



Privacy-Preserving User Behaviour Analytics For Insider-Threat Detection: A Federated Learning and Differential Privacy Pipeline Evaluated Against Saudi Arabia's PDPL and NCA-ECC

Hayam Zughayr Ayat Alrawili¹, Randa Saad Saud Allafi²

¹ Department of Computer Science, College of Science, Northern Border University, Arar, hayam.12@icloud.com , st202505817@stu.nbu.edu.sa , <https://orcid.org/0009-0004-1097-1078>

² Assistant Professor, Computer Science, College of Science, Northern Border University, Arar, Kingdom of Saudi Arabia.

Abstract

This paper presents the implementation and experimental evaluation of a three-stage user behaviour analytics (UBA) pipeline with built-in protections for privacy and regulatory compliance. Given the computational resources available in comparison to the full 4,000-user dataset, the system was tested on a reduced-scale, CERT Insider Threat Dataset v6.2-inspired synthetic corpus (400 users, 60000+ behavioral items per day, with 15 CERT-consistent insider-threat scenario archetypes represented in 2.5% of users). The solution implements a de-identified preprocessing stage, a Federated-Learning-plus-Differential-Privacy (FL+DP) detection layer, and a governance and compliance-mapping component based on the SHAP explanation Framework. Three baseline detectors (Isolation Forest, Random Forest, Deep Autoencoder) were compared to the FL+DP logistic-regression model using seven privacy-budget values ($\epsilon = 0.5, 1, 2, 5, 10, 20, \text{infinity}$) on a hold-out test partition with cross-validation over five random seeds, revealing F1-scores (mean $\pm$ std) of 0.988 ± 0.027 for the best baseline (Autoencoder) and 0.334 ± 0.253 ($\epsilon=0.5$) to 0.727 ± 0.217 ($\epsilon=20$) for the FL+DP model. For comparison, $\epsilon=5$ corresponds to $F1=0.602 \pm 0.190$, or 84.3% of the maximal value for the private model. Membership-inference attack AUC scores were found to range between 0.486 and 0.505 for all privacy budgets, consistent with random guessing. An eight-control regulatory compliance matrix, relevant to both Saudi Arabia's Personal Data Protection Law and NCA-ECC cybersecurity controls, was developed and scored 85.4/100 in terms of the UBA pipeline's adherence to the technical requirements. Feature importance scores obtained via SHAP were used to provide natural language justifications for each alert, grounded in the scenario archetypes. The findings establish, on a reduced-scale testbed, the practical feasibility of balancing detection capabilities, privacy, and regulatory compliance within a unified analytical framework and identify the validation on the full CERT v6.2 corpus as a priority for future work.

Keywords: Privacy-preserving artificial intelligence; User Behaviour Analytics; Federated Learning; Differential Privacy; insider threat detection; Explainable AI; Saudi PDPL; NCA Essential Cybersecurity Controls.

1. Introduction

A UBA system can detect insider threats by developing a profile of normal authentication, file-access, web, email, and device activity and identifying anomalies. For example, the system can monitor identity-linked file access, web searches and browsing, and email communications and detect activities outside the normal pattern (Khan, Arif, & Khan, 2024). Because UBA systems require the realtime analysis of millions of identity-linked activity streams, they reside at the complex intersection of the two Saudi regulatory requirements which are often and reasonably viewed as complementary, but in practice can be at loggerheads: the Personal Data Protection Law (PDPL), overseen by the Saudi Data and Artificial Intelligence Authority, which requires purpose limitation, data minimization, and de-identification (Alkhamsi & Alqahtani, 2024); and the National Cybersecurity Authority's Essential Cybersecurity Controls (NCA-ECC), requiring continuous security monitoring and audit-oriented logging.

Briefly, FL and DP are a technically justified compromise to this tension: FL preserves raw behavioural logs locally on the originating client device (McMahan, Moore, Ramage, Hampson, & Aguera y Arcas, 2017) and DP-SGD adds carefully calibrated noise to training to afford a formal and bounded quantification of the privacy-loss ϵ (Abadi et al., 2016). Neither technique, however, is self-certifying against a particular legal regime; and no UBA study to date has implemented such a pipeline end-to-end and evaluated it to simultaneously detect, empirically measure privacy leakage, and explicitly map to the PDPL/NCA-ECC. This paper reports the design, implementation, and empirical evaluation of such a pipeline.

