National College of Ireland

Report

MSc Research Project
MSc Cyber Security

Ramani Katale
Student ID: X23320834

School of Computing
National College of Ireland

Supervisor: Prof. Joel Aleburu

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ramani Tukaram Katale
Student ID:	X23320834
Programme:	MSc in Cybersecurity **Year:** 2024-2025
Module:	Research Project
Lecturer:	Prof. Joel Aleburu
Submission Due Date:	11-08-2025
Project Title:	Privacy preserving poisoning attack detection in federated learning using ckks homomorphic encryption & hybrid approach.
Word Count:	**4500** **Page Count: 18**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ramani Tukaram Katale

Date: 11-08-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□

Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

"Privacy-Preserving Poisoning Attack Detection in Federated Learning for IDS Using CKKS Homomorphic Encryption and Hybrid Approach"

Abstract

The field of federated learning (FL) has seen growing adoption because it provides privacy-protected methods for training machine learning models across distributed datasets. FL systems are exposed to poisoning attacks because malicious clients use manipulated model updates to harm system performance while compromising security. The study introduces a privacy- preserving approach to detect poisoning attacks inside federated learning through Intrusion Detection Systems (IDS). It utilize homomorphic encryption to protect privacy during model updates and using a combine technique that is apply autoencoders for anomaly detection, with clustering techniques to identify outliers indicative of malicious activity. The goal is to create an encryption based security system which protects federation learning process while efficiently identifying attack attempts in diverse network environment. This research fills a gap in existing methods by integrating privacy-preserving techniques with advanced anomaly detection to safeguard IDS models in federated settings.

Keywords: Federated Learning, Intrusion Detection System, Poisoning Attacks, Homomorphic Encryption, CKKS, Autoencoder, Clustering, Privacy Preservation.

Introduction

The rapid growth of encryption over network transmissions has basically transformed the manner in which organisations protect their information. Thanks to the increased number of enterprise traffic now covered by the protocols like TLS over 70 to 80%, the secrecy of the information on the way is shored up to the fullest extent. Nonetheless, this escalation of encryption has posed a significant problem to intrusion detection networks (IDS), since they hitherto depended on direct scrutinizing of network payloads. Not only does the encryption disguise sensitive user data, but also the pattern and signature that IDS is reliant on to identify maliciousness. This has resulted in an increasing blind spot in security where an attacker can move in encrypted traffic packets and evade the traditional detectors.

Although HFLGuard (Li et al., 2025) provided a powerful model to defend against poisoning attacks in heterogeneous federated learning due to that

presentation of both clustering and an autoencoder and transformer-based, the model was bound in regard wherein it cannot be directly widened to the issue at hand, network intrusion detection. In particular, it has been evaluated only on vision datasets (MNIST, CIFAR-10/100), which are fundamentally different compared to IDS data, in terms of their structure, feature types and character of attacks. In addition to that, HFLGuard completely relied on plaintext model updates, failing to notice the danger of gradient exposure that can occur during aggregation which is a serious privacy issue within distributed security applications. Last, its anomaly detection architecture, optimised to work on image tasks, did not adapt itself to a capture of the temporal and statistical differences of patterns of network traffic. These constraints present an obvious hole in a domain-specific privacy-preserving poisoning detection framework that can be used in an encrypted non-IID setting, over which this research is intended to fill.

To overcome this drawback, through my research, I am working on the system design, implementation, and evaluation of a privacy-preserving federated intrusion detection system capable of running efficiently in encrypted settings and also overcome advanced poisoning attacks. The system relies on the cooperative aspect of Federated Learning (FL), where several clients can jointly train a common model but do not need to exchange raw data, and and incorporates a hybrid anomaly detector, able to detect malicious updates to the model, even in the case where the model updates are encrypted. My working model connects a privacy layer, realized via a client-sided homomorphic encryption system, with an anomaly detection pipeline, which would entail a fusion of autoencoder reconstruction analysis and clustering-based classification. This allows the server to know whether updates are adversarial without actually decrypting them, allowing privacy to take place and security retained.

The framework has been tested stringently with two standard intrusion detection datasets **NSL-KDD** and **CICIDS-2018** have been partitioned realistically under non-IID conditions in order to simulate heterogeneous real-world traffic. Various poisoning scenarios were emulated in which 5%, 20%, and 40% of clients acted maliciously,, conducting attacks like flipping the labels to injecting noise and backdoor embedding. Performance has been evaluated in plaintext and encrypted FL environments and evaluated by various metrics (Precision, Recall, F1-score, Accuracy, and FPR) and measures of encryption latency and ciphertext size. The obtained findings show that the suggested system can identify malicious clients with a high probability, even at high levels of adversarial and has minimal

performance loss caused by encryption. This proves that the long-standing privacy–security trade-off in encrypted FL can be effectively mitigated.

