Federated Learning for DDoS Detection in SDN Networks with Explainable AI and Quantum-Inspired Models

Bouthaina Abid and Skandar Allouche
Department of Computer Engineering and Applied Mathematics
National Engineering School of Sfax (ENIS)
University of Sfax, Tunisia
{abidbouthaina2, alloucheskandar}@gmail.com

Dr. M. Adel Alimi
Department of Computer Engineering and Applied Mathematics
National Engineering School of Sfax (ENIS)
University of Sfax, Tunisia
adel.alimi@enis.tn

Takwa Omrani
REGIM-Lab: Research Groups in Intelligent Machines
University of Sfax, ENIS, BP 1173
Sfax, 3038, Tunisia
takwa.omrani@regim.usf.tn

***Abstract*— Software-Defined Networking (SDN) provides enhanced flexibility and centralized control but remains highly vulnerable to Distributed Denial of Service (DDoS) attacks. Conventional centralized machine learning solutions for DDoS detection raise critical concerns regarding data privacy, scalability, and transparency. To address these challenges, this paper proposes a Federated Learning (FL)–based framework for DDoS detection in SDN environments, enabling collaborative training without sharing raw data. Two federated models are evaluated: a classical linear Support Vector Machine (SVM) and a quantum-inspired learning model based on parameterized quantum circuits. In addition, Explainable Artificial Intelligence (XAI) is incorporated to enhance model interpretability by analyzing feature contributions to detection decisions. Experiments conducted on the Friday-WorkingHours-Afternoon-DDos subset of the CICDDoS2019 dataset demonstrate that federated learning achieves effective and privacy-preserving DDoS detection while highlighting the potential of quantum-inspired models and explainability for trustworthy SDN security systems.**

***Index Terms*— DDoS, SDN, Federated Learning, XAI, Quantum, SVM**

I. INTRODUCTION

SDN has become a key networking paradigm due to its programmability, centralized control, and dynamic resource management. By separating the control and data planes, SDN simplifies management and enables rapid security deployment, but its centralized nature also makes it vulnerable to DDoS attacks that can overload controllers and degrade performance.

Machine learning is widely used for DDoS detection to capture complex traffic patterns. However, most solutions rely on centralized learning, requiring large traffic aggregation, which raises privacy, trust, and scalability concerns in distributed SDN environments.

FL mitigates these issues by enabling collaborative training while keeping data local. In SDN security, FL allows domains to share knowledge without exposing sensitive traffic, offering a privacy-preserving solution for large-scale DDoS detection.

Model interpretability is also crucial, as operators need to trust decisions. Explainable AI (XAI) provides insights into model behavior and feature importance, essential in safety-critical SDN settings.

This paper proposes a federated learning framework for DDoS detection in SDN, comparing a classical linear SVM and a quantum-inspired variational circuit model, both trained with FedAvg. XAI analysis interprets model decisions and identifies key traffic features. Experiments on the Friday-WorkingHours-Afternoon-DDos subset of CICDDoS2019 show that FL enables accurate, privacy-preserving, and interpretable detection, while highlighting the potential of quantum-inspired models for SDN security.

II. LITERATURE SURVEY

A. Background

The rapid growth of SDN enables intelligent network management but also exposes critical security vulnerabilities, with DDoS attacks able to overwhelm centralized controllers. Machine learning is widely used to detect

anomalies in real time by capturing complex traffic patterns.

Traditional centralized ML faces privacy, scalability, and adversarial risks. FL addresses these by enabling collaborative training without sharing raw data, while XAI ensures transparency and interpretability, essential for trust in cybersecurity.

Quantum and quantum-inspired ML offer further potential for robustness and generalization. Combined, FL, XAI, and quantum-inspired techniques provide a promising foundation for next-generation SDN security solutions.

B. Literature Survey

Several recent studies have strongly influenced the design and direction of this work.

The study *Anticipating DDoS Incursions in Software-Defined Networking Using Explainable AI and Federated Learning* [1] demonstrated the effectiveness of combining FL with explainable models for DDoS detection in SDN environments. This work emphasized the importance of privacy-preserving collaboration across distributed controllers while maintaining interpretability for security operators. Its findings motivated the adoption of a FL framework in our project and reinforced the necessity of incorporating explainability to enhance trust in detection decisions.