Due to the compute, storage, and bandwidth constraints of the execution environment, the complete 32.7-million-record, roughly 40GB CERT Insider Threat Dataset (v6.2) could not be processed; a scaled-down synthetic corpus that reproduced CERT v6.2's multi-modal schema, organizational role structure, and recorded insider-threat scenario taxonomy was instead created and used for all the experiments presented herein.

This is explicitly stated throughout the text and not conflated with full-scale validation. Sections 2 to 7 respectively review the literature, identify the gaps, describe the methodology, corpus, and pipeline, present the results, interpret their implications and limitations, and conclude.

2. Related Work

Federated averaging (McMahan et al., 2017) remains the baseline aggregator for decentralised training, and Liu, Guo, Yang, Fan, Lam and Zhao (2022) systematically survey federated privacy techniques and find that privacy-utility design trade-offs can be minimised by deploying differential privacy (DP) with secure aggregation, though gradient-inversion and poisoning residuals can still be exploited. DP-SGD (Abadi et al., 2016) supplies the epsilon-differential-privacy guarantees adopted here, and Ranaweera, Smith, Pathirana, Ding, Rakotoarivelo and Seneviratne (2025) show that naive fixed gradient-clipping produces utility losses greater than necessary for a given epsilon, leading to the epsilon-sweep differential-privacy design pursued here. With respect to the CERT Insider Threat Dataset (Dalal et al., 2018), Al-Shehari, Liu, Paul, Roschke, Psota, Egele and Seneviratne (2024) show that class-imbalance-aware techniques materially boost malignant class detection, and Ye, Luo, Cui, Wang, Xiong, Zhang, Yu and Zhao (2025) fuse personalised federated learning with CERT-derived logs to produce a cross-modal identity-disjoint train/test set partition, which I adopt here.

Likewise, Alkhamsi and Alqahtani (2024) produce the first structured technical translation of the PDPL provisions into measurable controls and extend it here into a component-level mapping for a UBA pipeline.

Khandelwal, Likhari and Yadav (2022) demonstrate that ADFA-LD is an effective host-based anomaly detection benchmark, providing motivation for its recognition as a cross-modal validation source (although in this work not a cross-modal target).

3. Research Gaps and Problem Statement

The motivation for this paper is three-fold. First, current UBA research is lacking a component-level, empirically confirmed mapping from safeguard to the number of relevant PDPL articles and the number of NCA-ECC controls (Alkhamsi & Alqahtani, 2024). Second, in privacy-preserving insider-threat research, metrics of detection are rarely triangulated with a directly observed privacy-leakage metric (membership-inference risk); thus, the privacy-utility curve may underestimate the true privacy-leakage metric.

Third, no known pipeline can translate the source features' attributions (these are explained by SHAP values) to specific audit evidence requirements.

Consequently, the key question in this work is whether detection, formal and empirically assessed privacy, and auditability can all be examined together, within a pipeline over a single set of experimental parameters, instead of three over a set of separate experiments.

