

A REVIEW OF SPLIT LEARNING–BASED SECURE TRAFFIC CLASSIFICATION FOR PRIVACY-SENSITIVE NETWORKS

Annapurna Yadav¹, Mrs. Arifa Khan²

¹Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

²Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

Abstract - The rapid proliferation of encrypted communication and data-driven network services has significantly increased the demand for accurate and privacy-preserving traffic classification mechanisms. Traditional traffic analysis techniques, including deep packet inspection and centralized machine learning models, often require access to raw packet payloads, raising serious privacy, regulatory, and security concerns in privacy-sensitive environments such as healthcare, finance, and critical infrastructure networks. Recently, split learning has emerged as a promising distributed learning paradigm that partitions deep neural networks between clients and servers, enabling collaborative model training without direct sharing of raw data. This review systematically examines the state-of-the-art research on split learning–based secure traffic classification, focusing on architectural designs, privacy guarantees, communication overhead, and performance trade-offs. We analyze how split learning compares with related paradigms such as federated learning and differential privacy–based methods in the context of encrypted and large-scale network traffic. Furthermore, this review synthesizes existing datasets, evaluation metrics, and threat models adopted in prior studies, identifying limitations in benchmarking practices and security analysis. Key challenges, including gradient leakage, scalability constraints, and adversarial robustness, are critically discussed. Finally, potential research directions are outlined to guide future developments toward practical, secure, and high-performance deployment of split learning frameworks for traffic classification in privacy-sensitive networks.

Key Words: Split Learning, Secure Traffic Classification, Privacy-Preserving Machine Learning, Encrypted Network Traffic, Federated Learning, Network Security, Distributed Deep Learning

1. INTRODUCTION

The exponential growth of Internet-enabled services, cloud computing, IoT ecosystems, and mobile applications has intensified the need for efficient and intelligent network traffic classification. Traffic classification plays a fundamental role in network management, intrusion detection, quality-of-service (QoS) enforcement, and cybersecurity operations. However, the increasing deployment of encryption protocols and strict privacy regulations has significantly complicated traditional traffic analysis approaches. In this context, privacy-preserving

distributed learning paradigms—particularly split learning—have emerged as promising solutions. This review examines the evolution, challenges, and emerging role of split learning in secure traffic classification for privacy-sensitive networks.

1.1 Background and Problem Context

1.1.1 Evolution of Network Traffic Classification

Network traffic classification has evolved through several methodological paradigms. Early approaches relied on port-based identification, where traffic flows were categorized based on well-known transport layer port numbers. While computationally efficient, this technique became unreliable as applications increasingly adopted dynamic ports and port obfuscation mechanisms (Moore and Papagiannaki, 2005).

Subsequently, deep packet inspection (DPI) techniques were introduced, enabling payload-level inspection to achieve fine-grained classification. DPI significantly improved accuracy but required direct access to packet content, raising scalability and privacy concerns (Nguyen and Armitage, 2008).

With the proliferation of encrypted and high-volume traffic, statistical and machine learning-based methods gained prominence. These approaches leveraged flow-level features such as packet inter-arrival time, flow duration, and byte distribution patterns. More recently, deep learning architectures—including convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have demonstrated superior performance in encrypted traffic classification tasks by automatically extracting discriminative features (Lotfollahi et al., 2020).

1.1.2 Rise of Encrypted Traffic and Privacy Regulations

The widespread adoption of encryption protocols such as TLS has fundamentally altered the traffic classification landscape. Current reports indicate that a substantial majority of Internet traffic is encrypted, limiting visibility into packet payloads and rendering DPI ineffective (Anderson et al., 2017).

Simultaneously, regulatory frameworks such as the General Data Protection Regulation (GDPR) and sector-specific compliance mandates have imposed strict requirements on

how user data is processed and stored. These regulations emphasize data minimization, consent, and secure handling, thereby restricting centralized collection of raw network data for machine learning purposes (Voigt and Von dem Bussche, 2017).

Consequently, traffic classification systems must balance detection accuracy with privacy preservation, creating a complex design trade-off between visibility and compliance.

1.1.3 Limitations of Centralized ML-Based Traffic Analysis

Centralized machine learning frameworks require aggregation of raw traffic data at a central server for model training. Although effective in controlled environments, this approach introduces multiple limitations. First, transferring raw network data across administrative boundaries increases exposure risk and attack surfaces. Second, centralized storage creates a single point of failure vulnerable to breaches. Third, large-scale data transmission leads to significant communication overhead and latency (Kairouz et al., 2021).

Moreover, centralized deep learning models may inadvertently memorize sensitive traffic patterns, enabling model inversion or membership inference attacks. These vulnerabilities highlight the necessity for distributed privacy-aware learning architectures that minimize raw data exposure.

1.2 Motivation for Privacy-Preserving Traffic Classification

1.2.1 Privacy Risks in Raw Packet Inspection

Raw packet inspection inherently involves analyzing headers and payload contents, which may contain personally identifiable information (PII), authentication tokens, or proprietary communication details. Even flow-level metadata can reveal behavioral patterns when aggregated at scale. Studies have shown that traffic analysis techniques can infer user activities despite encryption, raising serious privacy implications (Shbair et al., 2016).

