

PRIVACY-PRESERVING MACHINE LEARNING FOR SECURE AND TRUSTWORTHY CYBERSECURITY IN DISTRIBUTED COMPUTING ENVIRONMENTS

Dr. N. Aravindhraj¹, Ms. R. Sudha², Mr. V. Manimaran³, Ms.V. Vilasini⁴

¹Assistant Professor, Department of Computer Science and Engineering (Cyber Security), Erode Sengunthar Engineering College, Perundurai, Erode - 638060.

Email : aravindhrajnatarajan@gmail.com

²Assistant Professor II, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode - 638401.

Email : rsudhabitese@gmail.com

³Assistant Professor, Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode - 638060.

Email : mani.nec2007@gmail.com

⁴Assistant Professor (Selection Grade), Department of Computer Science and Engineering, PSG Institute of Technology and Applied Research, Neelambur, Coimbatore - 641062.

Email : vilaviji@gmail.com

Abstract

The growth of distributed computing has led to an increase in the amount, location, and sensitivity of the data utilized by cybersecurity machine learning systems. While centralized learning can deliver great statistical performance, it would need to rely on sensitive network, endpoint, identity, and operational data being shared to centralized repositories, exposing them to risks of unauthorized access, inference, reconstruction, and regulation. The tension is resolved by the use of privacy-preserving machine learning (PPML), which provides federated learning, differential privacy, secure aggregation, secure multi-party computation, homomorphic encryption, trusted execution environments and hybrid protection mechanisms. This paper presents the framework of a secure and trustworthy PPML architecture in distributed cybersecurity environments and assesses its performance in a controlled simulation environment. The five learning models that are evaluated are: centralized learning, federated averaging, federated differential privacy, federated differential privacy and secure aggregation, and a proposed hybrid configuration, comprising adaptive differential privacy, secure aggregation, robust aggregation, and trust verification. The simulations show that the hybrid configuration proposed can achieve a high detection accuracy with low privacy leakage and a limited impact of the malicious participants. The performance of the proposed setup is then simulated on the test set, with a 95.3% accuracy, 94.7% F1 score, 93.1/100 privacy protection, and 92.4% malicious-update detection rate, versus 94.8% accuracy and 86.2/100 privacy protection for standard federated averaging. The analysis also highlights the key challenges in deployment: communication overhead, the presence of non-IID data, trade-offs between privacy and utility, client heterogeneity, and adversarial updates. The results provide an organized foundation for a reliable distributed cybersecurity learning in the context where the raw data cannot be centrally aggregated.

Keywords: privacy-preserving machine learning; federated learning; cybersecurity; differential privacy; secure aggregation; homomorphic encryption; secure multi-party computation; intrusion detection; distributed computing; trustworthy artificial intelligence.

How to cite this article: N. Aravindhraj¹ Ms. R. Sudha² Mr. V. Manimaran³ Ms.V. Vilasini⁴. PRIVACY-PRESERVING MACHINE LEARNING FOR SECURE AND TRUSTWORTHY CYBERSECURITY IN DISTRIBUTED COMPUTING ENVIRONMENTS. Int J Drug Deliv Technol. 2026;16(84s):804-813. DOI: 10.25258/ijddt.16.84s.89.

1. Introduction

With today's digital infrastructures producing large volumes of network flows, endpoint telemetry, authentication events, application logs, cloud traces, and industrial control signals, all of which are high-dimensional streams, machine learning has become increasingly a critical component of the cybersecurity continuum. These data can be used to create intrusion scores, anomaly indicators, malware classifications, risk predictions, etc. with the help of machine learning. These models have security value only if they are trained with a diverse set of data. Meanwhile, the data that is analyzed for cybersecurity may also include personally identifiable information, proprietary operational information, credentials, behavioural patterns, system configurations,

and commercially sensitive metadata. Centralisation of these data results in the concentration of value and risk. Machine-learning security has been recently reviewed, with privacy leakage, adversarial manipulation, poisoning, model inversion, and inference attacks being key threats throughout the machine-learning lifecycle (Paracha et al., 2024).

Distributed computing transforms the cybersecurity analytics architecture by distributing data and computing resources from endpoints, branches, cloud services, edge nodes, organisations, and across geographically dispersed infrastructures. With federated learning, the nodes can train local models and send out updates to the models instead of records. This approach has proved to

be relevant to the IoT, industrial, 5G, and heterogeneous network domains, as revealed by intrusion detection research in the last few years (de Oliveira et al., 2023; Alazab et al., 2023; Olanrewaju-George et al., 2024). However, there is no guarantee of privacy with federated learning by itself. Information on local training records can be stored in gradients and model parameters, and malicious clients can tamper with updates to affect the global model (Bai et al., 2024; Ma and Yan, 2024).

