

Federated Continual Learning for Intrusion Detection: A Drift-Aware Benchmark on the Windows-APT 2025 Dataset

Ilham Achmad Fiqrin¹, Taufiq Hidayat^{2*}

¹School of Computing and Informatics, Universiti Teknologi Brunei, Jalan Tungku Link, Gadong, Bandar Seri Begawan, Brunei Darussalam

²Department of Informatics, Faculty of Industrial Technology, Universitas Islam Indonesia Yogyakarta, Indonesia

Corresponding Author's e-mail: taufiq.hidayat@uii.ac.id

ARMADA
JURNAL PENELITIAN MULTIDISIPLIN

e-ISSN: 2964-2981

ARMADA : Jurnal Penelitian Multidisiplin

<https://ejournal.45mataram.ac.id/index.php/armada>

Vol. 04, No. 09 September, 2026

Page: 5385-5394

DOI:

<https://doi.org/10.55681/armada.v4i9.3928>

Article History:

Received: Juli 27, 2026

Revised: Agustus 11, 2026

Accepted: September 18, 2026

Abstract : Federated Learning (FL) enables collaborative intrusion detection among organizations without sharing raw telemetry data. However, existing FL approaches often assume static learning environments, limiting their effectiveness in detecting Advanced Persistent Threats (APTs) with continuously evolving behaviors. This study proposes a Federated Continual Learning (FCL) benchmark for APT technique detection using the Windows-APT 2025 dataset. Following a validation-first paradigm, the dataset characteristics, including client participation, label reliability, and behavioral drift, are analyzed to establish an appropriate evaluation framework. The benchmark evaluates Replay, EWC, and DER++ continual learning methods combined with FedAvg and FedProx optimization strategies, using MLP and 1D-CNN architectures with drift-triggered adaptation. Experimental results show that Replay and EWC achieve up to 23% higher multi-label F1 performance than DER++ under limited local training conditions. FedAvg consistently outperforms FedProx, while MLP surpasses 1D-CNN with approximately one-third of the communication overhead. These findings demonstrate the effectiveness of the proposed benchmark for evaluating drift-aware federated intrusion detection systems.

Keywords : Federated Learning; Continual Learning; Intrusion Detection; Drift-Aware Learning; Windows-APT 2025 Dataset; Network Security.

Abstrak : *Federated Learning* (FL) memungkinkan deteksi intrusi secara kolaboratif antarorganisasi tanpa perlu berbagi data telemetri mentah. Namun, pendekatan FL yang ada umumnya masih mengasumsikan lingkungan pembelajaran yang statis sehingga kurang efektif dalam mendeteksi *Advanced Persistent Threats* (APT) yang memiliki pola serangan yang terus berkembang. Penelitian ini mengusulkan *Federated Continual Learning* (FCL) benchmark untuk deteksi teknik APT menggunakan dataset Windows-APT 2025. Dengan pendekatan *validation-first paradigm*, karakteristik dataset seperti partisipasi klien, keandalan label, dan perubahan perilaku (*behavioral drift*) dianalisis untuk membangun kerangka evaluasi yang sesuai. Benchmark ini mengevaluasi metode pembelajaran kontinu Replay, EWC, dan DER++ yang dikombinasikan dengan strategi optimasi federasi FedAvg dan FedProx, menggunakan arsitektur jaringan MLP dan

1D-CNN dengan teknik adaptasi berbasis deteksi perubahan (*drift-triggered adaptation*). Hasil eksperimen menunjukkan bahwa Replay dan EWC mencapai peningkatan nilai *multi-label F1* hingga 23% dibandingkan DER++ pada kondisi pelatihan lokal terbatas. FedAvg secara konsisten memberikan kinerja lebih baik dibandingkan FedProx, sementara MLP mengungguli 1D-CNN dengan kebutuhan *communication overhead* sekitar sepertiga lebih rendah. Temuan ini menunjukkan efektivitas benchmark yang diusulkan untuk mengevaluasi sistem deteksi intrusi federasi yang adaptif terhadap perubahan pola serangan.