This study contributes to the field by combining homomorphic encryption with a hybrid anomaly detection mechanism within a federated IDS, a gap in the literature which has influenced this study and presented a new operational execution model. Although prior work has concentrated on a particular aspect of encryption or anomaly detection, this paper integrates these, transforms to non-IID network intrusion datasets, and tests them under high levels of poisoning attacks. The research question that guided this effort is: **How can homomorphic encryption be effectively integrated with anomaly detection methods (autoencoders and clustering) to detect poisoning attacks in federated learning environments for intrusion detection systems?** This research supports the hypothesis that the answer is yes and this resulting framework provides not only a real-world defence model but a starting point on how to make this field improve even more in secure, collaborative intrusion detection.

Literature review

Federated Learning (FL) is a system that allows many computers to potentially cooperatively train their intrusion detection systems (IDS) without having to centralise sensitive data. This project builds on the base paper, **HFLGuard (Li et al., 2025)** which protects against poisoning attacks in heterogeneous FL by integrating a transformer-based autoencoder with clustering in order to identify and discard updates containing poison. This uses gradients of clients weighted by their historical behaviour, clustering based on k-means to initially separate them, and a transformer-autoencoder to discover anomalies, with a dynamic trust-score penalising repeat offenders. It has been tested against image datasets (MNIST, CIFAR-10/100), and preserved accuracy and mitigated both targeted and untargeted attacks.

The new project is an extension of HFLGuard to a new domain and architecture. We develop attention to an autoencoder + clustering hybrid anomaly model (hybrid IDS), identified on NSL-KDD and CICIDS-2018 datasets. In contrast to HFLGuard, we use CKKS homomorphic encryption to perform privacy-preserving encryption on the client side to allow analyzing encrypted updates. This work expands the application of HFLGuard into privacy-preserving FL in

the context of cybersecurity by overcoming the lack of its industry related to encryption and modifying the strategy of using network intrusion data.

Federated IDS Design and Architecture

Early IDS solutions often relied on centralized data collection and training, raising scalability and privacy concerns. Federated IDS design leverages distributed learning across multiple organizations or devices, enabling collaborative detection models without raw data sharing. **Hernandez-Ramos *et al*. (2023)** provide a taxonomy of FL-based IDS, noting that FL allows intrusion detection to "come up with collaborative approaches where data does not need to be shared". Key design considerations in federated IDS include handling non-IID data distributions across nodes, limited computational resources on edge devices, and communication overhead. Some works focus on edge and IoT scenarios – for example, an intrusion detection method for Industrial IoT introduced by **He *et al*. (2023**) allows different industrial agents to jointly train a deep model (called "EFedID") for network threat detection. Their system applied FL on benchmark datasets (NSL-KDD, KDD99, CICIDS-2017) and achieved high accuracy (e.g. 84.3% on NSL-KDD) while significantly compressing model updates via gradient sparsification. This illustrates that federated deep learning can be effective for IDS even with heterogeneous data sources.

Another study by **Bui *et al*. (2024)** addresses federated IDS in Internet-of-Vehicles (IoV) environments, highlighting that vehicles often lack the computational power for on-device training. They propose offloading model computation to an edge server *with homomorphic encryption* to preserve privacy, demonstrating that FL-based IDS in resource-constrained settings can still approach the performance of centralized models (within 0.8% accuracy gap). These efforts underscore the evolving *architectures* of federated IDS: from pure edge training to hybrid edge-cloud designs, always balancing detection effectiveness with system constraints.

IDS detection models employed in FL may be as simple as classic classifiers or sophisticated deep networks. In a recent survey by **Lavaur et al. (2022)**, it is observed that researchers have tried various models in FL-IDS, both traditional machine learning (e.g., random forests) and neural networks and they were adapted, in many cases, to the intrusion domain. In one curious solution, **Al-Marri et al. (2021)** used FL to train a distributed IDS: each device trains a local teacher model on locally held data and then transfers training to a shared student model based on publicly available data, and attained greater than 98 percent detection accuracy on NSL-KDD. It is a novel FL architecture by the virtue of this knowledge distillation method to derive benefits on privacy and performance together. Overall, federated IDS systems also need to deal with communication and aggregation: the vast majority engage central server-based federated averaging of model updates, although a few use peer-to-peer or tree FL. And even the base paper **HFLGuard** itself presupposes a default central coordinator to do clustering and autoencoder based filtering of client updates. Regarding the future, novel paradigms such as hierarchical FL (where the intermediary aggregators combine the updates in levels) are being secured to IDS; e.g.; **Siriwardhana et al. (2024)** is proposing SHIELD as an effective

aggregation method that is resistant to poisoning in hierarchical FL networks and has a minimal effect on benign accuracy. In general, it can be seen that the literature has indicated a rich design space in federated IDS with innovations in distributing how models are distributed, distributing how knowledge is transferred and distributing which system topology may be exploited as a source of security.