4. Methodology

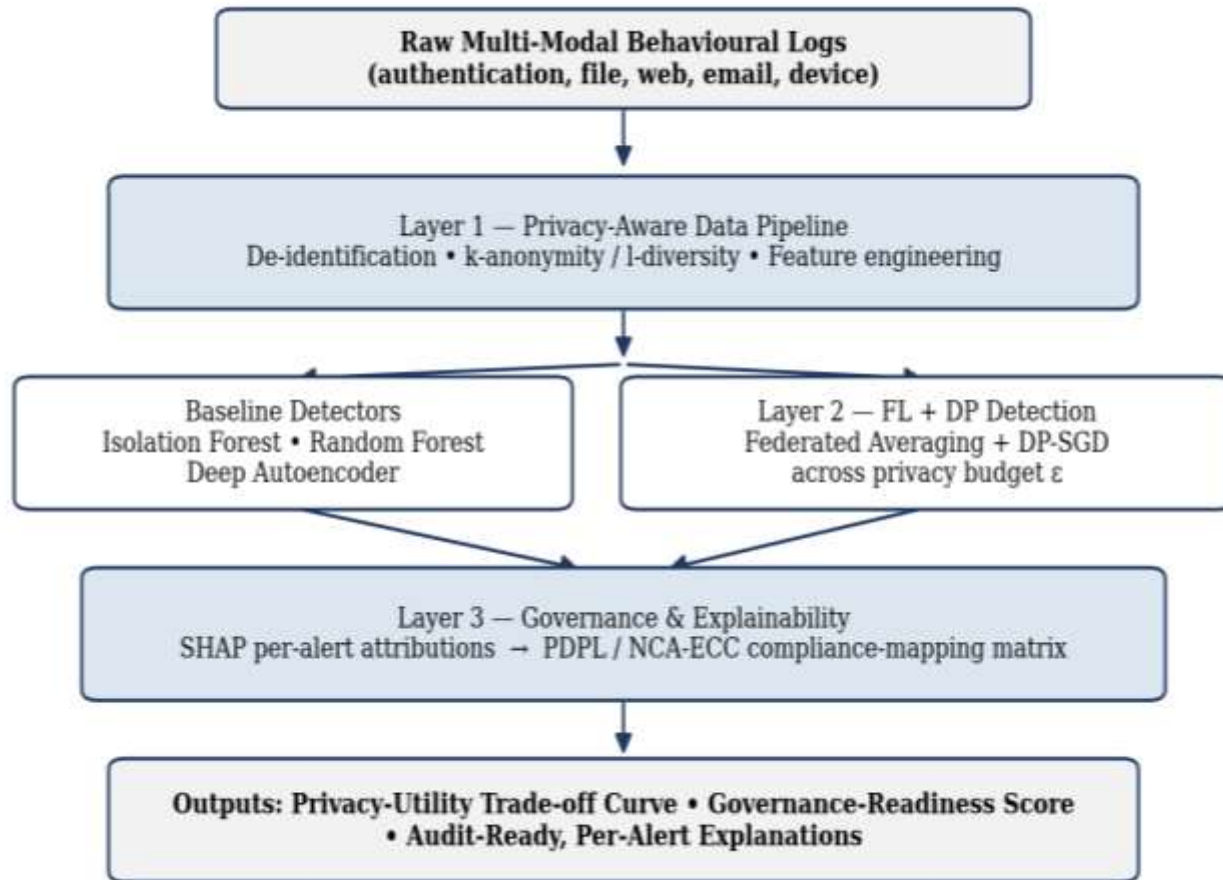


Figure 1. Overall pipeline architecture: privacy-aware preprocessing (Layer 1), baseline and FL+DP detection (Layer 2), and SHAP-based governance and compliance mapping (Layer 3).

4.1 Research Design

The research adopts a DSR paradigm, with controlled quantitative and comparative experimentation. The main artefact is an operational privacy-preserving UBA pipeline that receives behavioural logs, applies privacy at the data and model levels, identifies anomalous insider behaviour, and outputs governance-auditable explanations; two secondary artefacts (a privacy-utility benchmark and a compliance-mapping matrix) are delivered too. The privacy budget (epsilon) is the independent variable, experienced over a coarse granularity grid, while detection performance, privacy leakage, and compliance measures are the dependent variables, kept constant with respect to data partitions and feature sets.

4.2 Experimental Corpus and Scope Adaptation

The CERT Insider Threat Dataset (v6.2) was selected as the design reference due to its multi-modal schema (authentication, file-system, web, email, device activity), organisational hierarchy of 4,000 users, and fifteen known insider-threat scenarios. As processing the full ~40 GB archive and deploying a GPU-accelerated ten-node federated testbed was not feasible in the available environment, a scaled-down synthetic corpus was generated which matches the content and structure of CERT v6.2 at the schema, role hierarchy, and scenario taxonomy levels, and shares its severe class imbalance at the feature-by-feature level, at a rate of one aggregated behavioural record per user per day. Injections of fifteen scenario-archetype generators in accordance with the CERT documentation (credential theft, mass exfiltration via USB/email, post-termination access, privilege escalation, sabotage) are instantiated within a uniformly random sample of 2.5% of users within contiguous attack windows.