Kata Kunci : *Federated Learning; Continual Learning; Deteksi Intrusi; Pembelajaran Berbasis Drift; Dataset Windows-APT 2025; Keamanan Jaringan.*

INTRODUCTION

Advanced Persistent Threats (APTs) can be classified as some of the hardest forms of cyber attacks to defend against in today's world of information security: prolonged and multistage attacks that take place over weeks or even months and deliberately circumvent signature-based detections while adjusting their TTPs (Mat et al., 2024; Susanto, Figrin, Patchmuthu, Tajuddin, & Zai, 2025; Susanto, Figrin, Patchmuthu, Tajuddin, Waruwu, et al., 2025). The identification of such attack campaigns from host- and network-level data became more and more dependent on machine learning due to the inability of static rules-based systems to match the variety of attack behaviors described by frameworks such as MITRE ATT&CK (Apruzzese et al., 2023).

On the other hand, federated learning (FL) appears to be a viable strategy for use in the security domain. Rather than having organizations pool their sensitive log information into a single place and face potential risks associated with privacy, legal, and ownership issues, federated learning gives organizations the opportunity to train an intrusion detection system collaboratively without moving any of the original data out of their premises. Federated learning is appropriate for the purpose of intrusion detection as organizations can learn from the experience of others. (Agrawal et al., 2022; Lavaur et al., 2022). Nevertheless, existing FL approaches, among which the two popular algorithms considered in this paper, namely, FedAvg and its variation FedProx, which takes into account heterogeneity of data, imply that the learning process happens in quite stable conditions. Client's training sets might be non-IID (different clients' data sets), but the task itself remains the same throughout the entire process. It has been discovered that this condition does not hold in some cases even outside the scope of security applications. Moreover, it has also been noticed that such an approach, as FedProx, may impair learning from large local sets of clients' data. (Criado et al., 2022; Zhu et al., 2021)

Such an assumption fails in cybersecurity where the behavior of attackers keeps changing because they keep developing and adapting to new methods which results in concept drift (Agrahari & Singh, 2022; Han et al., 2022; Korycki & Krawczyk, 2023). This means that intrusion detection models that have been trained using static distribution of data become increasingly ineffective (Andresini et al., 2021). Continual Learning (CL) offers a solution in such a situation where a model is trained on new data without suffering from catastrophic forgetting of previous information (De Lange et al., 2022; L. Wang et al., 2024). There are many types of CL techniques that have been suggested to address this, from regularization-based techniques that impose penalties on the movement of parameters considered significant in the earlier tasks (Kirkpatrick et al., 2017) to replay-based techniques that replay an example buffer, to hybrid techniques that integrate the techniques of rehearsal and knowledge distillation of its own predictions (Buzzega et al., 2020).

Federated Continual Learning (FCL), which integrates federated learning and continual learning into one framework, is relatively new and under-explored research (Sun et al., 2025; Z. I. Wang et al., 2024; Zhong et al., 2025). Current research in FCL utilizes artificial benchmarks for

continual learning that include computer vision datasets with predetermined task boundaries and artificially generated client heterogeneity (De Lange et al., 2022; Hamed et al., 2025; Z. I. Wang et al., 2024). These benchmarks are not indicative of cybersecurity, whose dynamics imply the changing pattern of clients' behavior and labeling.

This study addresses the research gap by utilising the recently developed Windows-APT 2025 benchmark (Mozaffari et al., 2026), which contains approximately 102,000 log entries from 36 APT attack scenarios generated using Wazuh/Sysmon. The dataset is first validated for federated continual learning through analysis of client participation, label quality, and behavioural drift. The main contributions of this study are: (1) developing a realistic federated continual learning benchmark that considers sparse, intermittent, and delayed client arrivals; (2) defining behavioural drift based on changes in attack technique distributions to characterise evolving attacker behaviours; (3) proposing a drift-aware continual learning adaptation framework evaluated with FedAvg and FedProx aggregation methods; and (4) providing a comprehensive evaluation framework for binary attack detection and multi-label technique detection using federated and continual learning metrics.