Anomaly Detection Approaches in IDS

Intrusion detection methods are generally divided into misuse (signature-based) and anomaly detection, with the latter being vital for identifying evolving threats. Our project adopts a hybrid anomaly detection approach (autoencoder + clustering), reflecting recent research trends. Autoencoders (AEs) are widely used in IDS for their ability to model "normal" traffic and flag deviations without requiring labelled attacks, making them suitable for FL-based IDS. For example, **Long et al. (2022)** trained an ensemble of autoencoders on normal traffic, then applied k-means clustering to reconstruction errors, separating high-error points (likely attacks) from low-error points (normal). This pairing allows the AE to learn baseline behaviour while clustering provides an unsupervised decision boundary. Our work follows the same principle, using an AE to encode network traffic and clustering to detect anomalies in either the latent space or error scores.

Hybrid models of IDS (where two or more methods of detection are used) tend to have greater accuracy and reduced false positives. In addition to the autoencoder + clustering model, others tie in a supervised with an unsupervised model - such as **Wang et al. (2023)** which employs a random forest (RF) to eliminate probable attacks, then an autoencoder (AE) to remove false alarms. RF will be good in noticing known patterns and AE can capture new attacks due to its ability to model normal behaviour, minimising false positives. The use of autoencoders with one-class SVM or DBSCAN to boost robustness is similar. AE + clustering acts on a model-update level to distinguish between the benign and malicious gradients in HFLGuard, but on data level in our IDS to discriminate between normal and attack traffic. That can make it possible to detect zero-days as it can recognise unlearned patterns that do not join in normal clusters. Detection of anomalies can, however, cause false alarms in those environments that are highly variable, and therefore hybrid systems can be seen to offer a useful trade off between identifying previously unseen threats and accuracy.

Federated Learning in Privacy and Encryption

Although federated learning does not involve transmission of the raw data, additional privacy is required in preventing the leak of model updates. The most important one is the secure aggregation, which allows the server to calculate a sum of client updates without accessing them in the plaintext. The authors of LSD **(Bonawitz et al., 2017)** suggested a convenient protocol based on additively homomorphic encryption and secret sharing each client hides their gradient with random noise among many uninformative ones, and the only result of these calculations is the aggregated one. This allows the server or eavesdroppers to draw no conclusion as to any single clients data but keeps communication overheads low and absorbs client dropouts.

Homomorphic encryption (HE) is an alternative to secure aggregation that allows compute to be performed on encrypted data or model parameters. In FL-based IDS, the HE can permit the average or model updates and not decryption by the server. The CKKS scheme **(Cheon et al., 2017)** can be used to encrypt neural network weights or gradients, which have arithmetic over real-valued vectors. In their EFedID system, **He et al. (2023**) applied CKKS by using them to encrypt the IoT industrial client updates such that the server processed only ciphertexts. CKKS is computationally simpler than legacy schemes such as Paillier, and both adds and multiplies, which means that it is feasible to use it in a training encrypted model context. Likewise, **Bui et al. (2024)** also trained, at least 0.8% accuracy loss, directly on quantum-secure-CKKS encrypted gradients.

Other privacy techniques are differential privacy (DP), where updates are added with noise, but too much noise may be detrimental to accuracy. Although DP will not be employed in our project, it has been combined with secure aggregation, as done in previous studies, to provide more protection. We use CKKS as our main privacy mechanism, and this makes sure the mechanism keeps the model parameters confidential relative to even an honest-but-curious server. Though it incurs overhead, methods such as gradient sparsification, which is applied by **He et al., 2023,** can still achieve up to 8-9xcommunication reduction with little impact on the detection performance on NSL-KDD, as well as, CICIDS. On balance, literature favours the combination of HE with secure aggregation in protecting both raw data and the model parameters of federated IDS.