The paper *Federated Learning in Adversarial Environments: Testbed Design and Poisoning Resilience in Cybersecurity* [2] provided key insights into the resilience of FL systems under adversarial conditions. By highlighting the challenges of data poisoning and model manipulation in collaborative learning settings, this study guided our choice of a federated architecture and informed the experimental design, particularly regarding client isolation and controlled aggregation strategies.

Furthermore, the work titled *Advancing Adversarial Robustness in Cybersecurity: Gradient-Free Attacks and Quantum-Inspired Defenses for Machine Learning Models* [3]explored the potential of quantum-inspired techniques to enhance model robustness against adversarial threats. This research inspired the integration of a quantum-inspired learning model in our study, enabling a comparative analysis between classical and quantum approaches within a FL framework.

Collectively, these studies shaped the conceptual foundation of our work by highlighting the complementary roles of FL, XAI, and quantum-inspired models in addressing privacy, robustness, and interpretability challenges in SDN-based DDoS detection. Building upon these contributions, our work aims to provide a unified and practical evaluation of these paradigms within a single experimental framework.

III. METHODOLOGY

This section describes the proposed methodology adopted for detecting DDoS attacks in SDN environments using FL. The approach integrates both classical and quantum-inspired machine learning models, complemented by XAI techniques to enhance transparency and interpretability.

A. Overall System of Architecture

The proposed system follows a federated learning architecture composed of multiple distributed clients and a central aggregation server. Each client represents an independent data holder (e.g., SDN domains or network segments) that locally trains a model on its private data without sharing raw traffic information.

The learning process is iterative and consists of:

1) Local model training at each client,
2) Transmission of model parameters to the central server,
3) Aggregation of parameters using the Federated Averaging (FedAvg) strategy,
4) Redistribution of the updated global model to all clients.

B. Software and Tools

The implementation relies on the following technologies:

- Flower (FLWR) framework for federated learning simulation,
- Python as the primary programming language,
- Scikit-learn for classical machine learning models,
- PennyLane for quantum circuit modeling,
- NumPy and Pandas for numerical processing,
- Matplotlib for visualization,
- SHAP for explainable AI analysis.

All experiments are executed in a simulated environment without real-time SDN deployment due to resource constraints.

C. Dataset

The experiments are conducted using the Friday-WorkingHours-Afternoon-DDos subset of the CICDDoS2019 dataset, a widely adopted benchmark for network intrusion and DDoS detection research. The dataset contains 225,711 network flow samples, extracted from realistic packet capture files, with each record representing a network flow characterized by 85 traffic-related features.

The classification task is formulated as a binary problem, where:

- 0: Benign traffic
- 1: DDoS attack traffic

The dataset exhibits a moderately imbalanced class distribution, with approximately 57% attack traffic and 43% benign traffic. To ensure reliable and unbiased evaluation, the dataset is split into a training set of 180,568 samples and a test set of 45,143 samples, while preserving the original class distribution in both subsets. This balanced separation guarantees consistency between training and testing phases and allows fair comparison across different learning paradigms.

For the purposes of this study, a reduced subset of flow-level features is selected from the original 85 attributes to balance computational efficiency and detection performance. The feature selection process is guided by insights from the study *Anticipating DDoS Incursions in Software-Defined Networking Using Explainable AI and Federated Learning* [1], which informed the choice of representative traffic features, although the selected subset is adapted to the constraints and objectives of the present work. This ensures that the retained features preserve the essential temporal and statistical characteristics associated with DDoS attack behavior.

D. Feature Selection

A subset of eight flow-based features is selected based on prior studies on DDoS detection in SDN environments:

- Flow Duration
- Total Forward Packets
- Total Backward Packets
- Flow Inter-Arrival Time Mean
- Flow Bytes per Second
- Flow Packets per Second
- Forward Inter-Arrival Time Mean
- Backward Inter-Arrival Time Mean

These features capture temporal, volumetric, and behavioral characteristics commonly associated with DDoS attacks.

E. Federated Learning Strategy

FL is implemented using the FedAvg algorithm, where each client performs local training and sends model updates to the server. The server computes a weighted average of the received parameters based on the number of local samples.