RESEARCH METHOD

This study adopts an experimental research design to develop and evaluate a Federated Continual Learning (FCL) framework for Advanced Persistent Threat (APT) technique detection. The proposed methodology follows a validation-first paradigm, where the characteristics of the dataset are analysed before model development to establish a realistic evaluation environment. The study utilises the publicly available Windows-APT 2025 dataset (Mozaffari et al., 2026), which contains approximately 102,000 Wazuh/Sysmon log records collected from five Windows machines between October and December 2024. The dataset consists of 36 APT-inspired attack scenarios labelled according to MITRE ATT&CK tactics and techniques, together with a large number of background logs. The dataset is organised into six temporal windows representing federated communication rounds, with 33 MITRE ATT&CK techniques retained after frequency-based filtering. Data preparation is performed using only event telemetry features, including process, network, registry, logon, and file integrity monitoring information, while MITRE ATT&CK labels are excluded from model inputs to prevent information leakage. Categorical features are transformed using one-hot encoding, ordinal variables are standardised using z-score normalisation, and additional binary and count-based features are extracted from textual attributes.

The federated learning environment consists of five persistent clients, where each client represents a Windows host agent that maintains its local model state, replay buffer, and drift history throughout the training process. Client participation is determined based on local event availability, enabling evaluation under realistic federated conditions involving full participation, intermittent participation, and delayed client availability. Two federated optimisation strategies are evaluated, namely Federated Averaging (FedAvg) and FedProx. FedAvg aggregates client models using sample-weighted averaging, whereas FedProx introduces a proximal regularisation term to improve robustness under heterogeneous client conditions. To address the evolving nature of APT behaviours, behavioural drift is defined based on changes in attack technique distributions between federated rounds. Drift is measured using Jaccard similarity between technique sets from consecutive active rounds, allowing the framework to identify significant changes in attacker behaviour without being affected by intermittent client participation. Three continual learning approaches are investigated: Experience Replay (Replay), Elastic Weight Consolidation (EWC), and Dark Experience Replay++ (DER++). A drift-triggered adaptation mechanism is incorporated into all continual learning strategies, where detected behavioural changes modify replay contribution or parameter regularisation before local training.

The experimental evaluation employs two neural network architectures, namely Multilayer Perceptron (MLP) and 1D-Convolutional Neural Network (1D-CNN), using a shared multi-task framework with separate outputs for binary malicious/benign detection and multi-label MITRE ATT&CK technique classification. The MLP architecture consists of two fully connected layers with

128 and 64 units, while the 1D-CNN processes feature vectors through convolutional blocks with pooling and dropout layers. The dataset is divided using an 80:20 stratified train-test split for each client-round combination, with held-out test data preserved throughout training to evaluate continual adaptation performance. Model performance is assessed from three perspectives: classification effectiveness, continual learning capability, and federated learning efficiency. Binary detection is evaluated using F1-score, while multi-label technique detection uses macro- and micro-averaged F1 scores. Continual learning performance is measured using average accuracy, forgetting, backward transfer (BWT), and forward transfer (FWT), whereas federated learning efficiency is evaluated through communication cost and client heterogeneity based on local training loss variance.

RESULT AND DISCUSSION

This section examines the feasibility of the federated continual learning framework via four experiments. First, the comparison is carried out among continual learning methods, federated aggregation methods, then neural networks structures, and finally sensitivity analysis regarding the client participation threshold. Each experiment is run independently on the 'full_tagged' and 'high_confidence' label-reliability subsets.

Continual Learning Method Comparison

Table 1. Continual learning method comparison (FedAvg, MLP)

Metho d	Subset	Multilabe l F1	Multilabe l Forgettin g	Multilabe l BWT	Multilabe l FWT	Binar y F1	Binary Forgettin g
Replay	Full_tagged	0.755	0.0001	0.135	0.549	0.994	0.0003
Replay	High_confidenc e	0.703	0.0000	0.152	0.512	0.929	0.0008
EWC	Full_tagged	0.736	0.0000	0.127	0.542	0.994	0.0001
EWC	High_confidenc e	0.661	0.0000	0.162	0.465	0.926	0.0009
DER++	Full_tagged	0.523	0.0000	0.094	0.376	0.994	0.0002
DER++	High_confidenc e	0.513	0.0137	0.125	0.394	0.900	0.00651
No CL	Full_tagged	0.75	0.0000	0.14	0.537	0.994	0.000014
No CL	High_confidenc e	0.654	0.013	0.128	0.503	0.929	0.0008