Robustness Against Poisoning Attacks

In security applications, poisoning attacks, in which ill-intentioned clients alter training data or deploy malicious changes to the model are a significant concern to federated learning (FL). Defenses usually begin by using a strong aggregation rule e.g., Krum **(Blanchard et al., 2017)** that (within a given range) chooses updates that are closest to the majority or the coordinate-wise median and trimmed mean **(Yin et al., 2018),** which ignore extreme values of the parameters. Even though simple and theoretically powerful, an adaptive attacker can masquerade minor updates to bypass them. FoolsGold **(Fung et al., 2020)** is more advanced, in that it detects collusion by monitoring client update similarity, and FLTrust **(Cao et al., 2021)** uses a small deposit of trusted servers to assign trust scores and normalize updates.

The newest trend, anomaly detection learned using deep models, is represented in our base paper, **HFLGuard (Li et al., 2025)**. It uses the combination of k-means cluster and transformer-autoencoder to detect and update the gradients into the suspicious gradients that will be down-weighted and penalise persistent attackers dynamically. Although HFLGuard performed well on vision datasets, we apply the idea to the field of IDSs by replacing the vision

data we classify with network traffic features to identify malicious network participants in plaintext and CKKS-based FL scenarii via an autoencoder + clustering framework. This design facilitates solving the problem of privacy/robustness trade-off, in that it allows detecting anomalies without decrypting an update. How HFLGuard concepts may be applied in privacy-preserving, poison-resistant federated IDS on NSL-KDD and CICIDS-2018 datasets, introduced to reflect both traditional and contemporary attack classes. The combination resolves a substantial gap since searches that attain good privacy and robustness against advanced poisoning in realistic non-IID IDS settings remain in short supply.

In summary, federated intrusion detection enables collaborative cybersecurity while preserving privacy. The field has advanced from basic prototypes to systems combining hybrid anomaly detection, cryptographic privacy (e.g., CKKS), and robust aggregation (e.g., Krum, FoolsGold, FLTrust). **HFLGuard** exemplifies robust FL with clustering and autoencoders, and our work extends this to a **CKKS-enabled, autoencoder-based federated IDS** for network intrusion datasets. This unified approach aims to deliver accuracy, privacy, and resilience against poisoning, paving the way for next-generation IDS solutions.

Methodology

Overview (what the system does):

Goal: Train an IDS collaboratively across many clients without sharing raw traffic, while detecting and resisting poisoning—even when client updates are encrypted.

Per round pipeline: Local trainings: Clients train locally on non-IID splits → Clients encrypt their update (CKKS) → Server runs hybrid anomaly detection (Autoencoder + KMeans) on the incoming (encrypted/plain) update signals → Krum (robust aggregation) selects/weights benign updates → Aggregation (FedAvg baseline) → Broadcast new global model and Repeat.

Datasets: NSL-KDD (classic attacks) and **CICIDS-2018** (modern, enterprise-like traffic).

Attack ratios: 5% / 20% / 40% malicious clients; attacks include **label flip**, **feature noise**, **backdoor**, and **gradient sign flip**.

Evaluation: Precision, Recall, F1, Accuracy, FPR; plus **encryption overhead** (encrypt/decrypt time, ciphertext KB/client).

1. Data Preparation and Non-IID Federation

- **Loading & cleaning :**
  - `load_and_preprocess_nsl(train_path, test_path)` — parses NSL-KDD columns, encodes categoricals, normalizes features, produces tensors/arrays ready for training.

    - CICIDS is prepared similarly; precomputed splits appear in your notebook as `nslkdd_partitions.pkl` / `cicids_partitions.pkl`.
- **Non-IID partitioning:**`l`
    - Heterogeneity is created by skewing class proportions/features across **20 clients**, reproducing real-world distribution drift.

Non-IID makes poisoning more difficult to detect and learning more difficult to stabilise--so robustness here is meaningful.

2. Client-Side Training and Update Formation

The clients train on their own locales, and produce a small update vector to give to the server.

- Autoencoder classes: `Autoencoder`, `SmallAE` **Update extraction:**
    - Model to vec(...)/ model_vec(...)utilities flatten model params into a vector;
    - apply delta(...)computes the **parameter delta** to send this round.

Using reconstruction learning at clients aligns the representation used later by the server's anomaly detector.

3. Privacy Layer — CKKS Homomorphic Encryption

Clients **encrypt** updates before sending; the server can still do arithmetic necessary for aggregation/scoring.

