	0.41

Evaluated against six published systems representing the principal architectural lines reviewed in Section II—CNN-only (LUCID), attention-augmented LSTM, graph neural network, BERT Transformer, federated CNN, and Bi-LSTM attention—TrafficGuardNet reported a detection accuracy of 94.3%, precision of 0.93, recall of 0.91, F1-score of 0.92, AUC-ROC of 97.1%, false positive rate of 2.8%, and inference latency of 0.41 seconds per thirty-second window on Intel Xeon E5-2690 CPU hardware, an improvement of 11.2 percentage points over CNN-only baselines and approximately 8 percentage points over published state-of-the-art systems including LUCID and E-GraphSAGE [25]. On cross-dataset evaluation using the NSL-KDD KDDTest+ partition without dataset-specific fine-tuning, the framework achieved a macro-averaged F1-score of 0.871, within the targeted 5% cross-dataset generalisation margin.

DeepSHAP-based interpretability analysis confirmed that the dominant globally important features—flag entropy, connection frequency, inter-arrival time variance, and forward/backward byte ratio—correspond directly to the behavioural signatures of known attack categories documented in the MITRE ATT&CK framework, and that cross-stream attention weights redistribute coherently between the spatial and temporal streams depending on attack type

(mean attention 0.71 toward temporal features for Port Scan windows versus 0.68 toward spatial features for DDoS windows), directly addressing the interpretability gap identified in Section IV [25]. Positioned within the lightweight, edge-deployable design space discussed in Section II-B through II-D, this class of cross-stream attention CNN-BiLSTM framework represents a candidate resolution to the architectural, preprocessing, interpretability, and imbalance gaps identified in Section IV, though its accuracy remains below the highest figures reported for GPU-dependent graph-neural-network and Transformer pipelines in Table II.

Future Research Directions

Building on the gaps and the emerging direction discussed above, several concrete research priorities follow directly from the limitations identified in this review.

Model compression through pruning, INT8 quantisation, and knowledge distillation represents the most immediately actionable engineering priority, targeted at reducing parameter count and computational complexity by 70–80% to extend sub-second inference deployability to ultra-high-throughput (100 Gbps) backbone environments, where current 256-flow input capacity is insufficient for the concurrent flow volume observed.

Continual learning under concept drift constitutes a second priority. Architectures trained once offline on static historical traffic will degrade as production traffic distributions diverge from the training distribution. Sliding-window continual learning methods such as Elastic Weight Consolidation and Experience Replay offer a route to incremental adaptation without catastrophic forgetting of previously learned detection capability, avoiding the coverage gap introduced by periodic full retraining.

Third, the accelerating adoption of TLS 1.3 encryption, which hides application-protocol identifiers through encrypted header extensions and eliminates forward-secret decryption for forensic analysis, motivates redesigning behavioural feature-engineering pipelines around timing and volume attributes—inter-arrival time variance, byte-count ratios, flow duration, and connection frequency—that remain observable on fully encrypted connections, alongside systematic assessment of the resulting performance degradation.

Fourth, multi-dataset joint training across CICIDS-2018, NSL-KDD, and UNSW-NB15 offers a path toward narrowing the cross-dataset generalisation gap currently addressed only through conservative zero-imputation feature matching, and toward validating detection performance against the more heterogeneous, less stereotyped attack traffic characteristic of live enterprise networks rather than laboratory-emulated benchmarks.

Fifth, calibrated uncertainty estimation—through Monte Carlo dropout inference or post-hoc softmax calibration—would extend current binary severity scoring to provide analysts with per-class confidence intervals, distinguishing a genuinely ambiguous multi-class detection from a confidently resolved one.

Finally, the cross-domain evidence reviewed in Section II-G—spanning vehicular CAN bus intrusion detection, IoT device monitoring, industrial control systems, and building automation anomaly detection—suggests that systematic evaluation of the cross-stream spatial-temporal fusion template across Operational Technology and Cyber-Physical System environments, where dermatologist-equivalent scarcity of security analysts and hardware constraints are especially acute, represents a high-impact extension of this research beyond the enterprise IT network environment reviewed here [16], [24].

VI. CONCLUSION

This review has traced the methodological evolution of network traffic anomaly detection from signature-based and statistical baselining through deep convolutional, recurrent, graph-based, and Transformer architectures, to lightweight cross-stream attention CNN-BiLSTM frameworks. Across this trajectory, a persistent tension recurs between detection performance and operational deployability: the architectures reporting the strongest benchmark scores—graph neural

networks, Transformer models, and large-scale ensembles—are frequently the least suited to the resource-constrained, latency-sensitive, interpretability-critical environment of a production SOC or NOC, while architectures engineered for hardware efficiency have historically sacrificed either spatial or temporal representational capacity. The literature reviewed here suggests that this trade-off is not intrinsic to deep learning applied to network traffic analysis but is a consequence of single-modality architectural choices. Lightweight cross-stream attention CNN-BiLSTM frameworks, exemplified by TrafficGuardNet, represent a promising direction for reconciling spatial and temporal representational completeness with sub-second CPU-deployable inference and SHAP-grounded interpretability, though model compression for ultra-high-throughput deployment, continual learning under concept drift, TLS 1.3-adapted feature engineering, and calibrated multi-class uncertainty estimation remain open and operationally consequential research priorities.