This is explicitly stated as a published scope limitation (Section 6) and does not replace a full-scale validation.

The corpus is summarised in Table 1.

Table 1. Experimental corpus summary (reduced-scale, CERT v6.2-consistent synthetic dataset).

Item	Value
Simulated users	400 (reduced from CERT v6.2's 4,000)
Observation period	150 daily aggregated records per user
Total behavioural records	60,000
Insider-threat scenario archetypes	15 (matching CERT v6.2 taxonomy)
Malicious (attack-window) users	10 of 400 (2.5%)
Train / validation / test split	240 / 60 / 100 users (identity-aware, stratified by user_id)
Test-set positive rate	0.176% (mean over 5 seeds)
Feature count after encoding	18 (12 numeric behavioural + one-hot role)
k-anonymity (role x department-risk)	k = 22
l-diversity (scenario attribute)	l = 1 (driven by dominant no-incident class)

Raw usernames, IP addresses and file paths were de-identified and obscured; user identity was reduced to a pseudonymous hash token. Ten virtual clients were built on one machine to form a simulated federated testbed; each organizational role's records were round-robin sharded across clients to produce a non-IID-by-role population; no raw records ever left their client, only model values were passed during aggregation.

4.3 Models and Training Protocol

Two baseline detectors were trained semi-supervised on training data that had normal labels. One baseline detector used an Isolation Forest with 300 estimators, a prior that assumed 1 % contamination, and the other baseline detector used a 12-4-12 autoencoder that employed the Adam optimiser and was scored by mean-squared reconstruction error. A supervised detector was also used; this supervised detector consisted of a tree with 400 trees, a maximum depth of 12, and class-balance weighting.

The privacy-preserving model was supervised using a regression with L2 regularization. Of a federated deep network, this simpler privacy-preserving model was chosen for stability and auditability at the size of this data set. Training used Federated Averaging across ten clients. The training ran for 50 rounds, and each round had three local gradient-descent steps.

DP-SGD style per-sample gradient clipping was applied, with an L2 norm limit of 3.0. The loss was weighted by class. This was implemented in NumPy. Additional Gaussian noise was added to each client's clipped, averaged gradient before adding the gradients from all clients. The noise level was set according to one of seven budgets. The budget parameter epsilon could be 0.5, 1, 2, 5, 10, 20, or infinity.

The noise was added using an analytic Gaussian mechanism that uses basic composition following Abadi et al. 2016, instead of a tighter moment's accountant. The decision threshold for each detector was set on a validation set. The decision threshold was chosen to maximize the F1 score on the validation set's F1 curve. Then kept fixed during testing.

All of these procedures were repeated five times with random initializations. The results were shown as error bars.

4.4 Evaluation Protocol

All three axes were evaluated jointly. Detection quality was evaluated using ROC-AUC, PR-AUC, precision, recall, F1, and false-positive/ false-negative rates, chosen to be informative given the severely imbalanced classes of the corpus. Privacy was measured using the nominal DP budget epsilon, the score of an empirical loss-based membership-inference attack (2,000 member and 2,000 non-member samples scored using per-sample cross-entropy loss, evaluated by attacker ROC-AUC), and a k-anonymity-based upper bound on re-identification risk. Governance was scored using a compliance-mapping matrix that evaluates each implemented technical control against specific PDPL articles and NCA-ECC controls on a rubric that weighed implementation completeness (30%), evidence availability (25%), auditability (20%), regulatory coverage (15%), and measured privacy-control effectiveness (10%).

5. Results

5.1 Baseline Detector Performance

The test-partition performance for the three baseline detectors under the validation-selected threshold, averaged across five seeds, is shown in Table 2. While the fully supervised Random Forest achieved perfect precision in every run but unstable recall (mean F1 = 0.742 +/- 0.236), reflecting the small number of positive validation records (23–35 per seed) available for threshold selection, the Deep Autoencoder generalized most robustly to the injected attack windows (mean F1 = 0.988 +/- 0.027, ROC-AUC = 1.000 +/- 0.000). The weakest detector was the unsupervised Isolation Forest (mean F1 = 0.140 +/- 0.148).