The performance of Replay, EWC, DER++, and the No CL, which performs standard fine-tuning without any CL mechanism, baseline method under the optimal federated setup using FedAvg with MLP backbone is provided in Table 6. Multi-label and binary F1 scores for each of these algorithms on the full_tagged and high_confidence subsets are shown in Figure 6. Among all considered techniques, Replay demonstrates the best multi-label performance, scoring 0.755 and 0.703 F1 on the full_tagged and high_confidence subsets, respectively. The No CL baseline algorithm shows good performance, scoring 0.750 and 0.654 F1 on the full_tagged and high_confidence subsets, respectively, which is quite close to that of Replay on the full_tagged subset and comparable from EWC on both subsets but better than DER++. EWC demonstrates slightly worse multi-label F1 score compared to Replay, by approximately 2–5 percentage points lower, while DER++ is 18–24 percentage points lower. On the other hand, the performance of all the techniques for binary classification is very high, with 0.994 F1 for Replay, EWC, and No CL on the full_tagged subset and 0.926–0.929 F1 on the high_confidence subset, while DER++ has 0.900 F1 on the high_confidence subset.

Aggregation Strategy Comparison

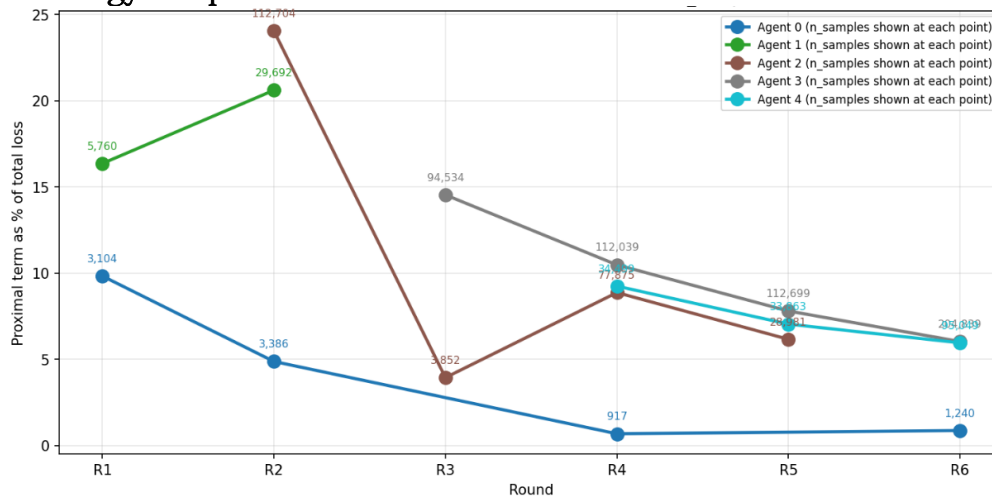


Figure 1. FedProx Proximal Term as a Share of Total Loss per Client and Round

Table 2. Aggregation strategy comparison (Replay, MLP)

Aggregation	Subset	Multilabel F1	Binary F1
FedAvg	Full_tagged	0.7555	0.994
FedAvg	High_confidence	0.703	0.929
FedProx	Full_tagged	0.521	0.992
FedProx	High_confidence	0.576	0.923

Table 2 shows the results for FedAvg and FedProx with $\mu = 0.01$ for the best continual learning technique which is Experience Replay. In Figure 7 we see a plot of the size of the proximal term used during training. FedAvg outperforms FedProx with a gain of 13-23 percentage points in multi-label F1. In order to analyze this gap, the contribution of the proximal term to the client's training loss was examined and illustrates in Figure 6. Up to 16.4% of the loss function of one client in round R1 and up to 24.0% of the loss function of another client in round R2 belong to the proximal term. The proximal contributions were observed to be larger for clients who have more local data points because it is determined by the distance of the parameter from the global model without normalization by client sample size. This implies that, for some client, the penalty term from the FedProx algorithm was high enough to impact the local model training. This may have decreased the adaptability of the model to changing local data.

