# Federated learning approach for multiple classes in network traffic classification

Aditya Kadam, Adesh Bhosale, Sejal Abraham, Prajwal Shankarshetty
*Prof. Sangeeta Alagi, Prof Milind Ankaleshwar ISBM COE, Pune, 412115*

*Abstract:* **Federated learning (FL) is an emerging paradigm in machine learning that allows decentralized model training across multiple devices or clients without the need to share raw data. This is especially valuable in scenarios where data privacy and security are paramount, such as in the case of network traffic classification. In the context of network traffic classification, FL enables multiple edge devices, such as routers, firewalls, or IoT devices, to collaboratively build a model capable of classifying traffic into multiple classes, such as normal traffic, Distributed Denial-of-Service (DDoS) attacks, malware, or other anomalous traffic. The key advantage of FL in this scenario is that it avoids the need to centralize sensitive traffic data, which can be privacy-sensitive and large in volume, especially when monitoring diverse network environments.**

**The federated learning process in network traffic classification begins with each client (e.g., an edge device or a network node) training a local model on its own network traffic data. The traffic data on each device may be highly heterogeneous, with some clients having more benign traffic while others might experience a higher volume of attack-related traffic. The local model is typically a neural network or another machine learning model that is trained to recognize patterns of normal and anomalous network behaviour Once the local model is trained on the device's data, only model updates (such as gradients or weight adjustments) are sent to a central server, rather than the raw network traffic data itself.**

## I. LITERATURESURVEY

Federated Learning in Network Traffic Classification.

Federated learning (FL) has emerged as a significant advancement in machine learning, particularly for applications where data privacy is paramount. Unlike traditional machine learning, which relies on centralized data gathering, FL trains models across decentralized devices without requiring data to be transmitted to a central server. This approach is particularly useful in network traffic classification, as network data is often sensitive and dispersed across various locations. The distributed nature of FL is ideal for handling this data while respecting user privacy, making it suitable for applications in mobile networks, IoT environments, and large-scale enterprise networks. Research has shown that by using federated learning, organizations can improve model robustness and classification accuracy while reducing privacy risks. Studies by researchers in the field highlight that FL's decentralized model also mitigates some of the costs associated with large-scale data transfers and storage, as data remains local to each device. This advantage is especially relevant in multiclass network traffic classification, where data is inherently diverse, sourced from a wide range of devices, and highly variable in its characteristics.

### A. Approaches and Algorithms for Multi-Class Network Traffic Classification

The challenge of multi-class classification in network traffic lies in accurately categorizing diverse traffic patterns that vary widely between applications, protocols, and user behaviours. Researchers have explored a variety of machine learning and deep learning algorithms adapted for federated learning frameworks to address this. Key techniques include federated averaging (Fed Avg) and more recent adaptive algorithms, which modify model weights dynamically based on the device's unique data distribution. Fed Avg, as one of the foundational algorithms in FL, has been widely used but is often limited by high variance in network traffic data. To counteract this, new algorithms incorporate methods to handle non-IID (non-independent and identically distributed) data, which is a common issue in multiclass traffic classification, as data collected from different devices and locations may not follow the same distribution. Additionally, some approaches leverage deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which have been adapted to work in federated settings and can better capture the spatial and temporal dependencies of network traffic. These methods have shown promise in achieving high accuracy in classifying multiple classes of network traffic while maintaining the benefits of federated learning.

### B. Challenges and Future Directions in Federated Learning for Network Traffic Classification

Despite the promise of FL, implementing it for multi-class network traffic classification presents several challenges. A primary concern is the need to address data heterogeneity, as network traffic data is typically non-IID across different devices and networks. This variation can lead to model inconsistency and performance degradation in federated settings. Another challenge involves managing the computational and communication overhead of federated learning, particularly in real-time applications where quick classification is essential. Additionally, ensuring data privacy remains a complex task, as FL models are still vulnerable to inference attacks and can unintentionally leak information about local data. Future research is focusing on developing more robust algorithms that can handle non-IID data more effectively and reduce communication costs through compression techniques and more efficient aggregation methods. Another promising direction lies in integrating advanced cryptographic methods, such as secure multiparty computation and differential privacy, to enhance data security further. Moreover, there is growing interest in hybrid models that combine federated learning with blockchain to add an additional layer of data integrity and auditability. By addressing these challenges, federated learning can become a more practical and scalable solution for multi-class network traffic classification across diverse real-world applications.