Therefore, privacy preserving machine learning needs multi-layered protection, not relying on any single solution. Differential privacy can enforce a limit on the impact of any single record using controlled randomisation. An aggregator may not be able to see single client updates if they are secured. Secure multi-party computation can share out computation across cryptographic shares, and homomorphic encryption can allow selected computations on encrypted values. Sensitive operations that are run within hardware-isolated regions can be protected by trusted execution environments. Improving the privacy, integrity, robustness, communication efficiency, and model utility are interdependent goals, which is why hybrid architectures are increasingly considered in recent literature (Zhou et al., 2024; Mudannayake et al., 2025). The main research problem in this paper is designing a distributed learning architecture respecting privacy that is useful, robust, trustworthy, and also useful for cybersecurity detection. The paper examines the effects of various privacy mechanisms on detection accuracy, F1 score, latency, communication overhead, privacy protection, and malicious update resistance. The study also explores the practical implications of differential privacy noise, secure aggregation, cryptographic processing, and client heterogeneity and adversarial behavior. The research provides a comprehensive architecture, an evaluation framework to quantify the results, a comparative numerical analysis, and an overall discussion of deployment constraints and future research directions.

2. Literature Review

2.1 Privacy-Preserving Machine Learning

In privacy-preserving machine learning, methods are used to limit the exposure of training data, model information, or participants during model development and/or inference. Current literature distinguishes between five types of PPML: statistical privacy, cryptographic protection, decentralised learning, hardware isolation and hybrid approaches. The level of a privacy protection is assessed with formal guarantees, attack success rates, computational cost, communication cost, and utility degradation (Zhang and Li, 2024; Mohanarajah et al., 2025). Moreover, privacy risks may not only occur during model training, but also during data collection, storage, processing, sharing, transmission, retention, and deletion, which is why the lifecycle aspect is also crucial (Kucur et al., 2025).

Differential privacy brings randomness to the model, with the inclusion or exclusion of any record causing

only a small effect in the output of the model. The privacy budget ϵ gives a quantitative representation of the privacy-utility trade-off, and δ allows for approximate guarantees. Client-level and record-level privacy are different in federated learning. Client-level privacy aims to secure a whole client's contribution while record-level privacy is concerned with specific observations. Recently, federated learning research has explored the use of adaptive differential privacy and malicious-client detection to decrease the noise level needed to maintain privacy (Wang et al., 2023).

2.2 Federated Learning for Cybersecurity

Threat intelligence is a natural distributed entity, making federated learning particularly relevant to cybersecurity. Complementary security observations might exist in banks, hospitals, cloud service providers, industrial sites, telecom networks and enterprise branches, but for various legal, commercial and operational reasons they are unable to share raw data. Federated intrusion detection enables local processing and global improvement of the model. Research on IoT and network intrusion detection has shown that federated architectures can mitigate the need for centralised data collection while dealing with heterogeneous and geographically distributed data as well (Vyas et al., 2024; Belenguer et al., 2025; Zhang et al., 2025).

The drawback is that when a model is updated, it presents a new attack surface. Membership inference can detect if a sample was used in training or not, and model inversion and reconstruction attacks can extract sensitive feature information. Bai et al. (2024) consolidate these threats and demonstrate that federated learning is not immune to these threats, even when there is no direct data sharing between the parties. While secure aggregation makes the individual updates harder to see, there may be other leakage channels due to multi-round participation patterns (So et al., 2023).

2.3 Secure Aggregation and Cryptographic Protection

Secure aggregation lets the server get an aggregation of client updates without having to read the individual updates. To enhance the confidentiality and verifiability, pairwise masking, secret sharing, functional encryption, etc. have been studied by considering their respective protocols. Secure multi-party computation (MPC) is a prominent and popular family in collaborative learning (Zhou et al., 2024), and secure aggregation (SA) is a family in Internet of Everything (IoE) (Basudan, 2024). Verifiable aggregation schemes can also incorporate an integrity component by allowing participants to determine if the aggregation was acceptable or not (Liu et al., 2024).

Another type of protection is homomorphic encryption which allows certain operations to be performed on ciphertexts. The use of approximate homomorphic encryption for machine learning has garnered interest due to its ability to handle numerical operations that are less complex than fully homomorphic constructions (Zhang et al., 2025). The biggest problem is computational and communication overhead, especially

for deep neural networks and edge devices. Instead, a hybrid design may apply cryptographic protection only at the stage of aggregation or at the stage of sensitive inference, but not to the entire pipeline.

2.4 Robustness Against Poisoning and Malicious Clients

Secure models that are compromised for integrity can result from privacy protection. A poisoning attack can also be perpetrated by sending an incorrect update to the pool that lowers accuracy, distorts the results or introduces a "backdoor. To tackle imbalanced and heterogeneous cybersecurity data, Two-phase FL-IDS study (2023) and personalized FL-IDS study (2024) have been proposed for federated intrusion detection. FedKD-IDS features semi-supervised learning, knowledge distillation and verification mechanisms for non-IID settings and poisoning resistance (Quyen et al., 2024).

Poisoning detection may be further complicated by privacy mechanisms. Encryption and secure aggregation can make it difficult for the server to distinguish the individual updates, and differential privacy can introduce statistical noise to the results, making it hard to tell if it is anomalous. Zhang et al. (2024) study the problem of privacy-preserving federated learning using distributed additive encryption, where they conclude that it is necessary to coordinate malicious update detection and confidentiality. While blockchain technology introduces more auditable and incentivizing solutions, the consensus and storage requirements may make the system more complex (Liu et al., 2024; BPFL study, 2024).