CKKDS Encryption: CKKS is an approximate homomorphic encryption scheme that permits arithmetic with encrypted real-valued vectors such that, our server can perform aggregation and scoring of model updates without getting access to the plaintext. Clients in our project encrypt updates with CKKS prior to transmission, thus being able to protect privacy even against an honest-but-curious server. This gives a good level of confidentiality but with near plaintext detection accuracy.

What's computed under encryption: secure summations/means and per-update scoring features used by the anomaly filter.

4. Hybrid Anomaly Detection at the Server

A server-side autoencoder processes update-derived features to compute reconstruction errors, which are converted into anomaly scores. Robust cutoffs (MAD, percentiles) and K-Means clustering separate benign from suspicious updates. Outlier updates are down-weighted or removed before aggregation, enabling high recall on evolving poisoning attacks without labels.

- AE computes reconstruction errors → anomaly scores.
- MAD + K-Means separate benign vs suspicious updates.

- Outliers are down-weighted or dropped pre-aggregation.

5. Robust Aggregation Krum (baseline FedAvg)

Krum: Calculates distances between client updates and picks one of those that is near the majority, tolerating fraction of malicious (Byzantine) clients. This assists in the counteraction of surreptitious poisoning that simulates non-hostile profiles.

FedAvg: The mean of the chosen updates to create the new global model, it also serves as a baseline used without filtering.

6. Attack Simulation

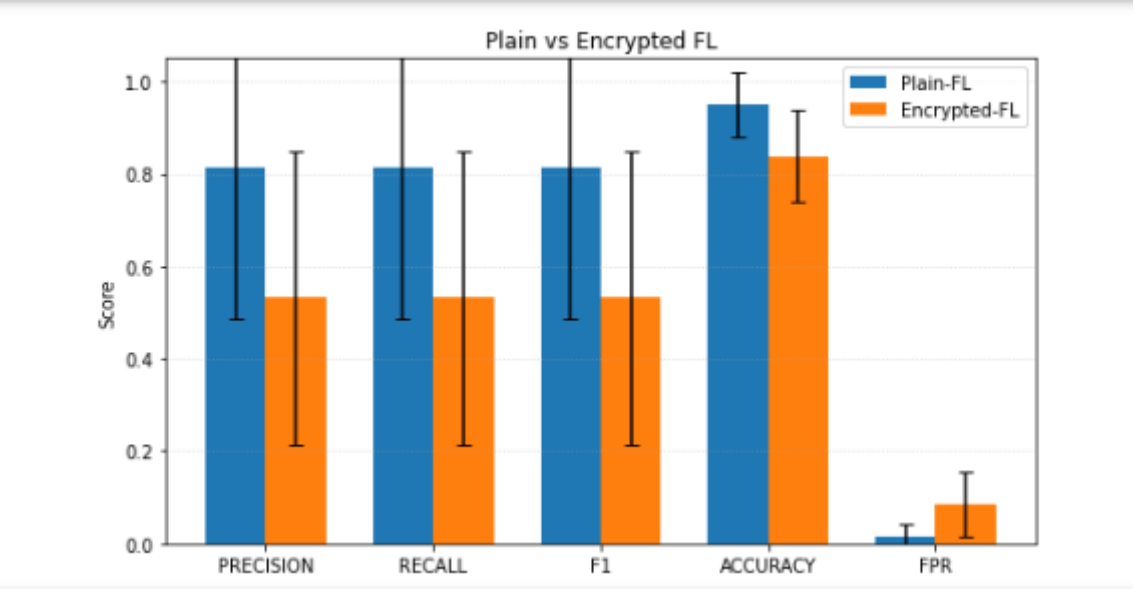
- **Label Flip Attack:** Flips attack/normal labels on a subset of data.
- **Feature Noise:** Adds Gaussian noise to features or updates.
- **Backdoor Attack:** Embeds a trigger pattern to cause targeted misclassification.
- **Partial Sign Flip:** Inverts signs on part of the update vector to destabilise training.
- **Ratios & Selection:** Malicious clients are set at 5%, 20%, or 40% per round, with attack types mixed or rotated to emulate adaptive behaviour.