Acknowledgment

The author gratefully acknowledges the guidance and supervision of Mr. Jitender Kumar, Assistant Professor, Department of Computer Science and Engineering, Ganga Institute of Technology and Management, Kablana, Jhajjar, throughout the preparation of the dissertation on which this review is based.

REFERENCES

1. F. M. Anis, M. AlAbdullatif, S. Aljbli, and M. Hammoudeh, "A Survey on the Applications of Deep Learning in Network Intrusion Detection Systems to Enhance Network Security," IEEE Access, vol. 13, pp. 1–22, Jan. 2025.
2. M. Khan and T. Mahmood, "Review of Deep Learning and Machine Learning in SDN-based NIDS," PMC, PMC9607035, Nov. 2022.
3. R. Doriguzzi-Corin, S. Millar, S. Scott-Hayward, J. Martínez-del-Rincón, and D. Siracusa, "LUCID: A Practical, Lightweight Deep Learning Solution for DDoS Attack Detection," IEEE Trans. Netw. Serv. Manag., vol. 17, no. 2, pp. 876–889, Jun. 2020.
4. H. Zhou, Y. Tang, and L. Wang, "Network Intrusion Detection Model Based on Bidirectional LSTM and Attention Mechanism," in Proc. IEEE Conf. Cyber Security, May 2025.
5. T. Nguyen and D. Tran, "Intrusion Detection in IoT Networks Using Bidirectional LSTM with Attention Mechanism," in Proc. IEEE Conf. IoT Systems, Sep. 2025.
6. X. Liu, J. Zhang, and Q. Zhao, "A Bi-Directional LSTM Model with Attention for Malicious URL Detection," in Proc. 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conf. (IAEAC), Dec. 2019.
7. J. Zhang and Y. Wang, "Attention-Based Bi-LSTM Model for Anomalous HTTP Traffic Detection," in Proc. 2019 15th Int. Conf. Service Systems and Service Management (ICSSSM), Jan. 2019.
8. S. Kim, D. Park, and J. Lee, "Graph Neural Network-Based Intrusion Detection for Backbone Network Traffic," in Proc. IEEE Conf. Network and Service Management, Nov. 2024.
9. Y. Zhao and R. Patel, "A BERT Transformer Approach for Encrypted TLS Metadata Classification," in Proc. IEEE Int. Conf. Communications, Jun. 2024.
10. P. Lanvin, S. Garcia, and E. Papadopoulos, "Inductive Graph Representation Learning for Network Intrusion Detection on CICIDS-2017," in Proc. IEEE Global Communications Conf., Dec. 2022.
11. K. Mani, B. Rao, and G. Vijay, "Insider Threat Detection Using GCN and Bi-LSTM with Explicit and Implicit Graph Representations," IEEE Journals & Magazine, vol. 12, pp. 3140–3152, Jun. 2024.
12. S. Krishnan and M. Naik S, "DDoS Attack Detection using CNN-BiLSTM with Attention Mechanism," Telematics and Informatics Reports, vol. 18, p. 100211, Jan. 2025.
13. A. Rahman, M. Ali, and S. Hasan, "Combining CNNs with BiLSTM and Attention Mechanism for Conflict Monitoring," arXiv preprint arXiv:2502.18555, Feb. 2025.
14. T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, Feb. 2020.

15. L. Carvalho and M. Proenca, "Mitigating Data Imbalance in Network Intrusion Detection Using Ensemble Models," MDPI Systems, vol. 12, no. 1, p. 2, Jan. 2024.
16. J. Abshari and S. Sridhar, "A Survey of AI-Based Anomaly Detection for Cyber-Attack Detection in ICS and CPS Environments," MDPI Sensors, vol. 15, no. 1, p. 20, Jan. 2026.
17. M. Nankya, A. Mugisa, Y. Usman, A. Upadhyay, and R. Chataut, "Security and Privacy in E-Health Systems: A Review of AI and Machine Learning Techniques," IEEE Access, vol. 12, pp. 148798–148815, Oct. 2024.
18. M. Islam and S. Ahmed, "AI-Enhanced Intrusion Detection Systems for IoD Architectures," MDPI Electronics, vol. 15, no. 6, p. 1204, Mar. 2026.
19. J. Bui and L. Yang, "Formation Control and Security in UAV Swarms," MDPI Drones, vol. 10, no. 4, p. 278, Apr. 2026.
20. H. Abbas and M. Farooq, "Requirements for Secure Steganography Systems," MDPI Mathematics, vol. 9, no. 21, p. 2829, Nov. 2021.
21. S. Yang and T. Nguyen, "Pairwise Cooperation in Wireless Powered Communication Networks," MDPI Electronics, vol. 15, no. 10, p. 1890, May 2026.
22. M. Ahmed, S. Rahman, and K. Osei, "Federated Convolutional Neural Networks for Privacy-Preserving Multi-NOC Intrusion Detection," in Proc. IEEE Conf. Distributed Computing Systems, Jul. 2025.
23. R. Sundaram and C. Lee, "Attention-Augmented LSTM for Network Intrusion Detection on NSL-KDD," in Proc. Int. Conf. Machine Learning Applications, Dec. 2023.
24. MarketsandMarkets, "Operational Technology (OT) Security Market — Global Forecast to 2031," MarketsandMarkets Research Report, Feb. 2026.
25. R. Singh, "Real-Time Network Traffic Anomaly Detection in Security and Network Operations Centers," M.Tech. dissertation, Dept. Comput. Sci. Eng., Ganga Inst. Technol. Manage., Kablana, Jhajjar, India, 2026.