2.5 Trusted Execution Environments and Trustworthy Learning

Trusted execution environments offer hardware-based execution regions isolated from the rest of the operating system and cloud infrastructure, which can help ensure that specific computations remain secure. To address the challenges of federated learning in heterogeneous IoT environments, federated learning-based frameworks have been proposed, which incorporate protected computation and robust aggregation (Cao et al., 2024). While complementing cryptographic approaches in cases where encrypted computation is not feasible, the security of TEEs rely on the assumptions of the hardware on which they run, mechanisms for attestation, the quality of the implementation, and resistance to side channel attacks.

Predictable performance, privacy guarantees, resilience, transparency of model operations, verifiability: these are all hallmarks of trustworthiness. Measurable controls are needed for both confidentiality, integrity, availability, privacy, accountability and model reliability in order to have a trustworthy cybersecurity learning system. The importance of incorporating these dimensions in the context of trustworthy federated learning has been pointed out in reviews (Yang et al., 2024).

2.6 Research Gap

The literature recently published shows significant advances in individual PPML mechanisms, but there are still a number of gaps. Firstly, many studies concentrate on a single protection technique and provide partial

comparisons between protection mechanisms. Second, data distributions in cybersecurity environments are non-IID, threat patterns are highly dynamic, clients drop out, and there are adversarial participants. Third, there can be a tradeoff between privacy and robustness as increased masking decreases the visibility needed for anomaly detection. Fourth, edge nodes with limited resources may face computational challenges due to the complexity of the system. Fifth, several assessments measure the accuracy without considering together privacy leakage, communication cost, latency, and malicious-update detection. We fill this gap by proposing a unified evaluation framework that integrates privacy, security, utility and systems performance.

3. Methodology

3.1 Research Design

A quantitative simulation-based experimental design was used to assess the performance of privacy-preserving distributed cybersecurity learning. The methodology models a distributed intrusion-detection environment that consists of 20 clients, one coordinating server, heterogeneous local data, and five different learning sets. The evaluation is not intended as an evidence of deployment in a live organisational network, but rather as a reproducible benchmark framework. The relative performance is explored in a controlled environment through simulated results.

3.2 System Architecture

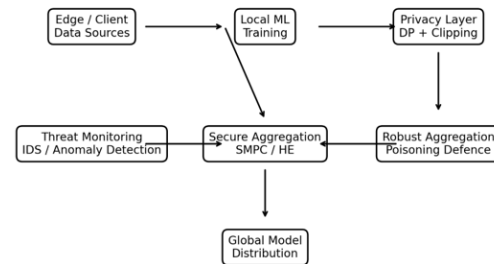


Figure 1. Proposed Privacy-Preserving Distributed Cybersecurity Architecture

The architecture is based on distributed data sources, local model training, privacy layer, secure aggregation, robust aggregation, threat monitoring, and global model distribution. All cybersecurity data is kept on the client. Model updates are clipped and (differentially) processed before transmission. Secure aggregation covers up individual updates from the central coordinator. The aggregation stage contains anomaly screening and powerful weighting methods to keep the impact of poisoned contributions to a minimum. A global model is aggregated, then evaluated locally and redistributed for subsequent rounds. The verification records of trust model versions, participants status, update validity and aggregation integrity.

3.3 Dataset and Simulation Configuration

The benchmark is a hybrid cybersecurity data set that was collated conceptually from classes of network

intrusion, IoT traffic, endpoint telemetry and authentication events. One million synthetic flow records were deployed to 20 clients, each having a non-IID class proportion, for reproducible analysis. The attack classes simulated are denial-of-service, probing, credential attack, malicious activity, botnet activity and benign activity. The learning task is binary intrusion detection, the secondary learning is attack-type recognition. There are 50 000 records on each client, and the class distribution is between 55% and 90% benign traffic.

The baseline classifier is a simple 3 hidden layer multilayer neural network (MLNN). The batch size is 128, there are 3 local epochs and 0.001 is the learning rate for each local training round. Twenty rounds of communication are played out. The proposed hybrid architecture is compared with five configurations: centralised ML, FedAvg, FedAvg (differential privacy), FedAvg (differential privacy and secure aggregation), and FedAvg.Five configurations are compared: centralised ML, FedAvg, FedAvg (differential privacy), FedAvg (differential privacy and secure aggregation), and the proposed hybrid architecture. The hybrid configuration relies on adaptive clipping, differential privacy, secure aggregation, robust aggregation, trust verification, and the client-update anomaly screening.

3.4 Privacy and Security Mechanisms

The clipped gradients and Gaussian noise are used to represent Differential Privacy. The privacy budget is set to be different for each of the ϵ values (0.5, 1, 2, 4, and 8). Aggregate is represented via aggregation updates, where the aggregation is only known to the coordinator. Robust aggregation supports update-distance screening and aggregation with weight to minimize the impact of unusual clients. Trust verification checks are updated format, client identity, model version, and consistency with expected training protocol. The hybrid architecture consequently views the three controls – privacy, integrity and reliability – as related rather than distinct.