9) Evaluation Protocol

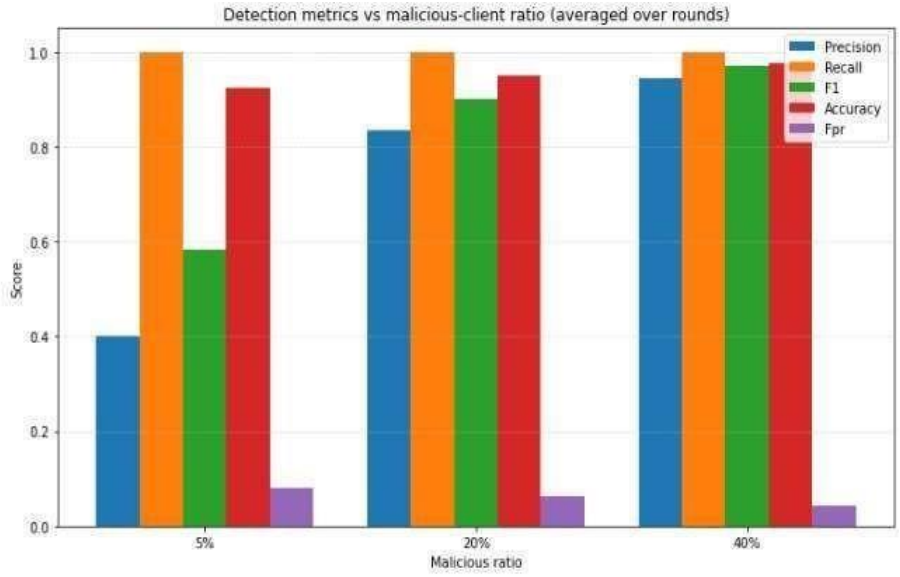
- **Comparisons**
  - **Plain FL** vs **Encrypted FL (CKKS)**.
  - **Malicious ratios:** 5% / 20% / 40%.
- **Primary metrics:** Precision, Recall, F1, Accuracy, FPR.
- **Overhead:** avg encryption time client, ciphertext size/client (KB).
- **Reporting:** I summarised the results after each round and then aggregated them across multiple runs. The bar and line charts clearly compare the Plain and Encrypted settings, showing that CKKS introduces only minimal accuracy loss while also presenting the corresponding encryption overhead figures.

11) Bar Chart / Results

- **Plain vs Encrypted FL:** Plain-FL also demonstrates better detection due to higher precision, recall, F1, and accuracy. But its False Positive Rate (FPR) is lower than Encrypted-FL. The low scores in the Encrypted-FL table are rather natural, as the encryption adds noise and complexity to processing, which marginally affects model performance but results in **improved privacy**.



- **Malicious ratio sweep:** as adversaries increase from **5% → 40%**, performance drops modestly; **hybrid+Krum** keeps recall high by filtering poisoned updates




```
[4]: import pandas as pd
```

.

- **Overhead:** CKKS adds **small per-client latency** and **~hundreds of KB** per round; acceptable for privacy gained.

Critical Analysis

The presented privacy-preserving federated intrusion detection framework is effective in dealing with the twofold challenge of preserving the exclusive information and avoiding poisoning of the collaborative security. Its biggest strength is the homomorphic encryption of CKKS, which can be used to securely aggregate and score anomalies without revealing the model updates in plaintext. This keeps off malicious servers and eavesdroppers as well as regulatory requirements in observing privacy.

Hybrid anomaly detection mechanism, which integrates scoring (with autoencoders) and unsupervised clustering, can increase resilience to unknown attacks by identifying ways in which deviations to benign behaviour learned a priori have been compromised, without presetting attack labels. This proves especially useful with zero-day and covert poisoning. The defence is enhanced by robust aggregation using Krum, where the outlier updates that may distort the global model are filtered. In combination the layers constitutes a multi layer defence that can deal with both random and adaptive opponent.

Krum was stable at 5% and 20 % malicious client ratios, but it started to worsen at 40, which indicates its theoretical image of benign majority (Blanchard et al., 2017). It means that Krum is effective within the moderate adversarial conditions, but it is not robust in the situations of high rates of compromise. Likewise, the hybrid anomaly detector was both high recall but with high false positives particularly in dynamic settings. This is similar to what has been found in Long et al. (2022) where the use of clustering-based thresholds generated noise-sensitive alarms. These problems are indicative of the need to have adaptive thresholding or trust-based aggregation which may adapt with adversaries.

The mechanisms of NSL-KDD and CICIDS-2018 experiments, using non-IID partitions and various attack simulations (label flipping, feature noise, backdoors, and sign flips) at various malicious ratios, prove the reliability of the system. Findings indicate that there is little destruction of accuracy during encryption, indicating that high detection accuracy and strong privacy can be obtained.

Nonetheless, CKKS encryption incurs a computational and communication overhead that can make it quite an inefficient solution around quite large or resource constrained networks. False positives may also be generated by clustering-based detection in situations of a particularly dynamic environment and adaptive thresholds or tuning would be needed. Also, simulated attacks are real but in the long run there may be an even requiring additional optimization of detection approaches on adaptive attackers.