Furthermore, in domains such as healthcare, financial systems, and critical infrastructure, traffic data may correspond to highly sensitive operational information. Unauthorized exposure may result in regulatory penalties and reputational damage. Therefore, minimizing raw data visibility during model training has become a fundamental requirement.

1.2.2 Need for Distributed Collaborative Learning

Modern networks are inherently distributed, spanning edge devices, IoT sensors, enterprise gateways, and cloud infrastructures. In such environments, data is naturally partitioned across multiple nodes. Distributed collaborative

learning enables local model training without centralizing raw data, thereby reducing exposure risks.

Paradigms such as federated learning have demonstrated the feasibility of decentralized training by sharing model updates instead of data (McMahan et al., 2017). However, even gradient sharing may leak sensitive information under certain threat models. These limitations motivate exploration of alternative architectures that further restrict information exchange, leading to the adoption of split learning for privacy-sensitive traffic classification scenarios.

1.3 Emergence of Split Learning

1.3.1 Concept and Basic Architecture

Split learning is a distributed deep learning paradigm in which a neural network is partitioned between client and server segments. Clients perform forward propagation up to a predefined cut layer and transmit intermediate activations—not raw data—to a central server. The server completes forward and backward propagation and returns gradients to the client for local parameter updates (Vepakomma et al., 2018).

This architecture reduces direct data exposure while maintaining collaborative training efficiency. Because raw inputs remain at the client side, split learning is particularly suited for privacy-sensitive applications where data cannot be shared across organizational boundaries.

1.3.2 Distinction from Other Distributed Paradigms

Unlike federated learning, where complete local models are trained independently and only model parameters are shared, split learning divides a single model across entities. This structural difference reduces computational requirements at the client side and may offer improved privacy under specific threat assumptions.

Additionally, compared to secure multi-party computation or homomorphic encryption—which introduce significant computational overhead—split learning provides a practical trade-off between privacy and efficiency (Singh et al., 2019). Nevertheless, recent research has identified potential risks such as activation leakage and gradient reconstruction attacks, indicating that split learning is not inherently immune to inference threats.

1.4 Contributions of the Review

This review provides a comprehensive synthesis of research on split learning-based secure traffic classification in privacy-sensitive networks. First, it systematically examines the evolution of traffic classification methodologies and contextualizes the transition toward privacy-preserving learning frameworks. Second, it critically analyzes split learning architectures, their security properties, and comparative advantages over alternative distributed

paradigms. Third, it evaluates existing literature with respect to datasets, performance metrics, threat models, and deployment feasibility. Finally, it identifies research gaps and outlines future directions aimed at improving robustness, scalability, and regulatory compliance in real-world network environments.

2. FOUNDATIONS OF SECURE TRAFFIC CLASSIFICATION

Secure traffic classification integrates traditional network traffic analysis with privacy-preserving computational paradigms. As modern networks increasingly handle encrypted and sensitive communications, the foundations of secure classification must encompass methodological accuracy, security guarantees, and regulatory compliance. This section reviews core traffic classification techniques, examines privacy and security concerns, and outlines distributed learning paradigms that enable privacy-aware model training.

2.1 Network Traffic Classification Techniques

2.1.1 Port-Based Methods

Port-based classification represents the earliest technique for identifying application-layer protocols in IP networks. This approach maps traffic flows to predefined transport layer port numbers (e.g., HTTP on port 80, HTTPS on port 443). Its primary advantage lies in computational simplicity and low processing overhead. However, the technique suffers from severe limitations due to dynamic port allocation, port masquerading, and tunneling mechanisms used by modern applications (Moore and Papagiannaki, 2005).

Furthermore, with the increasing adoption of encrypted protocols and content delivery networks, port numbers no longer reliably correspond to specific services. Consequently, port-based methods have become largely obsolete in contemporary high-speed and encrypted environments.

2.1.2 Statistical and Flow-Based Methods

To address the limitations of port-based identification, statistical and flow-based classification techniques were introduced. These methods rely on traffic flow characteristics such as packet size distribution, flow duration, inter-arrival time, and byte counts. By extracting features from NetFlow or similar metadata records, classifiers can infer application types without accessing payload contents (Nguyen and Armitage, 2008).

Machine learning algorithms—including Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors—have been widely applied to such feature sets. Although these approaches improved classification accuracy in encrypted environments, their performance depends heavily on feature engineering and dataset

representativeness. Additionally, flow-level metadata can still reveal sensitive behavioral information when aggregated at scale.

2.1.3 Deep Learning Approaches

Recent advancements in deep learning have significantly enhanced encrypted traffic classification. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) enable automated feature extraction from raw packet sequences or flow matrices, reducing dependence on manual feature design. Architectures such as 1D-CNNs and Long Short-Term Memory (LSTM) networks have demonstrated superior performance in identifying encrypted applications and intrusion patterns (Lotfollahi et al., 2020).

Deep learning models offer scalability and adaptability to evolving traffic patterns. However, they typically require large volumes of labeled data and centralized training infrastructure, which introduces privacy and security concerns in sensitive deployment scenarios.

2.2 Privacy and Security Considerations

2.2.1 Data Exposure Risks

Even when traffic payloads are encrypted, metadata such as timing information, packet lengths, and flow statistics may enable inference of user behavior or application types. Traffic analysis attacks have shown that encrypted communications can still leak sensitive information through side-channel characteristics (Shbair et al., 2016).