## II. OVERVIEW OF FEDERATED LEARNING APPROACH FOR MULTIPLE CLASSES IN NETWORK TRAFFIC CLASSIFICATION

### A. Introduction to Federated Learning for Network Traffic Classification

Federated learning (FL) is an innovative approach in machine learning that enables decentralized model training without centralizing the data. This is particularly beneficial for sensitive domains like network traffic classification, where data privacy and security are critical. In a typical network traffic classification task, data is spread across multiple devices and network nodes, each generating diverse traffic patterns based on user activities, applications, and protocols. Federated learning addresses this challenge by allowing the model to be trained on data stored locally on these devices, keeping the traffic data within its origin network, thus ensuring privacy. The global model is then aggregated periodically, without the need to transmit raw data across networks, which prevents data leakage and helps meet privacy regulations. This decentralized learning framework is highly suitable for the growing demand for privacy-preserving AI systems, especially in mobile networks, Internet of Things (IoT), and large enterprise environments. Federated learning for network traffic classification involves several steps: (1) local data processing on client devices or network nodes, (2) model training using local data, (3) model aggregation at a central server without sharing the raw data, and (4) the updated model is redistributed to all devices for further training. In the context of multi-class classification, each class represents a different type of network traffic, such as web browsing, email traffic, or real-time communication protocols like VoIP. The challenge lies in handling a wide variety of network traffic data that may exhibit different distributions across various devices. By leveraging FL, multiple devices can collaboratively train a model on these diverse traffic data without needing to share sensitive user information.

### B. Handling Multi-Class Traffic Classification Challenges through Federated Learning.

One of the most complex aspects of network traffic classification is handling the multiple types or "classes" of traffic. In real-world networks, traffic types range from benign traffic such as HTTP requests, video streaming, and social media usage to malicious traffic including DDoS attacks, botnet communications, and malware. Classifying traffic into multiple classes with high accuracy is an essential task in network security, yet it is fraught with challenges. These include data imbalance, the emergence of new traffic patterns, and the need for continuous adaptation. Federated learning plays a vital role in addressing these challenges, particularly in dealing with data imbalance. In network traffic datasets, some traffic classes are much more common than others, leading to imbalanced datasets that hinder the classification of less frequent traffic types. In traditional centralized learning, this imbalance can result in models that are biased towards the majority classes. In federated learning, however, the contribution of each participating device is essential, even for underrepresented traffic types. Each local model may specialize in its specific class of data, and through federated aggregation, the global model becomes more robust to imbalances. By allowing devices to individually focus on their local traffic data and then aggregating their contributions, FL reduces the risk of the global model favoring dominant classes and improves overall classification accuracy across all classes. In the context of multi-class traffic classification, FL is particularly effective because it allows devices to learn and classify traffic based on context-specific data. For instance, a smart home device may primarily generate traffic from video streaming or smart home applications, while a corporate network node may handle traffic related to enterprise software or secure communications. Federated learning enables these different devices to contribute to the global model, ensuring that the model accounts for the diversity of network traffic and can classify multiple traffic classes accurately. Moreover, FL enables local models to specialize in traffic classes that are unique to particular devices or networks, which enhances the precision of multi-class classification in a federated setting. Another advantage is that federated learning supports the continuous evolution of the model, which is crucial for keeping up with new and emerging traffic patterns. As new applications and protocols emerge, or as attack techniques evolve, federated learning allows for incremental updates to the model, ensuring that it adapts to new traffic types. This is in contrast to traditional methods that may require retraining from scratch or frequent finetuning, both of which are computationally expensive and time-consuming. In FL, new traffic patterns detected in a single device can be propagated across the entire system without the need to share raw traffic data, making it highly efficient for dynamic, real-time network environments.