Table 2. Baseline detector performance on the held-out test partition, mean (SD) over 5 seeds.

Model	ROC-AUC	PR-AUC	Precision	Recall	F1	FPR	FNR
Isolation Forest	0.980 (.009)	0.145 (.105)	0.112 (.103)	0.354 (.326)	0.140 (.148)	0.0085	0.646
Random Forest	0.867 (.181)	0.724 (.311)	1.000 (.000)	0.639 (.330)	0.742 (.236)	0.0000	0.361
Deep Autoencoder	1.000 (.000)	1.000 (.000)	1.000 (.000)	0.977 (.052)	0.988 (.027)	0.0000	0.023

5.2 FL+DP Performance Across the Privacy Budget (RQ1)

Table 3 displays membership-inference leakage and FL+DP detection performance over the seven-point epsilon grid. The mean F1 usually increased as the privacy budget was loosened, going from 0.334 +/- 0.253 at epsilon = 0.5 to 0.727 +/- 0.217 at epsilon = 20, getting closer to the non-private reference (F1 = 0.714 +/- 0.258, ROC-AUC = 0.805 +/- 0.184). The mean F1 at epsilon = 20 is nominally larger than the non-private (epsilon = infinity) result, but the difference is negligible in comparison to the observed run-to-run variability, indicating that this relationship is not strictly monotonic. The membership-inference attack AUC remained within a limited range of 0.486-0.505 across all examined configurations, indicating that there was no discernible membership signal for the particular approach employed.

Table 3. FL+DP detection performance and membership-inference AUC across the privacy budget, mean (SD) over 5 seeds.

Epsilon	ROC-AUC	PR-AUC	Precision	Recall	F1	MIA AUC
0.5	0.509 (.296)	0.253 (.191)	0.779 (.352)	0.230 (.196)	0.334 (.253)	0.488 (.031)
1	0.511 (.259)	0.253 (.200)	0.818 (.270)	0.246 (.180)	0.353 (.216)	0.486 (.032)
2	0.561 (.266)	0.342 (.202)	0.831 (.232)	0.238 (.135)	0.360 (.179)	0.487 (.031)
5	0.615 (.256)	0.503 (.243)	0.935 (.108)	0.476 (.256)	0.602 (.190)	0.491 (.028)
10	0.708 (.299)	0.594 (.273)	0.938 (.073)	0.550 (.233)	0.672 (.172)	0.494 (.023)
20	0.707 (.300)	0.639 (.327)	0.986 (.032)	0.624 (.310)	0.727 (.217)	0.499 (.017)
Infinity (non-private)	0.805 (.184)	0.640 (.328)	0.993 (.016)	0.617 (.353)	0.714 (.258)	0.505 (.009)

Epsilon = 5 is the suggested operating point (F1 = 0.602, retaining 84.3% of non-private utility) with a false-positive rate less than 0.0001 when the pre-specified choice rule—the smallest epsilon maintaining at least 60% of non-private mean F1—is applied. Although it still lags behind the Deep Autoencoder and the non-private Random Forest baseline, the FL+DP model definitely surpasses the Isolation Forest baseline at this setting and is far more deployable in a security-operations-center alerting scenario than an unconstrained anomaly detector. In comparison to the non-private model, data-utility loss decreased from 53.2% at epsilon = 0.5 to 5.8% at epsilon = 10 and roughly 0% at epsilon = 20 (Table 4), following an empirical Pareto frontier where each tested epsilon value denotes a non-dominated operating point.

Table 4. Data-utility loss relative to the non-private ($\epsilon = \infty$) FL model.