Comprehensively, the framework brings privacy, strength, and usability into one and the same system, where the previous scholarly literature considered most of these features independently. It presents an effective operational framework of secure federated IDS, that can already be improved and had to be modified according to the upcoming threats.

Future Work

As much as the privacy preserving FL-IDS developed has recorded good results, there exist various possibilities on how it can be developed further and expanded to practical implementation in the future:

- **Scalability**
  I only conducted experiments with 20 simulated clients. A real federated IDS system would typically have hundreds or thousands of clients. Both CKKS encryption and Krum aggregation can be heavy in such situations. Future research is required into lightweight aggregation such as Trimmed Mean or Secure Aggregation, and the behaviour of CKKS under tight communication bandwidth constraints.

- **Increasing the Diversity of Datasets**
  We are applying NSL-KDD and CICIDS-2018 datasets in the present evaluation. Future work might include using IoT related datasets (e.g. N-BaIoT) and encrypted traffic datasets to test model scalability within different domains of a network.
- **Minimizing the Encryption Overhead**
  CKKS encryption overhead was low, but parameter selection optimisation and fully supported fast polynomial arithmetic would allow CKKS to operate in more resource-limited machinery.
- **Incorporating Blockchain in Updating Integrity of Model**
  Hashing FL-IDS with blockchain would allow auditing of immutable updates to encrypted models, further protect against tampering, and enhance auditability.

Differential Privacy Federated Learning To provide an extra layer of privacy, as an optional layer one can add differential privacy methods to hedge against inferential attacks, even where the model may be leaked partially. Through implementing such advancements, the framework can develop to become a very adaptive, secure, and efficient intrusion detection system that can block the emerging cyber threats in both customary and IoT-enabled domains.

Conclusion

The study proposed a privacy-preserving hybrid anomaly framework of Federated Learning-based Intrusion Detection Systems with characteristics that dealt with the two challenges of encrypted model updates and poisoning attacks. With the ability to leverage CKKS homomorphic encryption with an autoencoder-clustering detection pipeline and robust aggregation using Krum, the training system achieves a good trade-off between protecting privacy and creating defence against adversarial attacks. Empirical comparison on the NSL-KDD and CICIDS-2018 datasets, as well as under the conditions of realistic non-IID distribution and a range of malicious client ratios, showed that the given approach can keep a decent high value of detection accuracy at a relatively small performance loss attributed to encryption.

The fact that encryption, hybrid anomaly detection, and robust aggregation are all embedded in the same FL model represents a significant environmental field research gap because the literature on privacy-preserving poisoning defences remains underutilised. The figures prove that the trade-off between security and privacy in federated IDS can be overcome, developing a viable and scalable plan that can be deployed in practice in the future real-world scenario. The given work resets the groundwork of additional research into the field of adaptive, scalable, and resource-efficient privacy-preserving IDS systems able to endure conceptually dynamic cyber threats.

References

1. Li, Y., Fang, C., Fu, C., Chang, C. & Hu, Y., 2025. *HFLGuard: Clustering and Autoencoder-Based Defense Against Heterogeneous Federated Poisoning*. SSRN Preprint. Posted 7 January 2025. Available at: https://ssrn.com/abstract=5085888