Key FL parameters include:

- Number of clients: 6
- Number of rounds: 10 for Federated SVM and 25 for Quantum Federated Learning
- Full client participation per round

This strategy enables collaborative training while preserving data locality and privacy.

F. Quantum-Inspired Federated Learning Model

A variational quantum circuit (VQC) is used as a quantum-inspired classifier within the federated learning framework.

Quantum Model Design:

- One qubit per feature,
- Feature encoding using parameterized rotation gates,
- Trainable quantum layers composed of rotation gates,
- Measurement performed on a Pauli-Z observable.

Each client trains its quantum parameters locally using gradient-based optimization. The trained parameters are aggregated centrally using FedAvg, producing a global quantum model shared across all clients.

This approach explores the feasibility of quantum-inspired learning under federated constraints using classical quantum simulators.

G. Classical Federated Learning Model (SVM)

To provide a baseline for comparison, a linear SVM classifier is implemented using stochastic gradient descent.

Key characteristics:

- Hinge loss function,
- L2 regularization,
- Mini-batch local training using partial-fit,
- Federated aggregation of model coefficients.

This classical federated model serves as a reference to evaluate the effectiveness and behavior of the quantum-inspired approach.

H. XAI

Explainability is incorporated as a separate post-hoc analysis module. A Random Forest classifier is trained centrally on the same dataset and features to facilitate interpretability.

The SHAP (SHapley Additive exPlanations) framework is employed to:

- Quantify feature importance,
- Explain individual predictions,
- Provide global and local interpretability.

XAI is not integrated into the federated training loop; instead, it is used to analyze learned traffic patterns and enhance trust in the detection system.

IV. EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the proposed framework, focusing on the behavior of two federated learning models applied to DDoS detection: a classical federated SVM and a quantum federated learning (QFL) model. The analysis emphasizes global accuracy and loss evolution across communication rounds, as well as client-level performance trends.

All experiments are conducted using six federated clients coordinated by a central server employing the FedAvg aggregation strategy. The same dataset, feature subset, and client partitioning strategy are used for both models to ensure a fair and consistent comparison.

A. Federated SVM Experimental Results

The classical baseline is implemented using a linear SVM trained in a federated manner via Flower. Each client trains a local SVM using stochastic gradient descent and shares model parameters with the server at each round.

Key settings:

- Number of clients: 6
- Number of federated rounds: 10
- Local epochs per round: 3
- Batch size: 512
- Aggregation strategy: FedAvg
- Evaluation: centralized server-side evaluation on a global test set

This centralized evaluation avoids bias introduced by heterogeneous client test sets and provides a consistent global performance indicator.

Figure 1 illustrates the evolution of global accuracy over the federated training rounds. Starting from an initial accuracy of approximately 57.9%, the model rapidly improves during the first round, reaching over 80% accuracy. Subsequent rounds show gradual and stable improvements, converging to a final global accuracy of approximately 81.2% after 10 rounds.

This rapid convergence demonstrates the effectiveness of federated optimization for linear models when data distributions are balanced across clients. The stability observed after the early rounds indicates that the SVM reaches a near-optimal solution under the given feature space and data partitioning.

Figure 2 presents the evolution of the global hinge loss across training rounds. The loss decreases sharply from an initial value of approximately 0.93 to 0.47 after the first round, followed by a slow and steady decline toward 0.44.

The monotonic reduction in loss confirms that the federated optimization process is stable and that model updates from clients are consistently improving the global decision boundary.

Although evaluation is performed centrally, client participation is uniform across rounds. Each client contributes equally to parameter updates, ensuring balanced aggregation. A summary table (Table 1) reports per-round client participation and data distribution, confirming that all six clients are active in each communication round.