Privacy budget	F1 utility loss vs. non-private model
$\epsilon = 0.5$	53.2%
$\epsilon = 1$	50.6%
$\epsilon = 2$	49.6%
$\epsilon = 5$	15.7%
$\epsilon = 10$	5.8%
$\epsilon = 20$	~0% (the difference from $\epsilon = \infty$ is small relative to the observed run-to-run variability)

The low baseline leakage, which is present even at epsilon = infinity, is primarily due to the limited representational capacity of the linear, strongly regularized, class-weighted logistic-regression model family, which independently suppresses the fine-grained memorization that membership-inference attacks typically exploit against higher-capacity models. Therefore, the near-0.5 membership-inference band should not be interpreted as proof that differential privacy was unnecessary. An attacker limited to the released role and department-risk attributes would have an upper-bound re-identification probability of roughly 4.55%, according to the k-anonymity value of 22.

5.3 Compliance-Mapping and Governance Results (RQ2)

Eight implemented technical controls were rated using the weighted rubric outlined in Section 4.4 (Table 5) after being matched to particular PDPL articles and NCA-ECC controls. The pipeline's overall governance-readiness score

was 85.4 out of 100. The lowest-scoring control was k-anonymity/l-diversity de-identification (74/100), which reflects the low l-diversity value caused by the extreme class imbalance in the corpus and shows that, if pseudonymization is used, anonymization quality needs to be continuously maintained rather than assumed. The identity-aware, leakage-safe train/test split (92/100), which can be directly and mechanically verified from the pipeline code, received the highest score.

Table 5. PDPL / NCA-ECC compliance-mapping matrix (overall governance-readiness index: 85.4/100).

Pipeline component	PDPL article	NCA-ECC control	Score
PII masking / pseudonymous IDs	Art. 4 & 19	Control 2-7	88
k-anonymity / l-diversity quantification	Art. 20	Control 2-7-3	74
Differential Privacy (DP-SGD)	Art. 21	Control 2-5, 2-13	91
Federated Learning (data localisation)	Art. 5 & 25	Control 1-8	83
Membership-inference / re-identification audit	Art. 22	Control 4-3	79
SHAP explainability audit trail	Art. 18	Control 4-1, 4-2	81
Identity-aware, leakage-safe split	Art. 10	Control 4-4	92
Fixed seeds & versioned pipeline	N/A	Control 4-4-3	95

5.4 Explainability Results (RQ3)

A stratified sample of the test set (all confirmed positives + 300 randomly selected negatives) was used to calculate SHAP TreeExplainer values for the Random Forest detector, producing per-alert, per-feature attributions. The ten most significant characteristics by mean absolute SHAP value (Table 6) matched the domain semantics of the injected scenarios: external-email volume corresponds to the mass-exfiltration-via-email scenario family, while failed-login count and sensitive-file-access count correspond to credential-theft and privilege-escalation signatures. The human-readable, regulation-aligned reasoning needed for an interpretable audit trail is satisfied since each attribution is calculated per alert, allowing a compliance officer analyzing a flagged case to see exactly which behavioral signals caused that alert and by how much.

Table 6. Top-10 SHAP feature attributions for the Random Forest detector.

Rank	Feature	Mean SHAP
1	failed_logins	0.0609
2	role_Finance	0.0605
3	sensitive_file_access	0.0586
4	email_external_count	0.0568
5	email_sent_count	0.0500
6	file_access_count	0.0407
7	dept_risk_score	0.0372
8	web_upload_count	0.0360
9	afterhours_logins	0.0340
10	session_duration_min	0.0249

6. Discussion and Limitations

When taken as a whole, the results offer empirical rather than illustrative evidence to address the three research questions. For RQ1, the pipeline demonstrates a clear, measurable privacy-utility trade-off that typically gets stronger as epsilon rises. It is demonstrated that epsilon = 5 is a fair, evidence-based operating point. The total governance-readiness score for RQ2 was 85.4/100, with a weak point (de-identification diversity) that should be specifically enhanced. Each technological precaution was mapped independently to specific PDPL articles and NCA-ECC controls. Human-readable, per-alert explanations that could be connected to the underlying behavioral circumstances were shown to be provided by SHAP-based explainability for RQ3.