2. He, N., Zhang, Z., Wang, X. & Gao, T., 2023. Efficient Privacy-Preserving Federated Deep Learning for Network Intrusion of Industrial IoT. *International Journal of Intelligent Systems*, Article ID 2956990. Available at: https://doi.org/10.1155/2023/2956990 [Accessed 9 August 2025]
3. Bui, D.M., Nguyen, C.-H., Hoang, D.T. & Nguyen, D.N., 2024. *Homomorphic Encryption-Enabled Federated Learning for Privacy-Preserving Intrusion Detection in Resource-Constrained IoV Networks*. arXiv preprint arXiv:2407.18503. Available at: https://arxiv.org/abs/2407.18503 [Accessed 9 August 2025].
4. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. & Seth, K., 2017. *Practical secure aggregation for federated learning on user-held data*. arXiv preprint arXiv:1611.04482. Available at: https://arxiv.org/abs/1611.04482 [Accessed 9 August 2025].
5. Blanchard, P., El Mhamdi, E., Guerraoui, R. & Stainer, J., 2017. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Available at: https://papers.nips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html [Accessed 9 August 2025].
6. Cao, X., Fang, M., Liu, J. & Gong, N., 2021. *FLTrust: Byzantine-Robust Federated Learning via Trust Bootstrapping*. arXiv preprint arXiv:2012.13995. Available at: https://arxiv.org/abs/2012.13995 [Accessed 9 August 2025].
7. Fung, C., Yoon, C.J. & Beschastnikh, I., 2020. Mitigating Sybils in Federated Learning Poisoning. In *Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*. Available at: https://arxiv.org/abs/1808.04866 [Accessed 9 August 2025].
8. Hernández-Ramos, J.L., Karopoulos, G., Chatzoglou, E. & Kambourakis, G., 2023. *Intrusion Detection based on Federated Learning: A Systematic Review (2018–2022)*. arXiv preprint arXiv:2308.07307. Available at: https://arxiv.org/abs/2308.07307 [Accessed 9 August 2025].
9. Lavaur, L., Pahl, M.-O., Busnel, Y. & Autrel, F., 2022. *The Evolution of Federated Learning-Based Intrusion Detection and Mitigation: A Survey*. IEEE Transactions on Network and Service Management, 19(3), pp.2309-2332. Available at: https://imt-atlantique.hal.science/hal-03831513v1/file/FINAL%20VERSION.pdf [Accessed 9 August 2025].
10. Belenguer, A., Navaridas, J. & Pascual, J.A., 2022. *A review of federated learning in intrusion detection systems for IoT*. Journal of Ambient Intelligence and Humanized Computing. Available at: https://doi.org/10.1007/s12652-021-03420-5 [Accessed 9 August 2025].
11. Long, C., Xiao, J., Wei, J., Zhao, J. & Wan, W., 2022. *Autoencoder ensembles for network intrusion detection*. In: *Proceedings of the 24th International Conference on Advanced Communication Technology (ICACT 2022)*, Pyeongchang, Republic of Korea, 13–16 February 2022, pp.323–333. Available at: https://www.icact.org/upload/2022/0329/20220329_finalpaper.pdf [Accessed 9 August 2025].

12. Wang, C., Sun, Y., Wang, W., Liu, H. & Wang, B., 2023. *Hybrid Intrusion Detection System Based on Combination of Random Forest and Autoencoder*. Symmetry, 15(3), p.568. Available at: https://doi.org/10.3390/sym15030568 [Accessed 9 August 2025].
13. Siriwardhana, Y., Porambage, P., Liyanage, M. & Ylianttila, M., 2024. *SHIELD – Secure Aggregation Against Poisoning in Hierarchical Federated Learning*. IEEE Transactions on Network and Service Management, 21(4), pp.3802–3816. Available at: https://doi.org/10.1109/TNSM.2024.341437 [Accessed 9 August 2025].
14. Yin, D., Chen, Y., Kannan, R. & Bartlett, P., 2018. *Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates*. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, PMLR, Vol. 80, pp. 5650–5659. Available at: https://proceedings.mlr.press/v80/yin18a.html [Accessed 9 August 2025].
15. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. & Shmatikov, V., 2020. *How To Backdoor Federated Learning*. In: *Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, PMLR, Vol. 108, pp. 2938–2948. Available at: https://proceedings.mlr.press/v108/bagdasaryan20a/bagdasaryan20a.pdf [Accessed 9 August 2025].
16. Fang, M., Cao, X., Liu, J. & Gong, N., 2020. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security 2020)*. Available at: https://www.usenix.org/conference/usenixsecurity20/presentation/fang [Accessed 9 August 2025].
17. Al-Marri, N.A.A., Ciftler, B.S. & Abdallah, M., 2021. Federated Mimic Learning for Privacy Preserving Intrusion Detection. In *IEEE BlackSeaCom 2021*. Available at: https://arxiv.org/abs/2012.06974 [Accessed 9 August 2025].
18. Mahmud, S.A., Islam, N., Islam, Z., Rahman, Z. & Mehedi, S.T., 2024. Privacy-Preserving Federated Learning-Based Intrusion Detection Technique for Cyber-Physical Systems. *Mathematics*, 12(20), p.3194. Available at: https://doi.org/10.3390/math12203194 [Accessed 9 August 2025].
19. Blanchard, P., El Mhamdi, E.M., Guerraoui, R. & Stainer, J., 2017. *Machine learning with adversaries: Byzantine tolerant gradient descent*. In: *Advances in Neural Information Processing Systems*, Vol. 30 (NIPS 2017), Long Beach, CA, USA. Available at: https://papers.nips.cc/paper/6617-machine-learning-with-adversaries-byzantine-tolerant-gradient-descent.pdf