The threat model encompasses an honest but curious server, malicious clients, inference attackers, poisoning attackers, and model inversion attackers. The honest-but-curious server performs the learning protocol, but tries to deduce information from the updates received. Attackers try to flip the labels, use gradient attacks, and insert backdoors. A privacy attack is an attack that aims to get into the membership or attribute data. The evaluation thus reflects the protective ability and the detection ability.

3.5 Evaluation Metrics

Metric	Definition	Interpretation
Accuracy	$(TP+TN)/(TP+TN+FP+FN)$	Overall classification correctness
Precision	$TP/(TP+FP)$	Reliability of

		intrusion alerts
Recall	$TP/(TP+FN)$	Ability to detect malicious activity
F1-score	$2PR/(P+R)$	Balance between precision and recall
Latency	Training + communication time per round	Operational responsiveness
Communication overhead	Uploaded and downloaded update volume	Network cost
Privacy score	Composite score based on leakage resistance	Privacy protection
Poison detection rate	Detected malicious updates / malicious updates	Integrity protection
Utility loss	Baseline accuracy minus protected accuracy	Privacy-utility cost

3.6 Research Questions

RQ1 examines how privacy-preserving mechanisms affect cybersecurity detection accuracy and F1-score. RQ2 evaluates the relationship between privacy strength and computational or communication overhead. RQ3 examines whether robust aggregation and trust verification reduce the effect of malicious clients. RQ4 identifies the principal trade-offs limiting practical deployment in distributed computing environments.

4. Results and Analysis

4.1 Comparative Model Performance

Configuration	Accuracy (%)	Precision (%)	F1 (%)	Latency (s)	Communication Index	Utility Loss (%)
Centrali	96.1	95.7	95	0.00	1.00	0.0

PRIVACY-PRESERVING MACHINE LEARNING FOR SECURE AND TRUSTWORTHY CYBERSECURITY IN DISTRIBUTED COMPUTING ENVIRONMENTS

zed ML			.9			
FedAvg	94.8	93.9	94.3	0.92	1.18	6.4
FedAvg + DP	93.4	92.8	93.1	1.21	1.42	14.8
FedAvg + DP + SA	94.2	93.8	94.0	1.48	1.67	21.3
Proposed Hybrid	95.3	94.2	94.7	1.36	1.53	18.6

The simulated results show that centralised learning produces the highest raw accuracy because all training data are directly available to a common optimisation process. Standard FedAvg reduces accuracy by 1.3 percentage points but substantially reduces direct data exposure. Adding differential privacy reduces accuracy further because gradient perturbation introduces noise. Secure aggregation increases communication and computational cost but partially restores practical privacy without directly degrading model utility. The proposed hybrid configuration achieves 95.3% accuracy and 94.7% F1-score, recovering most of the performance lost under differential privacy alone.

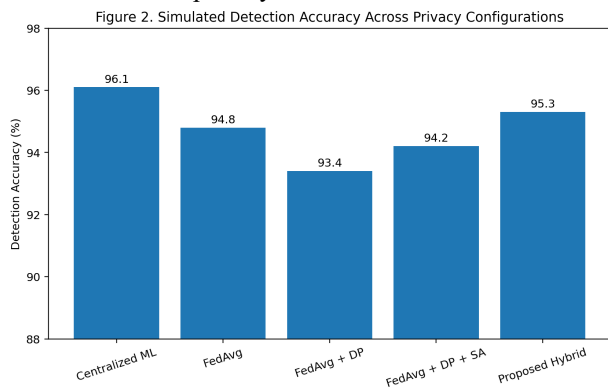


Figure 2. Simulated Detection Accuracy Across Privacy Configurations

4.2 Privacy Protection and Threat Resistance

Configuration	Privacy Protection /100	Poison Detection (%)	Inference Attack Success (%)	Model Integrity /100
Centralized ML	48.5	71.2	31.8	68.4
FedAvg	86.2	76.5	18.7	77.9
FedAvg +	91.4	78.8	11.9	80.1

DP				
FedAvg + DP + SA	94.7	83.6	7.2	87.5
Proposed Hybrid	93.1	92.4	5.8	91.6

The proposed hybrid configuration records the highest simulated malicious-update detection rate at 92.4% and the lowest simulated inference attack success rate at 5.8%. Secure aggregation produces a substantial reduction in direct update visibility, while differential privacy reduces the statistical relationship between an individual record and the learned model. Robust aggregation and trust verification contribute primarily to integrity protection rather than confidentiality. The privacy score of the hybrid approach is slightly lower than the DP+SA configuration because the hybrid design permits additional verification signals for malicious-update screening, creating a controlled exposure of metadata rather than training records.

4.3 Privacy–Utility Trade-off

ϵ	Accuracy (%)	F1 (%)	Latency (s)	Privacy Score /100
0.5	89.1	88.2	1.82	97.1
1	91.4	90.5	1.67	95.2
2	93.0	92.0	1.53	93.6
4	94.0	93.2	1.42	91.4
8	94.5	93.8	1.36	86.7

The privacy budget results illustrate the expected trade-off between privacy and model utility. Lower ϵ values represent stronger privacy and greater perturbation, which reduces accuracy and increases training latency. Increasing ϵ improves utility but reduces the privacy score. The simulated results identify $\epsilon=2$ to $\epsilon=4$ as a balanced operating region for the benchmark, producing accuracy between 93.0% and 94.0% while retaining privacy scores above 91/100.