Centralized storage of network traffic datasets further increases exposure risks by creating high-value targets for attackers. In addition, machine learning models themselves may leak information through model inversion or membership inference attacks, particularly when trained on sensitive datasets. Therefore, minimizing raw data sharing is a critical design objective in secure traffic classification frameworks.

2.2.2 Regulatory Implications

The regulatory landscape has intensified the importance of privacy-aware data processing. Frameworks such as the General Data Protection Regulation (GDPR) enforce strict requirements for lawful data collection, processing transparency, and data minimization (Voigt and Von dem Bussche, 2017).

For network operators and service providers, centralized traffic monitoring systems may conflict with compliance requirements if personal or behavioral data is stored without explicit consent. Consequently, privacy-by-design principles must be incorporated into traffic classification systems, encouraging decentralized and privacy-preserving learning approaches.

2.2.3 Threat Models in Traffic Analysis

Secure traffic classification must be evaluated under clearly defined threat models. Common adversarial assumptions include honest-but-curious servers, malicious clients, external eavesdroppers, and insider threats. Attack vectors may involve gradient reconstruction, activation leakage, or traffic correlation attacks.

Threat modeling is essential for understanding the privacy guarantees of machine learning-based classification frameworks. Without rigorous adversarial analysis, systems may provide a false sense of security while remaining vulnerable to inference-based attacks.

2.3 Distributed Learning Paradigms

2.3.1 Centralized vs Distributed Learning

Centralized learning aggregates raw data at a single server for model training. While this approach simplifies optimization and coordination, it introduces significant communication overhead and data exposure risks. In contrast, distributed learning retains data locally and shares model-related information instead of raw inputs.

Distributed paradigms enhance scalability and align with privacy requirements, particularly in multi-organization environments where direct data sharing is restricted (Kairouz et al., 2021). However, distributed training introduces challenges in synchronization, communication efficiency, and robustness.

2.3.2 Federated Learning Overview

Federated Learning (FL) is a decentralized learning framework in which clients locally train full models and share parameter updates with a coordinating server. The server aggregates updates—commonly using the Federated Averaging (FedAvg) algorithm—to construct a global model (McMahan et al., 2017).

FL reduces the need for raw data transmission and has been explored for network intrusion detection and traffic classification tasks. Nevertheless, research has demonstrated that shared gradients may still leak sensitive information under certain attack models, necessitating additional protective mechanisms such as secure aggregation.

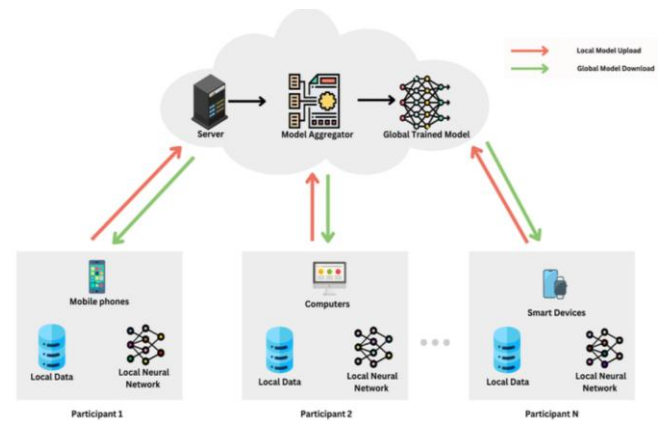


Figure-1: Federated Learning Architecture

2.3.3 Differential Privacy Mechanisms

Differential Privacy (DP) provides a mathematically rigorous framework for limiting information leakage from statistical outputs. By injecting calibrated noise into model gradients or query responses, DP ensures that the inclusion or exclusion of a single data record does not significantly affect model outputs (Dwork et al., 2014).

In traffic classification, DP mechanisms can protect sensitive traffic patterns during collaborative learning. However, privacy guarantees come at the cost of reduced model accuracy, especially when noise levels are high. Balancing privacy budgets and classification performance remains an active research challenge in secure network analytics.

3. SPLIT LEARNING: ARCHITECTURAL AND THEORETICAL OVERVIEW

Split learning has emerged as a promising distributed deep learning paradigm designed to address privacy, communication efficiency, and computational constraints in collaborative environments. Unlike traditional distributed learning frameworks that replicate full models across participants, split learning partitions a neural network between clients and a central server. This structural separation allows sensitive data to remain local while enabling joint model optimization. The following subsections discuss its architectural mechanisms, design variants, and associated privacy considerations.

3.1 Core Mechanism of Split Learning

3.1.1 Model Partitioning

The defining feature of split learning is model partitioning. A deep neural network is divided at a predefined “cut layer” into two segments: the client-side sub network and the server-side sub network. The client processes raw input data through its local layers and transmits intermediate activations to the server. The server completes the forward pass and computes the loss function.

This partitioning significantly reduces the need for transmitting raw data across network boundaries. The concept was formally introduced to enable privacy-preserving deep learning in distributed medical and enterprise environments, demonstrating that sensitive datasets can remain local without compromising collaborative training efficiency (Vepakomma et al., 2018). Furthermore, partitioning allows lightweight clients to offload computationally intensive layers to more powerful servers, improving scalability in edge-based deployments.