### C. Challenges, Solutions, and Future Directions in Federated Learning for Network Traffic Classification

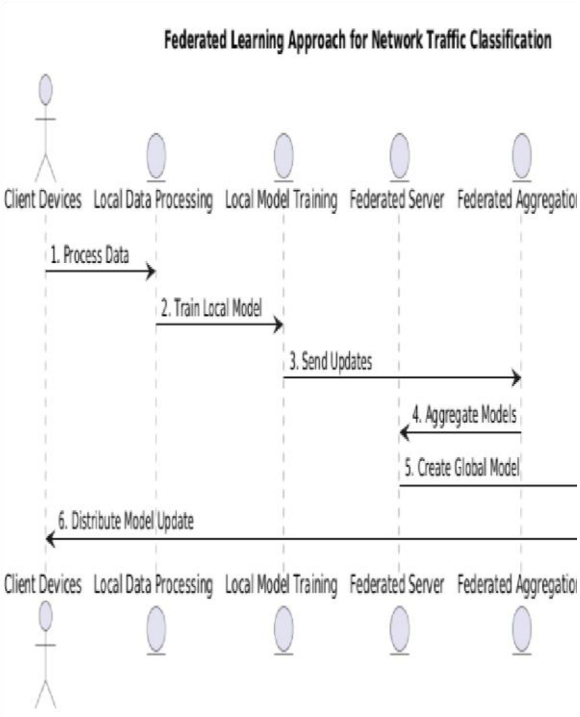
Federated learning for multi-class network traffic classification is not without significant challenges, particularly when dealing with the unique characteristics of network traffic data. A major issue is the problem of non-IID data. As mentioned, devices within a network generate different types of traffic based on user behavior, device type, and network environment. Since each local dataset is likely to be skewed, this heterogeneity can affect the performance of federated learning models. The aggregation of diverse models trained on such non-IID data may lead to poor generalization, especially for underrepresented traffic classes. To tackle this issue, various strategies are being proposed. One approach is to implement personalized federated learning. In this model, local devices have the ability to fine-tune their models based on the specific characteristics of the data they encounter. This personalization ensures that the model is better suited to the unique traffic patterns of each device, improving classification accuracy. Moreover, techniques like federated transfer learning are being explored, where a global model pre-trained on one set of traffic data can be finetuned locally on each device to account for the idiosyncratic nature of the local traffic patterns. Another challenge is communication overhead. Federated learning requires that local models communicate their updates to the central server, and as the number of devices scales, this communication can become burdensome. While FL reduces the need to transfer raw data, transmitting frequent model updates, particularly in networks with a large number of devices, still consumes substantial bandwidth and increases latency. To address this, research is focusing on techniques to reduce communication costs, such as model compression and federated distillation, where updates are compressed before being transmitted, or where a distilled model (a smaller, more efficient version) is shared instead of full models. These methods help maintain efficiency and reduce the time and resources needed to train global models. In terms of privacy and security, although federated learning inherently reduces risks related to data leakage by keeping data local, it still faces vulnerabilities, such as model poisoning attacks. In such attacks, malicious nodes send incorrect updates to the central server, which can degrade the performance of the global model. To mitigate this, researchers are looking at secure aggregation techniques, where model updates are encrypted before being sent to the central server, and only aggregated results are shared, ensuring the privacy of local updates. Additionally, federated learning frameworks can incorporate differential privacy to further prevent leakage of sensitive information from the model. Looking to the future, federated learning is expected to continue evolving, with a particular focus on improving its scalability and robustness. For instance, hybrid models that combine FL with blockchain technology are being explored as a way to provide both a decentralized and auditable system for traffic classification. Blockchain can enhance the integrity of the federated learning process by ensuring that model updates are transparent and traceable, which can be especially useful in scenarios where trust among the participants is essential. Additionally, the integration of edge computing with federated learning can enhance real-time traffic classification capabilities, as data processing and model training can be pushed closer to the devices, minimizing latency and reducing the reliance on central servers. As network traffic becomes increasingly diverse and sophisticated, federated learning will play a pivotal role in ensuring that classification models can keep pace with new developments in both network behaviour and attack vectors. By addressing the challenges of non-IID data, communication efficiency, and security concerns, federated learning will continue to be a powerful tool for scalable, privacy-preserving, multi-class network traffic classification.

## III. PROPOSED ARCHITECTURE

Proposed architecture for a federated learning approach in multiple classes for network traffic classification, I'll provide a structured format. Each major component and its role can be expanded upon based on research. Here's a starting point, divided into sections to achieve the in-depth detail you need. Each point will have a paragraph for elaboration.

### A. Decentralized Data Processing and Client Selection

In a federated learning (FL) framework for network traffic classification, the first stage of the architecture involves decentralized data processing. In this setting, data remains on each client device or local network segment, avoiding the need for data centralization while still allowing collaborative model building. Each client device represents a unique segment of network traffic, capturing various classes or types of network activity—such as HTTP, FTP, malicious traffic, or encrypted channels. To manage this diversity in traffic classes across clients, the architecture incorporates intelligent client selection mechanisms. These mechanisms select clients based on traffic class diversity, data quality, and device capability. By ensuring that clients represent a broad spectrum of traffic classes, the model achieves a balanced representation of network behaviors across devices, critical for effective classification. This decentralized approach provides privacy, minimizes data transfer overhead, and ensures that the federated model accurately reflects realworld network scenarios while maintaining data heterogeneity and privacy.