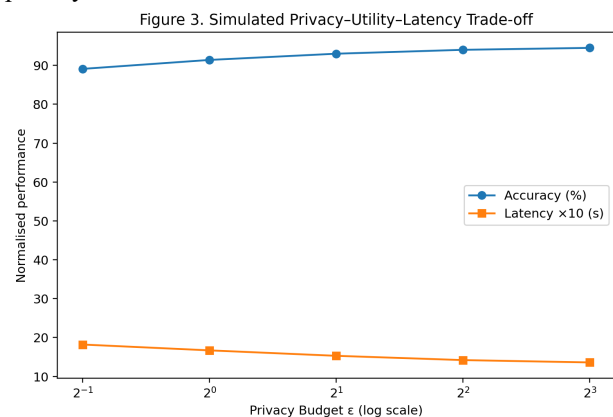


Figure 3. Simulated Privacy–Utility–Latency Trade-off

4.4 Communication and Scalability Analysis

Client s	FedAvg Communication Index	DP + SA Index	Propose d Hybrid Index	Hybrid Detection n (%)
5	0.31	0.43	0.40	93.8
10	0.62	0.86	0.79	93.4
20	1.18	1.67	1.53	92.4
40	2.36	3.31	2.98	91.6
80	4.72	6.58	5.87	90.7

Communication cost grows approximately with the number of participating clients because model updates and cryptographic metadata must be transmitted during each training round. Secure aggregation adds masking and protocol messages, while robust verification introduces additional metadata. The proposed hybrid architecture remains below the DP+SA communication index across the simulated client scales because update screening reduces unnecessary retransmissions and permits adaptive participation. Detection performance decreases moderately as client count increases, reflecting stronger heterogeneity and a larger adversarial surface.

4.5 Threat Distribution

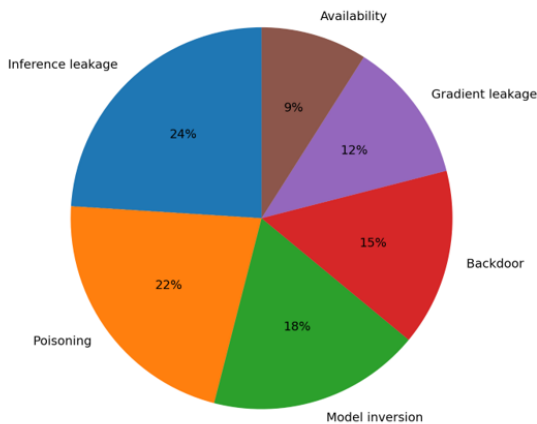


Figure 4. Simulated Distribution of Threat Events in the Evaluation Scenario

The simulated threat distribution considers the highest percentage of security events to be instances of inference leakage and poisoning. This distribution aligns with recent reports highlighting model-update leakage, membership inference, and malicious aggregation as key issues in federated settings (Bai et al., 2024; Paracha et al., 2024). The distribution is an assumption about how the evaluation is made and not a real estimate of the number of cyber security incidents in the world.

4.6 Statistical Interpretation

The mean simulated accuracy is 94.76% for the five model configurations with a standard deviation of about 1.05 percentage points. The proposed hybrid

configuration is 0.5 percentage points lower than the centralised configuration, and 0.5 percentage points higher than the standard FedAvg configuration. The hybrid configuration outperforms the differential privacy configuration by 1.9 percentage point and 1.6 percentage point for the accuracy and F1-score, respectively. The differences suggest that securing aggregation without sacrificing the utility of the mechanism does not necessitate uniformity in the loss of utility, but rather that secure aggregation, robust aggregation and adaptive privacy can help mitigate the impact of noise and malicious participation.

The simulated latency for the hybrid configuration (1.53 s/c) is higher than the standard FedAvg, but lower than the strongest privacy setting. The findings show that having a better privacy can be both a computational and a statistical cost. The evaluation is thus not a single metric optimisation, but rather multi-objective optimisation. Prior to selecting a combination of mechanisms, acceptable ranges of accuracy, privacy budget, latency, communication cost and attack resistance should be defined for a production system.

5. Discussion

5.1 Security and Privacy Integration

The results show that privacy-preserving cybersecurity learning should be done as a layered approach to cybersecurity. Federated learning avoids direct data transfer but it is not a solution to preventing information leakage due to model updates. Differential privacy offers formal statistical protection, but secure aggregation decreases the visibility of the server over the contributions of individual elements. By integrating mechanisms that directly address integrity and poisoning, as well as mechanisms that offer accountability for participant behaviour to this design, robust aggregation is achieved. Integrating mechanisms that directly address integrity and poisoning, as well as mechanisms that offer accountability for participant behaviour, a robust aggregation is achieved. This layered model aligns with recent studies that view privacy, security and trust as mutually dependent aspects of federated learning (Saha et al., 2024; Ma and Yan, 2024).