Model Architecture Comparison

Table 3. Model architecture comparison using Replay and FedAvg

Architecture	Subset	Multilabel F1	Binary F1	Total Parameter Transfers	Multilabel Forgetting
MLP	Full_tagged	0.750	0.994	1121796	0.0010
MLP	High_confidence	0.698	0.930	1012180	0.0000
1D-CNN	Full_tagged	0.747	0.994	3290180	0.0001
1D-CNN	High_confidence	0.655	0.918	3028244	0.0177

Table 3 presents a comparison between the MLP and 1D-CNN networks using the highest performance configuration which is Replay and FedAvg. Both networks have similar performance on full_tagged with difference of 0.3 percent points. MLP achieved better performance to the 1D-CNN in high_confidence with 0.698 and 0.655, respectively. On average, 1D-CNN needs around 3 times as many parameters to transfer per round as the MLP and shows higher forgetting on sparse, with forgetting values of 0.0177 vs. 0.000, respectively. Given that the input data in form of a

vector does not have any intrinsic spatial or temporal structure, the larger capacity of 1D-CNN brings no benefits at all.

Participation Threshold Sensitivity

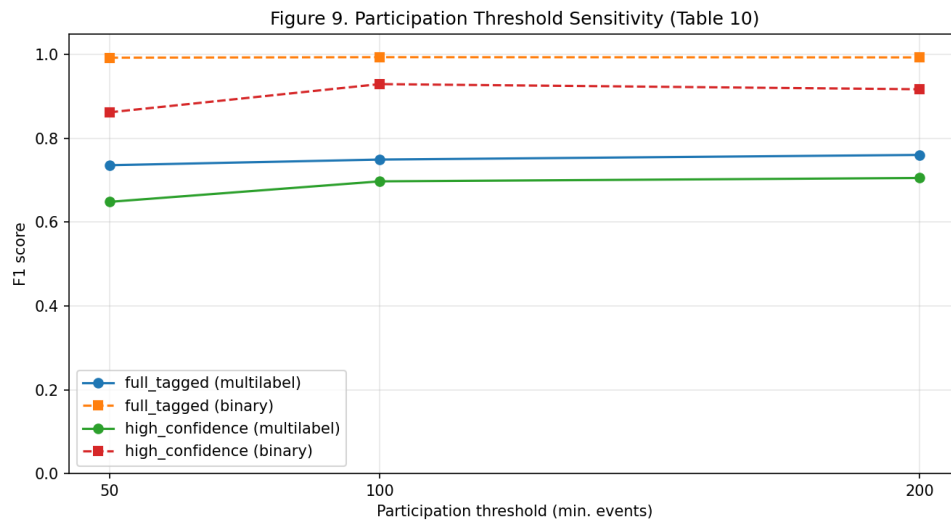


Figure 2. Line plot of F1 vs. participation threshold, by subset and task

Table 4. Reports Results Under The Confirmed Best Configuration (Replay, Fedavg, MLP) Across Three Client-Participation Thresholds.

Threshold	Subset	Multilabel F1	Binary F1	Client-Round Participations
50	Full_tagged	0.736	0.993	20
100	Full_tagged	0.750	0.994	17
200	Full_tagged	0.761	0.993	15
50	High_confidence	0.649	0.862	20
100	High_confidence	0.698	0.930	17
200	High_confidence	0.706	0.917	15

As shown in Table 10, the proposed model demonstrates stable performance across all evaluated participation threshold of 50, 100, and 200 minimum events. For the full_tagged data set, the improvement in multilabel F1 score from 0.736 to 0.761 with 2.5 percentage point, and the binary F1 score is unchanged and equals 0.993-0.994. In case of the high_confidence data set, multilabel F1 score increases from 0.649 to 0.706 with 5.7 percentage points and there is an increase in binary F1 from 0.862 to 0.930 at threshold 100 and decrease to 0.917 at threshold 200. In Figure 7, the increase in multilabel F1 is quite straightforward. Higher participation threshold allows using only the client-rounds which provide us with high-quality labeled data while the total number of clients participating in federated learning drops from 20 to 17 and 15, respectively. All these facts prove that the threshold 100 selected is not accidental or chosen selectively.