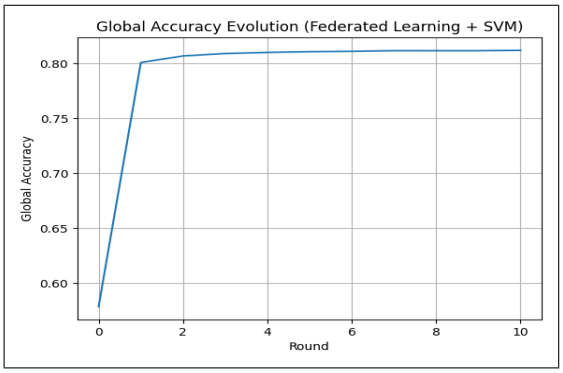

Fig. 1. Global Accuracy Evolution of Federated SVM

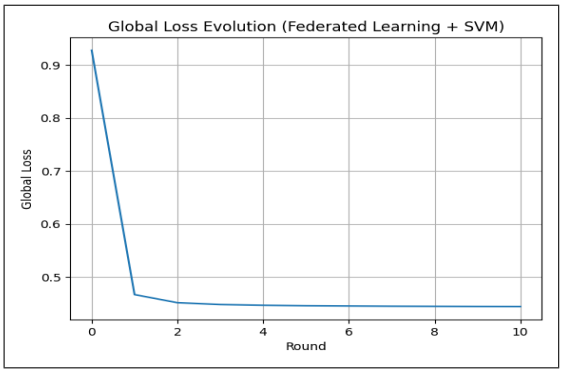

Fig. 2. Global Loss Evolution of Federated SVM

Round	Client ID	Accuracy
1	0	0.8066
	2	0.8093
	5	0.7957
3	1	0.8126
	3	0.8065
	4	0.8116
	5	0.8023
7	0	0.8173
	1	0.8146
	4	0.8143
10	0	0.8179
	1	0.8153
	2	0.8189
	3	0.8091
	4	0.8146
	5	0.8048

TABLE I

CLIENT PARTICIPATION AND ACCURACY PER ROUND

B. Quantum Federated Learning Experimental Results

The quantum model is implemented using a variational quantum circuit (VQC) with PennyLane, integrated into a federated learning setup using Flower.

Each client trains a local quantum model composed of:

- Angle encoding of classical features
- Parameterized rotation layers
- Measurement via expectation values

Key settings:

- Number of clients: 6
- Number of federated rounds: 25
- Aggregation strategy: FedAvg
- Evaluation: distributed client-side evaluation aggregated at the server

Unlike the SVM experiment, the quantum model requires a larger number of communication rounds to properly observe learning dynamics, due to the high variance, limited expressivity, and resource constraints inherent to near-term quantum models.

Figure 3 shows the evolution of global accuracy across 25 federated rounds. The model starts with an accuracy slightly above random guessing (54%). During the early rounds, accuracy fluctuates and temporarily decreases, reflecting the instability typical of variational quantum models trained with limited data and shallow circuits.

From round 10 onward, accuracy gradually stabilizes and improves, reaching a final global accuracy of approximately 57.2% at round 25. This behavior indicates that longer training horizons are necessary for quantum federated models to converge and exhibit meaningful learning trends.

Figure 4 depicts the evolution of the global loss over federated rounds. The loss initially increases during early rounds, highlighting optimization difficulty in the quantum parameter space. As training progresses, the loss gradually decreases and stabilizes after approximately 15 rounds.

Client-side accuracy values are computed at each round

and show noticeable variability across clients, especially in early rounds. This variability is expected due to:

- Small local datasets per client
- Sensitivity of quantum circuits to initialization
- Limited expressiveness of shallow quantum models

A summary of per-client accuracy per round is reported in Table II, illustrating the progressive alignment of client performances as federated training advances.