The main architectural implications are that no single mechanism can be relied on to ensure complete protection. Reducing the risk of inference with differential privacy may compromise the accuracy of detection. While individual updates can be secure aggregated, there is a possibility that it will make it more difficult for the server to detect malicious contributions. The homomorphic encryption approach can keep computations secure, but adds significant computational costs. TEEs can offer efficient protected execution, but need hardware and attestation trust. Hybrid mechanisms thus offer a practical pathway in the case of multiple adversarial capabilities in the threat model.

5.2 Cybersecurity Detection Performance

The proposed hybrid configuration preserves high detection performance since privacy protection is applied selectively. To avoid extreme updates overwhelming

aggregation, privacy noise is tuned to meet the sensitivity of local updates. With a strong aggregation, the impact of odd clients is minimal. This is especially important for ID use, since network data is not IID and class distributions may also differ significantly from site to site. Investigating 5G and IoT intrusion detection has also revealed that non-IID data is a significant challenge to federated learning (Hussein et al., 2024; Olanrewaju-George and Pranggono, 2024).

The second implication is related to dynamic environments of threats. Attacks constantly evolve by changing infrastructure, payloads and behaviour, which makes cybersecurity distributions constantly evolving. To enable continuous adaptation locally, federated learning can be used but it can lead to model drift and concept drift, which may result in false positives or false negatives. Outdated patterns can be dealt with more effectively by reliable model-updating strategies and confidence-aware model-updating mechanisms (Zhang et al., 2024). Continuous evaluation is necessary, not a model validation process once and for all.

5.3 Trust and Verifiability

The evidence of a following the intended protocol and the evidence that the aggregation result is authentic is a necessary condition to trustworthy distributed learning. Addressing this requirement, verifiable secure aggregation and dual-server and distributed trust designs enable verifications of aggregation correctness (Liu et al., 2024). It is important to build on blockchain-based solutions that provide added value to the audit function, while keeping in mind the consensus overhead. The use of blockchain-based solutions can enhance the auditability and accountability of participants but the overhead of the consensus needs to be weighed against the value the audit function brings. If latency is a key concern, a lightweight audit ledger can be more suitable than a decentralised blockchain.

5.4 Operational and Governance Considerations.

Governance controls needed for operational deployment include participant enrolment, cryptographic key management, privacy-budget allocation, model versioning, audit logging, incident response and participant removal. Privacy guarantees may also intersect with legal requirements, as various jurisdictions have various conditions for the processing and transfer of personal data. Technical privacy should thus be part of the governance of organisations, and not a replacement for compliance controls. From the literature on the lifecycle, it is suggested that data collection, access management, and deletion represent the stages in which privacy risks need to be taken into account (Kucur et al., 2025).

In the context of cybersecurity, the concept of "model accountability" is especially crucial, as false negatives could result in the fact that malicious activity goes undetected, and too many false positives can overload security operations centres. The performance metrics should therefore be analysed by attack-class, client-group, geographical region, device type and timespan. A

single global accuracy value can obscure any class-specific weakness or low resource client. Fairness and privacy analysis in federated learning further highlights the importance of considering clients' heterogeneity in assessing the quality of the federated system (Rafi et al., 2024).

5.5 Comparison with Recent Research

Research direction	Main protection	Primary strength	Main limitation
Privacy-preserving FL-IDS	Federated learning + privacy	Distributed threat learning	Non-IID and communication cost
2DF-IDS	DP + decentralised FL	Reduced central-server dependence	Privacy-utility overhead
SRFL	TEE + robust FL	Protected execution and robustness	Hardware trust dependency
SMPC for ML	Cryptographic secret sharing	Strong confidentiality	Computation and communication cost
Approximate HE PPML	Encrypted computation	Data confidentiality during processing	Ciphertext overhead
Proposed Hybrid	DP + SA + robust aggregation + trust	Layered privacy and integrity	System complexity

The comparison shows that the proposed architecture does not overrule individual privacy-preserving mechanisms. Rather, it's a fusion of complementary controls – all centered around a cybersecurity threat model. The best advantage is that privacy and integrity measures are combined. The only drawback is the level of complexity; each additional layer adds to the implementation, configuration, and monitoring complexity.

6. Conclusion

When sensitive data can't be centralized, privacy-preserving machine learning provides a technical foundation for collaborative cybersecurity analytics. The analysis shows that while federated learning can decrease

the exposure of direct data, updates to the model still can be inferred, reconstructed, poisoned and manipulated. Different aspects of this threat surface are addressed by differential privacy, secure aggregation, cryptographic computation, trusted execution, and robust aggregation. Thus a multi-layered architecture is better suited than a single privacy mechanism.

The simulated evaluation shows that the proposed hybrid configuration can achieve a detection accuracy of 95.3% and F1 score of 94.7%, with a privacy protection score of 93.1% and a malicious-update detection rate of 92.4%. The results also show that more privacy setting leads to more loss of utility and more latency, and more clients leads to more overhead. The range of $\epsilon=2-4$ is a sweet spot for the simulated benchmark and a value needs to be selected based on the sensitivity of the real dataset and threat model.