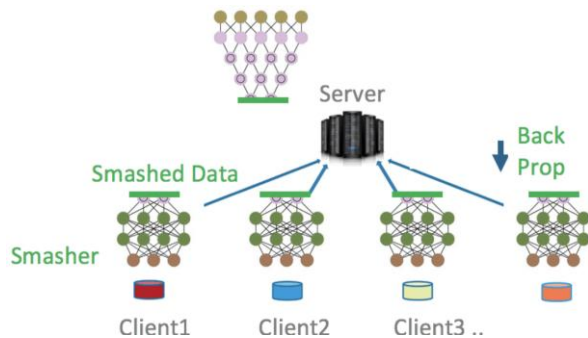


Figure-2: Split Learning Architecture

3.1.2 Forward and Backward Propagation Workflow

In split learning, training follows a coordinated forward-backward propagation workflow. During the forward pass, the client computes activations up to the cut layer and transmits them to the server. The server completes the forward propagation, evaluates the loss, and initiates backpropagation. Gradients corresponding to the cut layer are sent back to the client, which then updates its local parameters.

This sequential interaction ensures that raw input data never leaves the client environment. Compared to federated learning—where complete model gradients are shared—split learning only exchanges intermediate representations and partial gradients. This design reduces client-side computational load and may lower communication complexity in certain configurations (Gupta and Raskar, 2018). However, the sequential dependency between client and server introduces latency considerations that must be managed in large-scale deployments.

3.2 Variants of Split Learning

3.2.1 Vanilla Split Learning

Vanilla split learning represents the original implementation of the paradigm, involving a single client and a central server. In this configuration, only one participant performs local computation before transmitting activations. Although conceptually straightforward, vanilla split learning primarily serves as a baseline architecture for evaluating privacy and efficiency characteristics.

Its applicability is limited in multi-client scenarios unless extended with scheduling mechanisms. Nevertheless, it provides foundational insights into communication overhead and information leakage dynamics.

3.2.2 Multi-Split Configurations

To accommodate multiple data owners, multi-client split learning architectures have been proposed. In these configurations, multiple clients sequentially or asynchronously interact with a shared server-side model. Some designs incorporate client-side weight synchronization or parallelized cut-layer strategies to improve scalability.

Advanced variants also introduce techniques such as activation compression and pipelining to mitigate communication bottlenecks. These enhancements aim to make split learning viable in bandwidth-constrained or latency-sensitive environments, particularly in IoT and edge computing scenarios (Singh et al., 2019). However, coordination complexity increases as the number of participating clients grows.

3.2.3 Hybrid FL-Split Architectures

Hybrid architectures combine elements of federated learning and split learning to leverage the advantages of both paradigms. In such systems, clients may train partial models locally using federated averaging while still splitting deeper layers with a server. This approach reduces sequential dependency and can enhance parallelization.

Hybrid frameworks are particularly useful when computational resources are heterogeneous across clients. By integrating model partitioning with federated aggregation, these architectures aim to balance privacy, scalability, and training efficiency. Recent studies have explored hybrid designs to strengthen resilience against inference attacks while maintaining acceptable model convergence rates (Thapa et al., 2022).

3.3 Privacy Guarantees and Vulnerabilities

3.3.1 Intermediate Activation Leakage

Although split learning prevents direct sharing of raw data, the transmission of intermediate activations introduces potential privacy risks. Research has demonstrated that under certain conditions, adversaries may reconstruct approximate input data from shared activations, especially when the cut layer is shallow (Hitaj et al., 2017).

The risk of activation leakage depends on network architecture, depth of partitioning, and adversarial capabilities. Deeper cut layers generally reduce reconstruction feasibility but may increase client computational burden. Therefore, selecting an optimal partition point is critical for balancing privacy and efficiency.

3.3.2 Gradient-Based Inference Risks

In addition to activation leakage, gradient-based attacks pose significant threats to distributed learning frameworks. Gradient inversion techniques can potentially recover sensitive input information from shared gradients during backpropagation. Although such attacks were initially studied in federated learning contexts, similar vulnerabilities may arise in split learning if adversaries gain access to intermediate gradients (Zhu et al., 2019).

Mitigation strategies include adding noise, applying differential privacy mechanisms, or encrypting intermediate representations. However, these defenses introduce trade-offs in computational cost and model accuracy. Consequently, while split learning enhances privacy compared to centralized training, it does not inherently guarantee complete protection against sophisticated inference attacks.

4. LITERATURE REVIEW

This section critically synthesizes prior research on traffic classification and its evolution toward privacy-preserving distributed learning paradigms, with particular emphasis on split learning-based secure traffic analytics.

4.1 Conventional and ML-Based Traffic Classification Studies

4.1.1 Survey of Classical ML Models

Before the widespread adoption of deep learning, traffic classification primarily relied on classical machine learning algorithms trained on flow-level statistical features. Techniques such as Support Vector Machines (SVM), Decision Trees, Random Forests, and k-Nearest Neighbors were widely used due to their interpretability and moderate computational requirements. Early empirical evaluations demonstrated that statistical feature-based classifiers could achieve competitive accuracy in identifying application-layer protocols without payload inspection (Nguyen and Armitage, 2008).

However, these approaches were heavily dependent on manual feature engineering and struggled with encrypted or obfuscated traffic. Additionally, model performance often degraded when exposed to evolving traffic patterns or zero-day applications, highlighting generalization limitations in dynamic network environments.