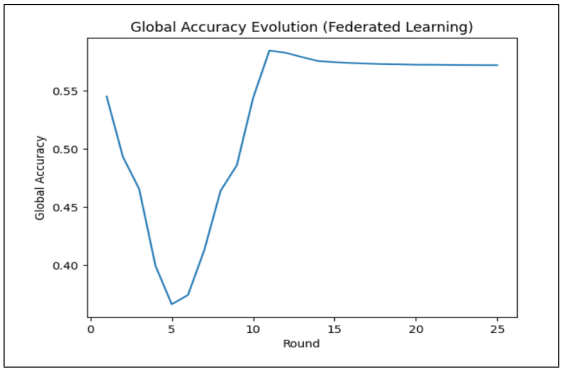

Fig. 3. Global Accuracy Evolution of Quantum Federated Learning

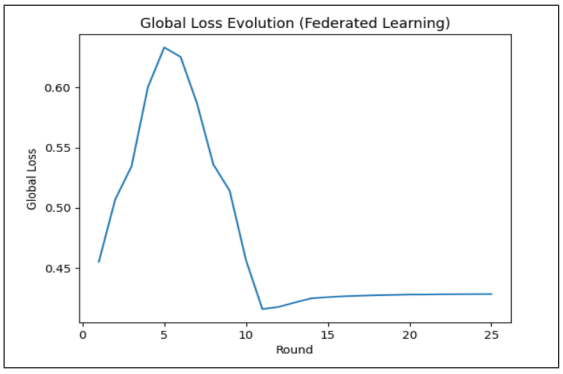

Fig. 4. Global Loss Evolution of Quantum Federated Learning

Round	Client ID	Accuracy
1	3	0.5444
	1	0.5428
	5	0.5460
2	3	0.4991
	2	0.4813
	4	0.4997
10	0	0.4876
	2	0.5413
	3	0.5420
25	5	0.5726
	2	0.5766
	4	0.5766
	1	0.5719

TABLE II

CLIENT PARTICIPATION AND ACCURACY PER ROUND

C. Comparative Analysis of Federated Models

This subsection compares the classical federated SVM and the quantum federated learning model in terms of convergence and learning behavior. The federated SVM converges quickly, reaching stable accuracy within the first rounds, while the quantum model requires more rounds to stabilize, which is characteristic of variational quantum algorithms.

Although the classical SVM achieves higher accuracy under the current setup, the quantum model shows consistent learning progression despite operating in a simulated environment and under stricter computational constraints. These results should be viewed as a proof of feasibility rather than a final performance benchmark.

Overall, the classical model offers efficiency and stability, whereas the quantum model highlights promising future directions for federated and privacy-preserving learning as quantum technologies mature.

D. XAI Analysis Using SHAP

To enhance the interpretability of the proposed federated SVM model, XAI is integrated using SHAP . SHAP is applied to the final global federated model obtained after the last training round, allowing explanations to be generated without accessing client-level data, thus preserving privacy.

Figure 5 presents the global feature importance using a SHAP bar plot. This visualization highlights which flow-level features contribute the most to the DDoS detection decision across the entire test set. Features related to packet rates, flow duration, and inter-arrival times are shown to have the strongest influence.

Figure 6 shows the SHAP summary plot, which illustrates both the importance and directional impact of each feature on the model's decision. This figure provides insight into how high or low feature values push predictions toward either benign or attack classes.

Finally, Figure 7 depicts a SHAP waterfall plot for an individual network flow. This local explanation decomposes a single prediction, showing how each feature contributes positively or negatively to the final decision, offering fine-grained interpretability at the sample level.

Overall, this XAI analysis demonstrates that the federated SVM model is not only effective but also transparent and interpretable, making its decisions understandable for security analysts in SDN environments.