Summary

In each of the four experiments, the best model was identified as Replay continual learning with FedAvg aggregation and MLP architecture, tested on both the full tagged dataset and the high-confidence dataset. The multilabel technique attribution F1 scores for these configurations were 0.755 and 0.703, whereas the binary detection F1 scores were 0.994 and 0.929 in full_tagged high_confidence respectively, with zero catastrophic forgetting and positive backward and forward transfer in all federated iterations. In all four of the conducted experiments, the best overall results have been obtained in the configuration of the Replay continual learning method along with the FedAvg aggregation and MLP backbone that consistently yielded the highest multilabel F1-scores on both the full_tagged and high_confidence datasets. These F1-scores for the multilabel setup are 0.755 and 0.703 respectively, while for the binary detection setting they are 0.994 and 0.929

respectively. Replay also demonstrated negligible catastrophic forgetting and consistently positive transfer in both directions. Although the baseline model with naive fine-tuning that achieves binary detection results comparable to those of Replay, Replay provided the best option to improve on the multilabel problem.

Discussion

Interpretation of Key Findings

Simplicity can be better than more regularizing complexity when the budget for training is restricted. Replay and EWC performed similarly and were superior to DER++ by 15 to 23 percentage points on multi-label F1. This trend held true for both label-confidence subsets after solving the problems of class imbalance and numerical stability. One potential reason for this could be that DER++ includes two other learning goals apart from classification, which compete with the latter under limited local training with 3 epochs per round.

The relatively good performance of the No CL baseline indicates that the problem of catastrophic forgetting in the proposed federated continual learning scenario is inherently constrained. The client updates are rather short by only using three local epochs, tasks change gradually rather than abruptly, and the model aggregation via FedAvg implicitly maintains already learned representations among the participating clients. In this context, the contribution of EWC's regularization is rather small and may even limit the adaptation to the new attack behavior patterns, leading to only slight improvement in multi-label performance. On the other hand, the replay approach still has some small but persistent advantage, as the replay samples help to reinforce the representation of minority attacks and the rare combinations of labels, which are harder to learn in the multi-label setting. This advantage is manifested in the better multi-label F1 score but is significantly smaller for the binary classification task, where the decision boundary just separates the benign from the malicious traffic and is therefore much simpler to maintain. The combination of replay with knowledge distillation (DER++) seems to be penalized by the small local training budget with three epochs per round.

Low forgetting can be interpreted in many different ways. All the approaches to continual learning showed very low forgetting with under 0.014 and positive backward transfer. This means that the tested algorithms managed to maintain the previously learned knowledge, although it might be due to the nature of the benchmark used. The change of the set of techniques in between the rounds shown in Figure 3 suggests the existence of some drift, but the sequence of tasks used is not disjoint in time and campaigns. The performance of FedProx is highly dependent on the relationship between the regularization effect and client heterogeneity. FedAvg performed better than FedProx in this benchmark, and the contribution of the proximal term in Aggregation Strategy Comparison subsection helps to explain the reason behind this result. The un-normalized proximal regularization constrained some large volume clients, since their updates are expected to be far apart from the global model. This suggests that the success of FedProx is highly dependent on the relationship among the factors mentioned above.

Technique-set drift gives a concrete way of representing the drift in the dataset. Validation of the dataset has confirmed the consistency and measurability of change in the technique-set turnover from one round to another, while drift at sequence level could not be estimated accurately because of insufficient volume of data in each campaign. Therefore, the benchmark adopts technique-set changes as the operational definition of behavioral drift

Limitations

The overwhelming majority with 88.1% of the technique-tagged records in the dataset belong to the lowest confidence category for attribution, having a median posterior margin of 0.01 shown in Table 2. Although this influenced the choice of high_confidence and full_tagged setup, the high_confidence set represents only 5.4% of the data, leading to higher variance in results. Results on the full_tagged set requires interpretation with the understanding that some technique labels have limited attribution. The benchmark includes five federated clients, representing the number of host agents, but fewer than half of the six rounds involve several of these clients. This is representative of the source data set. However, it constrains the ability to perform statistics at the client level, and the outcomes can be considered a small-scale example.