4.1.2 Deep Learning-Based Encrypted Traffic Classification

The rise of encrypted traffic accelerated the transition toward deep learning architectures capable of automated feature extraction. Convolutional Neural Networks (CNNs) have been employed to analyze raw packet byte sequences, while Recurrent Neural Networks (RNNs) and Long Short-

Term Memory (LSTM) networks capture temporal dependencies in traffic flows. Deep Packet, for example, demonstrated the feasibility of end-to-end deep learning for encrypted traffic classification without handcrafted features (Lotfollahi et al., 2020).

Despite improved classification accuracy, centralized deep learning models introduce privacy risks due to the need for large-scale data aggregation. Moreover, high computational demands and storage requirements limit their deployment in distributed or resource-constrained environments.

4.2 Privacy-Preserving Traffic Classification Approaches

4.2.1 Federated Learning-Based Solutions

Federated Learning (FL) has been widely explored as a privacy-aware alternative to centralized training. In FL-based traffic classification systems, each client trains a local model on its own network data and shares model updates with a central aggregator using algorithms such as Federated Averaging (McMahan et al., 2017).

Several studies have applied FL to intrusion detection and encrypted traffic classification, demonstrating reduced data exposure while maintaining competitive performance. However, subsequent analyses revealed that shared gradients may leak sensitive information under adversarial settings, raising concerns about gradient inversion and membership inference vulnerabilities (Zhu et al., 2019).

4.2.2 Differential Privacy-Based Methods

Differential Privacy (DP) introduces mathematically quantifiable privacy guarantees by injecting calibrated noise into model parameters or gradients. In traffic classification, DP mechanisms have been integrated into distributed learning frameworks to limit information leakage from shared updates (Dwork et al., 2014).

While DP enhances privacy protection, the added noise may degrade classification accuracy, particularly in complex encrypted traffic scenarios. Achieving an optimal privacy-utility balance remains an ongoing research challenge, especially when strict privacy budgets are enforced.

4.2.3 Secure Multi-Party Computation Approaches

Secure Multi-Party Computation (SMPC) enables collaborative computation without revealing private inputs among participating entities. Cryptographic techniques such as secret sharing and homomorphic encryption have been explored to train traffic classification models securely across multiple organizations.

Although SMPC provides strong theoretical security guarantees, its computational overhead and communication complexity often limit scalability in high-throughput

network environments. Consequently, practical deployment in real-time traffic monitoring systems remains challenging.

4.3 Split Learning–Based Traffic Classification

4.3.1 Early Adoption Studies

Architectural Setups

Initial studies applying split learning to network security tasks adopted client–server neural network partitioning models, where early layers were executed locally and deeper layers were processed centrally. This configuration aimed to prevent raw packet exposure while leveraging centralized computational resources. Foundational work demonstrated the feasibility of split learning in privacy-sensitive domains, including healthcare analytics, establishing the architectural principles later adapted for network security applications (Vepakomma et al., 2018).

Dataset Usage

Early implementations primarily relied on benchmark intrusion detection and traffic classification datasets such as NSL-KDD, UNSW-NB15, and CICIDS2017. These datasets provided labeled flow-level features suitable for evaluating distributed architectures. However, limited experimentation on real-world encrypted traffic datasets constrained external validity.

Reported Performance

Empirical results indicated that split learning could achieve classification accuracy comparable to centralized deep learning models while reducing raw data exposure. Nonetheless, communication overhead and sequential training dependencies were identified as potential bottlenecks.

4.3.2 Advanced Frameworks and Enhancements

Communication-Efficient Models

Recent advancements have focused on reducing communication latency through activation compression, model pruning, and pipelining techniques. These enhancements aim to optimize bandwidth utilization in distributed environments, particularly for IoT-based traffic monitoring systems (Singh et al., 2019).

Secure Aggregation Mechanisms

To mitigate intermediate information leakage, secure aggregation and encryption mechanisms have been integrated into split learning frameworks. These methods encrypt activations or gradients before transmission, strengthening resilience against eavesdropping and malicious servers.

Adversarial Defense Integration

Advanced frameworks incorporate adversarial training and noise injection strategies to defend against inference attacks. Hybrid architectures combining split learning with federated learning have also been proposed to improve robustness and parallelization efficiency (Thapa et al., 2022).

4.3.3 Comparative Performance Analysis

Accuracy vs Privacy Trade-Off

Comparative analyses indicate that split learning offers a favorable balance between privacy preservation and classification accuracy relative to purely centralized approaches. However, privacy gains depend heavily on cut-layer depth and threat assumptions.

Computational Complexity

Client-side computational load in split learning is typically lower than in federated learning, as only partial models are trained locally. Nevertheless, sequential client–server interactions may increase training latency.

Scalability

Scalability remains a key concern in multi-client deployments. While model partitioning reduces data transmission volume, coordination overhead grows with the number of participants. Efficient scheduling and asynchronous communication protocols are therefore critical for large-scale network applications.

4.3.4 Dataset and Benchmark Analysis

Commonly Used Datasets

Most reviewed studies rely on public intrusion detection and traffic classification benchmarks such as CICIDS2017, UNSW-NB15, and ISCX VPN-nonVPN datasets. Although these datasets facilitate reproducibility, they may not fully capture contemporary encrypted traffic characteristics.