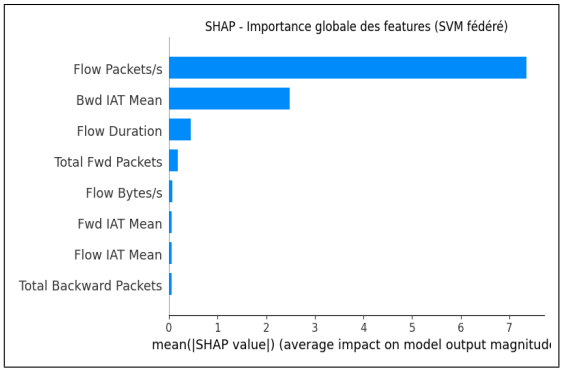

Fig. 5. Global Feature Importance for Federated SVM Using SHAP

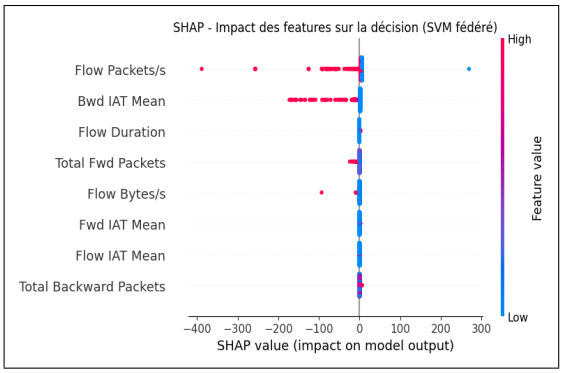

Fig. 6. SHAP Summary Plot Showing Feature Impact on DDoS Predictions

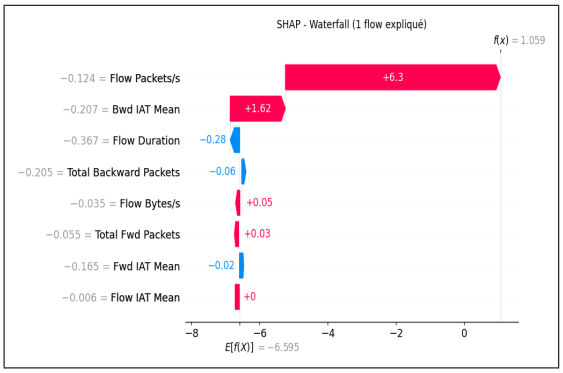

Fig. 7. Local Explanation of a Single Network Flow Using SHAP Waterfall Plot

V. CONCLUSION

This paper presented a privacy-preserving framework for DDoS detection in SDN by combining federated learning with explainable AI and a quantum-inspired learning baseline. Using the CICDDoS2019 Friday-WorkingHours-Afternoon-DDos subset, we evaluated a federated linear SVM and a federated variational quantum circuit model under a common FL architecture.

The federated SVM converged quickly and achieved stable global performance, while the quantum-inspired model demonstrated feasibility but required longer training and remained more sensitive to optimization constraints in simulation.

To support operational trust, SHAP explanations were generated on the final global model, providing both global feature importance and local, sample-level decision breakdowns suitable for analyst interpretation.

Overall, the study shows that federated learning can enable effective DDoS detection without centralizing sensitive traffic data, and that explainability adds transparency that is critical for SDN security deployment.

Future work will focus on evaluation in more realistic SDN testbeds, robustness to adversarial/poisoning behaviors, and improving quantum model stability and accuracy as quantum tooling and resources mature.

ACKNOWLEDGMENT

We would like to thank Dr. M. Adel Alimi for his valuable comments and suggestions that improved the quality of this paper. We are also grateful to Miss Takwa Omrani for her help in regularly reviewing our work. Finally, we thank the Department of Computer Engineering and Applied Mathematics, National Engineering School of Sfax, Tunisia, for providing the support that made this work possible.

REFERENCES

[1] N. Mehta, Anticipating DDoS Incursions in Software-Defined Networking Using Explainable AI and Federated Learning, Master's thesis, May 2023, doi: 10.13140/RG.2.2.33736.12803. Available: https://www.researchgate.net/profile/Nisarg-Mehta-7/publication/371125881_Anticipating_DDoS_Incursions_in_Software-Defined_Networking_Using_Explainable_AI_and_Federated_Learning/links/647476de59d5ad5f9c8377a1/Anticipating-DDoS-Incursions-in-Software-Defined-Networking-Using-pdf. [Accessed 25 Sept. 2025].

[2] H. J. Huang, H. T. Otal, and M. A. Canbaz, "Federated Learning in Adversarial Environments: Testbed Design and Poisoning Resilience in Cybersecurity," arXiv:2409.09794v2, Apr. 2025.Available: https://ieeexplore.ieee.org/abstract/document/11162297?casa_token=9XE0_QDaEk8AAAAA:TqtAMEV5mXopXZwhnose7wt8kD4-xWFwJPtGwqkh-QOqsaZoVcLJRpHwjennx-0pFK [Accessed 25 Sept. 2025].

[3] D. K. Kejriwal, A. Goel, and A. Sharma, "Advancing Adversarial Robustness in Cybersecurity: Gradient-Free Attacks and Quantum-Inspired Defenses for Machine Learning Models," International Journal of Innovative Science and Research Technology, vol. 10, no. 4, pp. 54–65, Apr. 2025, doi: 10.38124/ijisrt/25apr469. Available: https://www.researchgate.net/profile/Dk-Kumar-2/publication/390676557_Advancing_Adversarial_Robustness_in_Cybersecurity_Gradient-Free_Attacks_and_Quantum-Inspired_Defenses_for_Machine_Learning_Models/links/68026fd9df0e3f544f42b867/Advancing-Adversarial-Robustness-in-Cybersecurity-Gradient-Free-At pdf. [Accessed: 25 Sept 2025].