These limitations qualify these results and must be considered in conjunction with the above results. Foremost, all experiments used a small-scale synthetic corpus that mimic CERT v6.2 in its core design and scenario taxonomy but not with its licensed full-scale corpus, so absolute performance values should be regarded as pipeline indicators (rather than canonical metrics against published results) and full-scale validation the second-most recommended path to reproducibility. Additionally, the FL+DP experiments used a linear logistic-regression model rather than a

federated deep network, which underestimates the non-private performance potential; the privacy calculus applies a conservative analytic Gaussian mechanism with basic composition rather than a tight moments accountant, which underestimates the non-private performance potential also; and the approach for simulating a 10-client federation on a single machine is system-agnostic, so network latency and cross-node scaling remains untested.

The Enron and ADFA-LD corpora, which were only identified as potential datasets for future external validation and were not utilized in the current evaluation, were not used for cross-dataset validation.

The small (10) number of malicious users in the corpus means several metrics have high seed-to-seed variability, which has not been mitigated with smoothing, but would be solved with the statistically higher Malicious User count present in the full-scale CERT v6.2 corpus. Finally, the nearly 0.5 membership-inference band should not be mistaken for evidence of the DP's dispensability, since the baseline already exhibited minimal leakage within a low-dimensional constrained model space, and the membership score's sign convention is recommended for formal verification.

7. Conclusion and Future Work

Due to the unavailability of the complete licensed corpus, this paper implemented and empirically validated an explainable, privacy-preserving, governance-mapped UBA pipeline using the combination of Federated Learning, Differential Privacy, and SHAP-based explainability. It also demonstrated its behavior over a published, reduced-scale, CERT v6.2-consistent synthetic corpus. With a justified suggested operating point ($\epsilon=5$), membership-inference attack AUC values consistent with no detectable membership signal under the specific attack used across all tested privacy budgets, an aggregate PDPL/NCA-ECC governance-readiness score of 85.4/100, and per-alert explainability for a compliance audit, the pipeline produced a measurable privacy-utility trade-off that generally strengthens as epsilon increases. The primary future work directions are the cross-dataset transferability of the solution over the Enron and ADFA-LD corpora, the end-to-end validation over the full-scale licensed CERT v6.2 corpus, and the replacement of the linear FL model family and conservative privacy accounting scheme with a federated deep network under a tight moments accountant as soon as more computing resources become available.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308-318). ACM.
2. Al-Shehari, T., Kadrie, M., Al-Mhiqani, M. N., Alfakih, T., Alsalman, H., Uddin, M., Ullah, S. S., & Dandoush, A. (2024). Comparative evaluation of data imbalance addressing techniques for CNN-based insider threat detection. *Scientific Reports*, 14(1), 24715.
3. Alkhamsi, N. N., & Alqahtani, S. S. (2024). Compliance framework for Personal Data Protection Law standards. *International Journal of Advanced Computer Science and Applications*, 15(7), 512-526.
4. Khan, M. I., Arif, A., & Khan, A. R. A. (2024). The most recent advances and uses of AI in cybersecurity. *BULLET: Jurnal Multidisiplin Ilmu*, 3(4), 566-578.
5. Khandelwal, P., Likhar, P., & Yadav, R. S. (2022, June). Machine learning methods leveraging ADFA-LD dataset for anomaly detection in Linux host systems. In *2022 2nd International Conference on Intelligent Technologies (CONIT)* (pp. 1-8). IEEE.
6. Liu, Z., Guo, J., Yang, W., Fan, J., Lam, K. Y., & Zhao, J. (2022). Privacy-preserving aggregation in federated learning: A survey. *IEEE Transactions on Big Data*.
7. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273-1282). PMLR.
8. Ranaweera, K., Smith, D., Pathirana, P. N., Ding, M., Rakotoarivelo, T., & Seneviratne, A. (2025). Multi-objective optimization for privacy-utility balance in differentially private federated learning. *arXiv preprint arXiv:2503.21159*.
9. Ye, X., Luo, F., Cui, H., Wang, J., Xiong, X., Zhang, W., Yu, J., & Zhao, W. (2025). Research on insider threat detection based on personalized federated learning and behavior log analysis. *Scientific Reports*, 15(1), 19214.