Evaluation Metrics

Performance is typically assessed using accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). However, few studies report privacy leakage metrics or communication cost evaluations, limiting comprehensive comparison.

Reproducibility Concerns

Reproducibility remains a challenge due to variations in preprocessing pipelines, feature extraction strategies, and split-layer configurations. Inconsistent experimental setups hinder fair benchmarking and cross-study validation.

5. CRITICAL ANALYSIS AND RESEARCH GAPS

Although split learning-based secure traffic classification has demonstrated promising capabilities in privacy-sensitive environments, several unresolved challenges remain. These challenges span technical efficiency, privacy robustness, and evaluation standardization. A critical examination of the literature reveals that existing frameworks often prioritize accuracy over holistic system design, leaving important research gaps in scalability, security guarantees, and benchmarking consistency.

5.1 Technical Limitations

5.1.1 Communication Overhead

One of the primary technical constraints of split learning is communication overhead resulting from iterative client-server interactions. During each training cycle, intermediate activations must be transmitted from the client to the server, followed by gradient updates returned to the client. In high-frequency traffic monitoring systems, this repeated bidirectional communication can generate significant latency and bandwidth consumption, particularly in wide-area or edge-based deployments (Singh et al., 2019).

Unlike federated learning, where model updates are typically exchanged periodically, split learning requires synchronous communication at every batch iteration. This sequential dependency limits throughput and may hinder real-time traffic classification. Although compression and quantization techniques have been proposed to reduce activation size, these methods introduce additional computational overhead and may affect numerical stability.

5.1.2 Model Convergence Challenges

Model convergence in split learning depends on stable coordination between distributed segments of the neural network. The partitioning of layers alters gradient propagation dynamics, potentially affecting optimization stability. Research in distributed optimization has shown that asynchronous updates, non-IID data distributions, and heterogeneous client resources can slow convergence or cause oscillatory behavior (Kairouz et al., 2021).

In traffic classification scenarios, where network behavior varies across domains, non-identically distributed data further complicates convergence. Existing studies often evaluate models under controlled experimental conditions, leaving open questions regarding robustness under real-world variability and large-scale multi-client deployments.

5.2 Privacy and Security Gaps

5.2.1 Gradient Leakage

Although split learning prevents direct sharing of raw traffic data, exchanged gradients and intermediate representations

may still leak sensitive information. Gradient inversion attacks have demonstrated that input samples can be reconstructed from shared gradients under certain assumptions (Zhu et al., 2019).

In split learning, similar vulnerabilities arise when adversaries gain access to cut-layer gradients. The extent of leakage depends on model architecture, activation dimensionality, and adversarial knowledge. While defensive strategies such as gradient clipping, noise injection, and encryption have been proposed, their integration into traffic classification systems remains limited. Consequently, the privacy guarantees of many split learning implementations are empirical rather than formally proven.

5.2.2 Model Inversion Attacks

Model inversion attacks attempt to reconstruct sensitive input features by exploiting access to trained model parameters or outputs. In distributed learning frameworks, adversaries may leverage prediction confidence scores or intermediate activations to infer private traffic characteristics. Studies in collaborative learning contexts have demonstrated that deep models can inadvertently memorize training data patterns, enabling partial reconstruction of inputs (Hitaj et al., 2017).

In privacy-sensitive network environments, such leakage may reveal behavioral signatures or communication metadata. Despite these risks, many existing split learning studies evaluate privacy qualitatively rather than through rigorous adversarial testing. The absence of standardized attack simulations represents a significant research gap.

5.3 Standardization and Benchmarking Issues

A critical limitation in current literature is the lack of standardized benchmarking frameworks for evaluating split learning-based traffic classification. Studies frequently employ different datasets, preprocessing pipelines, feature extraction strategies, and split-layer configurations. As a result, direct performance comparison across publications becomes challenging.

Moreover, evaluation metrics predominantly focus on classification accuracy, precision, recall, and F1-score, while neglecting privacy leakage quantification and communication cost analysis. Few works report metrics such as bandwidth consumption, latency, or privacy budgets, limiting comprehensive assessment. The absence of unified experimental protocols undermines reproducibility and impedes fair comparison among distributed learning paradigms.

6. OPEN CHALLENGES

Despite notable progress in split learning-based secure traffic classification, several open challenges hinder its seamless deployment in privacy-sensitive network

environments. These challenges extend beyond architectural design and encompass operational feasibility, generalization across heterogeneous domains, adversarial robustness, and regulatory compliance. Addressing these issues is essential for translating experimental prototypes into production-grade network security systems.

6.1 Real-Time Deployment Constraints

One of the most pressing challenges involves real-time deployment in high-throughput network environments. Traffic classification systems in enterprise backbones, 5G infrastructures, and cloud data centers must process massive volumes of packets with minimal latency. Split learning introduces sequential client-server interactions during training, which can generate additional communication delays compared to fully local inference systems (Singh et al., 2019).

Although inference can be optimized after training, dynamic environments often require continuous model updates to adapt to evolving traffic patterns. This iterative retraining process may conflict with strict latency requirements in real-time intrusion detection systems. Furthermore, edge devices participating in distributed training may have limited computational and energy resources, constraining the feasibility of deep model partitioning. Efficient scheduling, lightweight architectures, and hardware-aware optimization remain areas requiring further investigation.