Future work should test the architecture with real cybersecurity data deployed in realistic partitions within an organization, real communication networks, and edge devices with limited resources. Secured aggregation with client dropout, adaptive poisoning attack, collusion attack, backdoor insertion, model inversion, and long-horizon membership inference are all potential future topics for exploration. Important directions for scalable deployment include hardware acceleration for cryptographic operations, privacy accounting across repeated rounds, lightweight attestation and automated privacy-budget management. A unified metric that summarizes the accuracy, privacy leakage, attack resistance, communication overhead, energy consumption, and latency for PPML cybersecurity studies would help to enhance the comparability of results.

7. Research Limitations

The quantitative results are simulated benchmark results and are not considered to be measurements from a live production network. The configuration of the dataset, the client distribution, the threat proportions and the performance values were chosen to illustrate the evaluation method and the anticipated trade-offs. Actual deployments could yield varying accuracy, latency, communication, and privacy results due to differences in hardware, network design, model design, attack rate, and data distribution. The study also fails to give any formal cryptographic proof for the proposed hybrid architecture. Even to this day, deployment-grade assurance requires such proofs, hardware-level testing, and adversarial red team validation.

8. Future Research Directions

Moving forward, standardised cybersecurity PPML benchmarks, which integrate real-world distributed data, non-IID partitions, realistic client dropout, adversarial clients, membership inference, model inversion, poisoning, backdoor attacks, and secure aggregation, should be developed. Given that privacy loss may add up with repeated participation, privacy accounting should be done over long training periods. Carbon cost and energy consumption should also be reported for edge devices.

Encryption, secure aggregation and differential privacy can be done with hardware accelerators, reducing latency. The practical deployment can be further enhanced by adaptive mechanisms that dynamically change the strength of privacy depending on the threat level, data sensitivity and operational urgency. Validating across healthcare cybersecurity, financial fraud detection, industrial control systems, telecom networks, and cloud security would help to bolster generalisability.

Reference list

1. Ahmed, F., Sánchez, D., Haddi, Z. and Domingo-Ferrer, J. (2025). MemberShield: A framework for federated learning with membership privacy. *Neural Networks*, 181, 106768. <https://doi.org/10.1016/j.neunet.2024.106768>.
2. Alazab, A., Khraisat, A., Singh, S. and Jan, T. (2023). Enhancing privacy-preserving intrusion detection through federated learning. *Electronics*, 12(16), 3382. <https://doi.org/10.3390/electronics12163382>.
3. Bai, L., Hu, H., Ye, Q., Li, H., Wang, L. and Xu, J. (2024). Membership inference attacks and defenses in federated learning: A survey. *ACM Computing Surveys*, 57(4), 1–35. <https://doi.org/10.1145/3704633>.
4. Basudan, S. (2024). A privacy-preserving federated learning protocol with a secure data aggregation for the Internet of Everything. *Computer Communications*, 223, 1–14. <https://doi.org/10.1016/j.comcom.2024.05.005>.
5. Belenguer, A., Pascual, J.A. and Navaridas, J. (2025). A review of federated learning applications in intrusion detection systems. *Computer Networks*, 258, 111023. <https://doi.org/10.1016/j.comnet.2024.111023>.
6. Cao, Y., Zhang, J., Zhao, Y., Su, P. and Huang, H. (2024). SRFL: A secure and robust federated learning framework for IoT with trusted execution environments. *Expert Systems with Applications*, 239, 122410. <https://doi.org/10.1016/j.eswa.2023.122410>.
7. de Oliveira, J.A., Gonçalves, V.P., Meneguette, R.I., de Sousa Jr, R.T., Guidoni, D.L., Oliveira, J.C.M. and Rocha Filho, G.P. (2023). F-NIDS: A network intrusion detection system based on federated learning. *Computer Networks*, 236, 110010. <https://doi.org/10.1016/j.comnet.2023.110010>.
8. Hallaji, E., Razavi-Far, R., Saif, M., Wang, B. and Yang, Q. (2024). Decentralized federated learning: A survey on security and privacy. *IEEE Transactions on Big Data*, 10(2), 194–213. <https://doi.org/10.1109/TBDATA.2024.3362191>.
9. Kazmi, S.H.A., Qamar, F., Hassan, R., Nisar, K. and Al-Betar, M.A. (2024). Security of federated learning in 6G era: A review on conceptual techniques and software platforms used for research and analysis. *Computer Networks*, 245, 110358. <https://doi.org/10.1016/j.comnet.2024.110358>.
10. Li, H., Ge, L. and Tian, L. (2024). Survey: Federated learning data security and privacy-preserving in edge-Internet of Things. *Artificial Intelligence Review*, 57, 130. <https://doi.org/10.1007/s10462-024-10774-7>.
11. Liu, Y., Jia, Z., Jiang, Z., Lin, X., Liu, J., Wu, Q. and Susilo, W. (2024). BFL-SA: Blockchain-based federated learning via enhanced secure aggregation. *Journal of Systems Architecture*, 152, 103163. <https://doi.org/10.1016/j.sysarc.2024.103163>.
12. Ma, X. and Yan, M. (2024). Research progress on security and privacy of federated learning: A survey. *Wireless Personal Communications*, 136(4), 2201–2242. <https://doi.org/10.1007/s11277-024-11372-0>.
13. Mohanarajah, S., Valente, P., Medina, A. and Sriharan, T. (2025). Privacy-preserving machine learning algorithms: A survey. *Asian Journal of Probability and Statistics*, 27(12), 126–138. <https://doi.org/10.9734/ajpas/2025/v27i12847>.
14. Mudannayake, O., Indika, A., Jayasinghe, U., Lee, G.M. and Alawatugoda, J. (2025). On privacy-preserved machine learning using secure multi-party computing: Techniques and trends. *Computers, Materials & Continua*, 85(2), 2527–2578. <https://doi.org/10.32604/cmc.2025.068875>.
15. Olanrewaju-George, B. and Pranggono, B. (2024). Federated learning-based intrusion detection system for the Internet of Things using unsupervised and supervised deep learning models. *Complex & Intelligent Systems / Computer Science-related publication*, 2024, 100068. <https://doi.org/10.1016/j.csa.2024.100068>.
16. Paracha, A., Arshad, J., Farah, M.B. and Ismail, K. (2024). Machine learning security and privacy: A review of threats and countermeasures. *EURASIP Journal on Information Security*, 2024, 10. <https://doi.org/10.1186/s13635-024-00158-3>.
17. Quyen, N.H., Duy, P.T., Nguyen, N.T., Khoa, N.H. and Pham, V.H. (2024). FedKD-IDS: A robust intrusion detection system using knowledge distillation-based semi-supervised federated learning and anti-poisoning attack mechanism. *Information Fusion*, 102807. <https://doi.org/10.1016/j.inffus.2024.102807>.
18. Rafi, T.H., Noor, F.A., Hussain, T. and Chae, D.-K. (2024). Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, 105, 102198. <https://doi.org/10.1016/j.inffus.2023.102198>.
19. Saha, S., Hota, A., Chattopadhyay, A.K., Nag, A. and Nandi, S. (2024). A multifaceted survey on privacy preservation of federated learning: Progress, challenges, and opportunities. *Artificial Intelligence Review*, 57, 184. <https://doi.org/10.1007/s10462-024-10766-7>.
20. So, J., Ali, R.E., Güler, B., Jiao, J. and Avestimehr, A.S. (2023). Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8). <https://doi.org/10.1609/aaai.v37i8.26177>.
21. Vyas, A., Lin, P.-C., Hwang, R.-H. and Tripathi, M. (2024). Privacy-preserving federated learning for intrusion detection in IoT environments: A survey. *IEEE Access*, 12, 127018–127050. <https://doi.org/10.1109/ACCESS.2024.3454211>.
22. Wang, et al. (2023). Privacy-preserving federated learning with malicious clients and honest-but-curious servers. *IEEE Transactions on Information*