A penalty cap for EWC was necessary to avoid instability whenever intermittently participating clients came back when the global model had changed, thereby bringing about a contrast between the classical notion of a single learner and that of federated participation. Also, sequence-level drift and sequential-specific architectures were not well-developed due to unreliable sequence drift measurement. FedProx was tested by using only one value for the proximal strength of $\mu = 0.01$. As the size of the proximal term affects the algorithm's performance, varying μ can result in different outcomes. The experiments were carried out using Windows-APT 2025. Despite validation of the dataset prior to modeling, generalizability of the approach to other intrusion detection datasets, telemetry sources, and real-world attackers has not been validated.

Future Work

Several directions for exploration are possible. Firstly, a study of FedProx hyperparameter choice and other federated-aware regularizations can clarify whether this difference is associated with a particular choice of proximal strength or represents a general limitation when using this benchmark setup. Secondly, continual learning approaches for periodic federated learning can be considered which include alternative EWC implementations and different replay techniques. Thirdly, sequence-aware models can be evaluated on campaigns where there is enough historical data to see whether the attack progression modeling will help in addition to technique set drift. Fourthly, more complex federation simulations, more dataset, and uncertainty-aware training objectives can be used to investigate whether the obtained results generalize and utilize uncertain labels effectively.

CONCLUSION

This paper addressed the gap between federated learning, continual learning, and how these concepts are applied in APT intrusion detection, where clients participate intermittently, have no well-defined labels, and exhibit behavioral drift, contrary to some assumptions that are made. The proposed benchmark was designed through empirical analysis of the Windows-APT 2025 dataset. From the findings, it is clear that simple methods still remain very competitive. Replay method performed better compared to complex continual learning algorithms. FedAvg outperformed FedProx because of the effect of non-normalized proximal term used, and convoluted networks offered no benefit over feed-forward networks even if they were more computationally expensive. Further analysis is needed to find out why that is the case even though there was no catastrophic forgetting. Overall, this paper argues that federated continual learning techniques for use in security applications needs a clear understanding of real-world factors including client availability, labeling accuracy, and non-stationarity. The proposed design and evaluation framework can serve as a starting point for investigating privacy-preserving and adaptive learning systems in other environments.

ACKNOWLEDGEMENT

We would like to express our deepest gratitude to all parties who contributed to this research. We would like to thank our colleagues who provided advice, support, and inspiration throughout the research process. We would also like to thank all participants and respondents who took the time to participate in this research. We would also like to thank the institutions that provided support and facilities for conducting this research. All contributions and assistance were invaluable to the smooth running and success of this research. Thank you for all your hard work and collaboration.

REFERENCES

- Agrahari, S., & Singh, A. K. (2022). Concept Drift Detection in Data Stream Mining : A literature review. *Journal of King Saud University - Computer and Information Sciences*, 34(10), 9523–9540. <https://doi.org/10.1016/J.JKSUCI.2021.11.006>
- Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Alazab, M., Bhattacharya, S., Maddikunta, P. K. R., & Gadekallu, T. R. (2022). Federated Learning for intrusion detection