6.2 Cross-Domain Traffic Adaptation

Network traffic characteristics vary significantly across domains, such as enterprise networks, IoT ecosystems, and mobile carrier infrastructures. Models trained on one dataset often experience performance degradation when deployed in a different operational environment due to domain shift and non-identically distributed (non-IID) data (Kairouz et al., 2021).

In split learning scenarios, cross-domain heterogeneity further complicates convergence and generalization. Clients may possess vastly different traffic distributions, encryption protocols, or application usage patterns. Without domain adaptation mechanisms, global models risk bias toward dominant participants. Techniques such as transfer learning, domain adversarial training, and meta-learning have been proposed in broader machine learning contexts, but their integration with split learning for traffic classification remains underexplored. Developing adaptive architectures capable of handling heterogeneous network environments constitutes a critical research direction.

6.3 Robustness Against Adversarial Attacks

Adversarial robustness represents another significant open challenge. Machine learning models for traffic classification are vulnerable to evasion and poisoning attacks, where

adversaries manipulate traffic patterns or training data to degrade detection accuracy. In distributed learning frameworks, malicious clients may inject poisoned updates to influence global model behavior (Bhagoji et al., 2019).

In split learning, threats may arise from compromised clients, curious servers, or external eavesdroppers intercepting intermediate activations. Additionally, adversarial examples crafted to mimic legitimate encrypted traffic can bypass classifiers without altering encryption protocols. Defensive strategies—including robust aggregation, anomaly detection for gradient updates, and adversarial training—require further adaptation to split learning architectures. Formal security proofs and systematic red-team evaluations are largely absent in current traffic classification studies.

6.4 Regulatory Compliance and Practical Integration

Regulatory compliance poses both technical and organizational challenges. Data protection frameworks emphasize accountability, transparency, and user consent in data processing activities. While distributed learning reduces raw data transfer, it does not automatically guarantee compliance if intermediate representations or model outputs reveal sensitive information (Voigt and Von dem Bussche, 2017).

Moreover, practical integration of split learning into existing network management systems requires interoperability with legacy monitoring tools, standardized APIs, and scalable orchestration frameworks. Enterprises must balance compliance requirements with operational efficiency, cost considerations, and cybersecurity objectives. Auditable privacy guarantees, explainable model behavior, and clear governance policies are essential for real-world adoption.

7. CONCLUSION

This review has systematically examined the evolution of secure traffic classification with a particular focus on split learning-based distributed architectures for privacy-sensitive networks. Traditional port-based and statistical approaches have gradually been replaced by deep learning models capable of handling encrypted and high-dimensional traffic patterns. However, centralized training paradigms introduce substantial privacy risks and regulatory challenges. Distributed learning frameworks, including federated learning and differential privacy mechanisms, represent important milestones toward privacy-aware analytics. Within this landscape, split learning offers a distinctive architectural advantage by partitioning neural networks between clients and servers, thereby minimizing direct exposure of raw traffic data.

The literature indicates that split learning can achieve classification performance comparable to centralized deep

learning while reducing data sharing risks. Nevertheless, its effectiveness depends heavily on model partitioning strategies, threat assumptions, communication efficiency, and deployment context. Comparative analyses reveal trade-offs among accuracy, scalability, and privacy guarantees, particularly in heterogeneous network environments. Although promising, split learning does not inherently eliminate risks such as gradient leakage or activation reconstruction. Future advancements must therefore integrate formal privacy guarantees, communication optimization, and adversarial robustness to enable practical large-scale adoption. Overall, split learning represents a viable and evolving paradigm for secure traffic classification, but sustained research efforts are necessary to address its architectural and operational constraints.

7.1. Limitations of the Review

This review is subject to several limitations. First, it primarily synthesizes peer-reviewed academic literature and may not fully capture proprietary industrial implementations of split learning in operational network environments. Second, variations in experimental setups, datasets, and evaluation metrics across studies limit the ability to provide strict quantitative comparisons. Third, rapid advancements in distributed learning and privacy-preserving techniques mean that newly emerging methods may not be comprehensively represented. Additionally, while privacy risks and adversarial threats are discussed conceptually, detailed empirical validation of attack resilience is beyond the scope of this review. Finally, the focus on split learning may underrepresent alternative cryptographic or hardware-based secure computation approaches that could also contribute to privacy-sensitive traffic classification.