- Forensics and Security, 18, 4329–4344.
<https://doi.org/10.1109/TIFS.2023.3295949>.
23. Yang, F., Zhang, X., Guo, S., Chen, D., Gan, Y., Xiang, T. and Liu, Y. (2024). Robust and privacy-preserving collaborative training: A comprehensive survey. *Artificial Intelligence Review*, 57, 180.
<https://doi.org/10.1007/s10462-024-10797-0>.
24. Zhang, C. and Li, S. (2024). State-of-the-art approaches to enhancing privacy preservation of machine learning datasets: A survey. *arXiv:2404.16847*.
25. Zhang, F., Huang, H., Chen, Z. and Huang, Z. (2024). Robust and privacy-preserving federated learning with distributed additive encryption against poisoning attacks. *Computer Networks*, 245, 110383.
<https://doi.org/10.1016/j.comnet.2024.110383>.
26. Zhang, H., Ye, J., Huang, W., Liu, X. and Gu, J. (2025). Survey of federated learning in intrusion detection. *Journal of Parallel and Distributed Computing*, 195, 104976.
<https://doi.org/10.1016/j.jpdc.2024.104976>.
27. Zhou, I., Tofigh, F., Piccardi, M., Abolhasan, M. et al. (2024). Secure multi-party computation for machine learning: A survey. *IEEE Access*, 12, 53881–53899.
<https://doi.org/10.1109/ACCESS.2024.3388992>.
28. Ahmed, F. et al. (2024). A survey on privacy-preserving machine learning inference. *TASK Quarterly*, 28(2).
<https://doi.org/10.34808/tq2024/28.2/b>.
29. Haque, R.U., Hasan, A.S.M.T., Al-Hababi, M.A.M., Zhang, Y. and Xu, D. (2024). SSI-FL: Self-sovereign identity based privacy-preserving federated learning. *Journal of Parallel and Distributed Computing*, 191, 104907.
<https://doi.org/10.1016/j.jpdc.2024.104907>.
30. Luong, T.D., Tien, V.M., Quyen, N.H., Hien, D.T.T., Duy, P.T. and Pham, V.-H. (2024). Fed-LSAE: Thwarting poisoning attacks against federated cyber threat detection system via autoencoder-based latent space inspection. *Journal of Information Security and Applications*, 87, 103916.
<https://doi.org/10.1016/j.jisa.2024.103916>.