- system: Concepts, challenges and future directions. *Computer Communications*, 195, 346–361. <https://doi.org/10.1016/J.COMCOM.2022.09.012>
- Andresini, G., Pendlebury, F., Pierazzi, F., Loglisci, C., Appice, A., & Cavallaro, L. (2021). INSOMNIA: Towards Concept-Drift Robustness in Network Intrusion Detection. *AISeC 2021 - Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, Co-Located with CCS 2021*, 111–122. <https://doi.org/10.1145/3474369.3486864>
- Apruzzese, G., Laskov, P., Montes De Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos, A. V., & Di Franco, F. (2023). The Role of Machine Learning in Cybersecurity. *Digital Threats: Research and Practice*, 4(1). <https://doi.org/10.1145/3545574/ASSET/C4586A69-4437-435A-9DE6-E5F12828847D/ASSETS/GRAPHIC/DTRAP-2021-0064-F15.JPG>
- Buzzega, P., Boschini, M., Porrello, A., Abati, D., & CALDERARA, S. (2020). Dark Experience for General Continual Learning: a Strong, Simple Baseline. *Advances in Neural Information Processing Systems*, 33, 15920–15930. <https://github.com/aimagelab/mammoth>.
- Criado, M. F., Casado, F. E., Iglesias, R., Regueiro, C. V., & Barro, S. (2022). Non-IID data and Continual Learning processes in Federated Learning: A long road ahead. *Information Fusion*, 88, 263–280. <https://doi.org/10.1016/J.INFFUS.2022.07.024>
- De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2022). A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3366–3385. <https://doi.org/10.1109/TPAMI.2021.3057446>
- Hamed, P., Razavi-Far, R., & Hallaji, E. (2025). Federated continual learning: Concepts, challenges, and solutions. *Neurocomputing*, 651, 130844. <https://doi.org/10.1016/J.NEUCOM.2025.130844>
- Han, M., Chen, Z., Li, M., Wu, H., & Zhang, X. (2022). A survey of active and passive concept drift handling methods. *Computational Intelligence*, 38(4), 1492–1535. <https://doi.org/10.1111/COIN.12520>
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13), 3521–3526. https://doi.org/10.1073/PNAS.1611835114/SUPPL_FILE/PNAS.201611835SI.PDF
- Korycki, Ł., & Krawczyk, B. (2023). Adversarial concept drift detection under poisoning attacks for robust data stream mining. *Machine Learning*, 112(10), 4013–4048. <https://doi.org/10.1007/S10994-022-06177-W/METRICS>
- 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), 2309–2332. <https://doi.org/10.1109/TNSM.2022.3177512>
- Mat, N. I. C., Jamil, N., Yusoff, Y., & Kiah, M. L. M. (2024). A systematic literature review on advanced persistent threat behaviors and its detection strategy. *Journal of Cybersecurity*, 10(1). <https://doi.org/10.1093/CYBSEC/TYAD023>
- Mozaffari, M., Yazdinejad, A., & Dehghantanha, A. (2026). Windows-APT 2025: A dataset for APT-inspired attack scenarios on windows systems. *Data in Brief*, 65, 112569. <https://doi.org/10.1016/J.DIB.2026.112569>
- Sun, H., Zhang, W., Zhang, R., Xu, L., Guan, H., Zhao, H., Zhang, B., Yang, S., & Yu, X. (2025). Federated Continual Learning Based on Weakly Supervised Diffusion Models for Disease Diagnosis. *IEEE Internet of Things Journal*, 12(22), 45958–45971. <https://doi.org/10.1109/JIOT.2025.3535628>
- Susanto, A. K. S., Figrin, I. A., Patchmuthu, R. K., Tajuddin, S. T. B. H., Waruwu, D. S., & Laia, O. (2025). Performance Comparison Analysis of Conventional and Advanced Machine Learning Algorithms for APT Malware Detection. *2025 7th International Conference on Applied Computational Intelligence in Information Systems: Intelligent and Resilient Digital Innovations for Sustainable Living, ACIIS 2025 - Proceedings*. <https://doi.org/10.1109/ACIIS66255.2025.11402962>

- Susanto, A. K. S., Figrin, I. A., Patchmuthu, R. K., Tajuddin, S. T. B. H., & Zai, M. S. P. (2025). A Comparison of Deep Learning Algorithms for APT Malware Detection. *2025 7th International Conference on Applied Computational Intelligence in Information Systems (ACIIS)*, 1–6. <https://doi.org/10.1109/ACIIS66255.2025.11402960>
- Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8), 5362–5383. <https://doi.org/10.1109/TPAMI.2024.3367329>
- Wang, Z. I., Wu, F., Yu, F., Zhou, Y., Hu, J., Min, G., & Wang, Z. (2024). Federated Continual Learning for Edge-AI: A Comprehensive Survey. *ACM Computing Surveys*, 1. <https://doi.org/XXXXXXX.XXXXXXX>
- Zhong, Z., Bao, W., Wang, J., Chen, J., Lyu, L., & Yang Bryan Lim, W. (2025). SacFL: Self-Adaptive Federated Continual Learning for Resource-Constrained End Devices. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9), 17169–17183. <https://doi.org/10.1109/TNNLS.2025.3565827>
- Zhu, H., Xu, J., Liu, S., & Jin, Y. (2021). Federated learning on non-IID data: A survey. *Neurocomputing*, 465, 371–390. <https://doi.org/10.1016/J.NEUCOM.2021.07.098>