REFERENCES

- 1) Anderson, B., Paul, S. and McGrew, D. (2017) 'Deciphering malware's use of TLS (without decryption)', *Journal of Computer Virology and Hacking Techniques*, 13(3), pp. 195–211.
- 2) Bhagoji, A.N., Chakraborty, S., Mittal, P. and Calo, S. (2019) 'Analyzing federated learning through an adversarial lens', *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 634–643.
- 3) Dwork, C., Roth, A., et al. (2014) 'The algorithmic foundations of differential privacy', *Foundations and Trends in Theoretical Computer Science*, 9(3–4), pp. 211–407.
- 4) Gupta, O. and Raskar, R. (2018) 'Distributed learning of deep neural network over multiple agents', *Journal of Network and Computer Applications*, 116, pp. 1–8.
- 5) Hitaj, B., Ateniese, G. and Perez-Cruz, F. (2017) 'Deep models under the GAN: Information leakage from collaborative deep learning', *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 603–618.
- 6) Kairouz, P., McMahan, H.B., Avent, B., et al. (2021) 'Advances and open problems in federated learning', *Foundations and Trends in Machine Learning*, 14(1–2), pp. 1–210.
- 7) Lotfollahi, M., Siavoshani, M.J., Zade, R.S.H. and Saberian, M. (2020) 'Deep Packet: A novel approach for encrypted traffic classification using deep learning', *Soft Computing*, 24(3), pp. 1999–2012.
- 8) McMahan, H.B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A.Y. (2017) 'Communication-efficient learning of deep networks from decentralized data', *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282.
- 9) Moore, A.W. and Papagiannaki, K. (2005) 'Toward the accurate identification of network applications', *Passive and Active Network Measurement (PAM)*, LNCS 3431, pp. 41–54.
- 10) Nguyen, T.T. and Armitage, G. (2008) 'A survey of techniques for internet traffic classification using machine learning', *IEEE Communications Surveys & Tutorials*, 10(4), pp. 56–76.
- 11) Shbair, W., Cholez, T., Francois, J. and Chrisment, I. (2016) 'A multi-level framework to identify HTTPS services', *IEEE Conference on Communications and Network Security (CNS)*, pp. 240–248.
- 12) Singh, A., Vepakomma, P., Gupta, O. and Raskar, R. (2019) 'Detailed comparison of communication efficiency of split learning and federated learning', *arXiv preprint arXiv:1909.09145*.
- 13) Thapa, C., Arachchige, P.C.M., Camtepe, S. and Sun, L. (2022) 'SplitFed: When federated learning meets split learning', *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8), pp. 8485–8493.
- 14) Vepakomma, P., Gupta, O., Swedish, T. and Raskar, R. (2018) 'Split learning for health: Distributed deep learning without sharing raw patient data', *arXiv preprint arXiv:1812.00564*.
- 15) Voigt, P. and Von dem Bussche, A. (2017) *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Cham: Springer.
- 16) Zhu, L., Liu, Z. and Han, S. (2019) 'Deep leakage from gradients', *Advances in Neural Information Processing Systems (NeurIPS)*, 32, pp. 14774–14784.

- 17) Li, P., Guo, C., Xing, Y., Shi, Y., Feng, L. and Zhou, F. (2024) 'Core network traffic prediction based on vertical federated learning and split learning', *Scientific Reports*, 14(1), p. 4663, doi:10.1038/s41598-024-53193-y.
- 18) Jin, Z., Duan, K., Chen, C., He, M., Jiang, S. and Xue, H. (2024) 'FedETC: Encrypted traffic classification based on federated learning', *Heliyon*, 10(16), e35962.
- 19) Trivedi, D., Boudguiga, A., Kaaniche, N. and Triandopoulos, N. (2026) 'SplitML: A unified privacy-preserving architecture for federated split-learning in heterogeneous environments', *Electronics*, 15(2), p. 267.
- 20) Alnasser, W., Beigi, G., Mosallanezhad, A. and Liu, H. (2022) 'PPSL: Privacy-preserving text classification for split learning', in *4th Int. Conf. on Data Intelligence and Security (ICDIS)*, IEEE, pp. 160–167.
- 21) Anonymous (2026) 'PrivRobust-SL: A privacy-preserving and adversarially robust split learning framework for IoT intrusion detection', *Journal Article*, ScienceDirect.
- 22) Shalabi, E., Khedr, W., Rushdy, E. and Salah, A. (2025) 'A comparative study of privacy-preserving techniques in federated learning: Performance and security analysis', *Information*, 16(3), p. 244.
- 23) Chaudhary, D., Rajasegarar, S. and Pokhrel, S.R. (2025) 'Towards Adapting Federated & Quantum Machine Learning for Network Intrusion Detection: A Survey', *arXiv preprint*.
- 24) Peng, Y., He, M. and Wang, Y. (2021) 'A federated semi-supervised learning approach for network traffic classification', *arXiv preprint*, arXiv:2107.03933.
- 25) Jiang, L., Wang, Y., Zheng, W., Jin, C., Li, Z. and Teo, S.G. (2022) 'LSTMSPLIT: Effective split learning based LSTM on sequential time-series data', *arXiv preprint*, arXiv:2203.04305.
- 26) Nguyen, P.T., Dao, N.N., Do, Q.T., Nguyen, T.V. and Cho, S. (Year) 'Privacy-Preserving Traffic Flow Prediction: A Split Learning Approach', *Conference Proceeding*, ElsevierPure.
- 27) (2025) 'Efficient privacy-preserving ML for IoT: Cluster-based split federated learning scheme for non-IID data', *Journal of Network and Computer Applications*, doi:10.1016/j.jnca.2025.104105.
- 28) (2024) 'Encrypted network traffic classification based on machine learning', *Ain Shams Engineering Journal*, 15(2), 102361.
- 29) (2025) 'Enhanced IoT security: Privacy-preserving federated learning model for accurate, real-time intrusion detection across devices', *ScienceDirect Article*.
- 30) (2024) 'Federated distributed network traffic classification based on deep mutual learning', *MDPI Electronics*, 14(24), p. 4928.