### B. Model Aggregation and Federated Averaging

Once the data is processed locally at each client, the model enters the federated learning cycle, where models trained on different client devices are aggregated at a central server. The core of the FL framework lies in this step, called federated averaging, where model parameters from each client are combined to form a global model. This federated averaging step takes into account weighted contributions from each client, considering the data volume and class diversity each client represents. The architecture ensures that clients receive periodic model updates, allowing them to retrain on local traffic while benefiting from global knowledge. The federated averaging approach also supports frequent model updates, enabling the architecture to adapt to evolving network traffic patterns dynamically. Given that network traffic constantly changes, frequent updates help the model stay relevant, accurately classifying new traffic patterns that emerge over time.

### C. Privacy Preservation and Security Measures

In federated learning, privacy preservation is essential, particularly when dealing with sensitive network traffic data. The proposed architecture incorporates differential privacy and secure aggregation methods to ensure data confidentiality. Differential privacy adds noise to model updates, making it challenging to trace specific data back to individual clients. Secure aggregation, on the other hand, ensures that client updates are encrypted, only revealing aggregate information to the central server. To handle malicious attacks—like model poisoning or adversarial attacks—the architecture includes anomaly detection techniques that identify and mitigate contributions from compromised clients. This security framework not only protects data privacy but also safeguards the model's integrity, ensuring that it remains robust even if some clients attempt to manipulate model updates. With privacy and security measures in place, the federated model can deliver reliable, unbiased traffic classification without compromising individual network data.

## IV. CONCLUSION

Federated Learning (FL) Represents a transformative approach to machine learning, particularly in the context of sensitive data such as network traffic. The growing demand for privacy-preserving technologies, combined with the increasing scale and diversity of network data, has driven the exploration of decentralized machine learning models. In this survey, we explored the potential of Federated Learning to classify multiple classes of network traffic in a way that ensures both privacy and accuracy while maintaining scalability and adaptability to dynamic network conditions. The approach’s key benefit lies in the ability to train models locally on client devices (such as routers, IoT devices, or other edge devices), keeping raw data within these devices and preventing the need for data centralization. This fundamentally shifts the paradigm of how machine learning models are traditionally developed and deployed, especially in the context of network traffic classification. One of the most significant challenges in network traffic classification is the vast heterogeneity of traffic types, which includes everything from benign web traffic to malicious intrusions, as well as encrypted communication. traffic data in one place often struggle to generalize across this diverse spectrum of traffic patterns, and they also pose privacy and scalability concerns. Federated Learning addresses these issues by decentralizing the training process, allowing each client device to train its local model on its specific subset of data. This localized training approach enables the model to adapt to local traffic patterns, improving classification accuracy without the need to share sensitive raw data. By aggregating the locally trained models at a central server, Federated Learning facilitates the creation of a robust global model that incorporates the knowledge learned across different clients. This process ensures that the resulting model is both personalized to each client’s traffic patterns and generalized enough to handle the variability inherent in network data from diverse sources. In terms of classification performance, the federated approach allows for the efficient handling of multiple classes in network traffic. Network traffic can be categorized into a variety of classes such as benign (normal) traffic, malicious traffic (e.g., DDoS, phishing), encrypted traffic, and anomalous traffic. Federated Learning allows for the incorporation of all these classes into a single framework, enhancing the system’s capability to differentiate between them accurately. The inclusion of multiple classes makes the network traffic classification more comprehensive and robust, as the model can learn to distinguish not only between regular and malicious traffic but also between different types of attacks or anomalies. This is crucial because new attack types or traffic patterns are continuously emerging, and a model that is trained using Federated Learning can quickly adapt by incorporating new client-specific data, keeping it up-to-date and relevant. Additionally, as more clients participate in the training process, the model benefits from an increasingly diverse set of data, which contributes to improved accuracy and generalization.

## REFERENCES

[1] https://ieeexplore.ieee.org/document/8473260

[2] https://ieeexplore.ieee.org/document/7724836

[3] https://ieeexplore.ieee.org/document/90687 17/