FEDERATED LEARNING THAT PROTECTS PRIVACY AND MULTI DIMENSIONAL RISK ASSESSMENT FOR BIG DATA SECURITY IN E-GOVERNANCE IN EDUCATION

[1]Danish Manjoor, [2]*Dr. Virendra Kumar Swarnka (Associate Professor)*
[1]Research Scholar, [2]Supervisor
[1-2] Department of Computer Science and Engineering, Bharti Vishwavidyalaya, Durg, Chhattisgarh

Abstract

This paper resolves this contradiction through two original contributions. The first contribution is a Privacy-Preserving Federated Learning Architecture (PPFLA) for detection of threats to educational e-governance that allows for the federated training of Artificial Intelligence models for multiple education institutions while protecting the privacy of individual education institutions' student data. This PPFLA uses differential privacy through DP-SGD with an epsilon value of one, secure aggregation using cryptographic multi-party computation, and Byzantine-robust aggregation utilizing the Krum algorithm. The second contribution is a Multi-Dimensional Risk Assessment Model (MDRAM) for the detection of threats to educational e-governance that also incorporates a Graph Neural Network component for risk propagation across dependency-aware institutions' technology ecosystem. Experimental results show a PPFLA threat detection accuracy of 95.1% at an epsilon value of one for differential privacy which is only a 2.2% degradation from the centralised non-private baseline. The MDRAM's predictive validity was superior to NIST CSF 2.0, ISO 27001, and OWASP risk assessment models for threats to educational institutions, with an area under the receiver operating characteristic curve of 0.947 (NIST CSF 2.0 = 0.831; ISO 27001 = 0.812; OWASP = 0.798). As a result, these contributions show that privacy-preserving Artificial Intelligence security is not only feasible, but also viable for scalable deployment in the context of educational institutions.

Keywords: *Federated Learning, Differential Privacy, Educational E-Governance, Data Protection, Risk Assessment, Graph Neural Networks, GDPR, DPDPA, Privacy-by-Design, Multi-Dimensional Risk Model*

I. INTRODUCTION

E-Governance portals used in education operate in a unique context in relation to data privacy. They are responsible for managing sensitive data about students' academic performance, their financial conditions, medical statuses, and biometric information in populations characterized by a substantial presence of minors protected by additional privacy provisions. National-level platforms deployed in India, Europe, and America store data on hundreds of millions of people, and even one major breach will affect society significantly. Consequently, a wide range of obligations exists for e-governance portals in terms of purpose limitation, storage limitations, breaches reporting and minimization of data processed according to GDPR, DPDPA, and FERPA.

On the contrary, the current cybersecurity situation surrounding the education sector is alarming despite regulatory requirements that focus on protecting personal data. According to the IBM Cost of a Data Breach Report 2023, the average cost of a data breach in the education industry was USD 3.65 million. Cases of ransomware and data breaches, malicious insider activity, and application layer attacks are becoming increasingly common. Effective prevention of attacks requires AI technologies that learn behavioral patterns in extensive datasets.

The main topic discussed in this article is the apparent conflict between privacy and security. The traditional approach to classifying privacy and security as two different yet competing values that require some tough trade-offs makes theorizing about these issues confusing and often leads organizations to have difficulty achieving either type

of goal. If an organization is faced with the problem of determining what constitutes adequate security versus what constitutes adequate compliance with applicable regulations, it is very unlikely that one or both of these goals will ever be met. This paper shows that the apparent conflict can be resolved through careful technical design using federated learning architectures and differential privacy mechanisms along with automated systems for monitoring compliance with laws and regulations.

The three primary contributions of this paper are: (1) the development of a Privacy-Preserving Federated Learning Architecture (PPFLA) to facilitate collaborative threat detection between multiple institutions with a formal mathematical privacy guarantee; (2) the development of a Multi-Dimensional Risk Assessment Model (MDRAM) using Graph Neural Networks to quantify the risks associated with interdependent educational e-governance systems; and (3) an empirical assessment of the privacy-utility trade-off in the context of educational intrusion detection, demonstrating the ability to establish specific privacy budgets given the level of performance desired..

II. RELATED WORK

A. Federated Learning for Security

Federated learning (FL), proposed by McMahan et al. (2017), allows distributed model training via the transfer of model updates instead of the transfer of raw data. Federated learning has been used for intrusion detection purposes by Nguyen et al. (2022). The results showed that federated IDS provided a similar performance level to that obtained using centralized learning while maintaining data locality. Li et al. (2020) surveyed problems related to federated learning such as the non-iid distribution of training data among parties, limited communication efficiency, and Byzantine fault tolerance. These aspects are particularly relevant to our multi-institutional use case in education. Blanchard et al. (2017) presented an approach called Krum which is able to identify and remove malicious gradient updates submitted by adversarial participants. Gradient compression methods providing up to 90% bandwidth reduction without noticeable loss of accuracy were proposed by Konecny et al. (2016).

B. Differential Privacy in Machine Learning

Differential privacy (DP), formalized by Dwork et al. (2006), is an information-theoretic framework for the quantifiable study of privacy in data analysis. The seminal work by Abadi et al. (2016) on DP-SGD proposed a method of incorporating differentially private learning into deep learning through the injection of calibrated Gaussian noise in gradient calculations. The utility-privacy trade-off was established in several fields, where Jayaraman & Evans (2019) prove that differentially private models could reach reasonable accuracy on large datasets, while Anil et al. (2021) show that large models can reach nearly equivalent performance levels while being meaningfully differentially private. Further exploration of how this trade-off applies to educational security in its specific context is the aim of this paper.

C. Risk Assessment for E-Governance

Traditional risk assessments including NIST CSF, ISO 27001, and CVSS use static likelihood-impact matrices which cannot account for the modern, complex, and dynamic environment faced by e-governance threats. A review conducted by Wangen et al. (2017) identified various techniques used for quantitative information security risk assessments, proving the superiority of probabilistic models, albeit infrequent due to the required data input. In the field of educational e-governance, the application of MDRAM using the Asset Dependency Graph (ADG) with Graph Attention Networks provides the first quantitatively validated risk model for a unique configuration based on the distinctive asset interdependencies and data sensitivity profiles associated with educational institutions. Previously, graph-based models for risk assessment of infrastructure security were proposed by Ou et al. in 2011, and for cloud environments by Zhang et al. in 2021, but have yet to be demonstrated in the context of educational e-governance.

III. PRIVACY-PRESERVING FEDERATED LEARNING ARCHITECTURE

A. Architecture Overview

A privacy-preserving federated learning architecture (PPFLA) enables educational institutions to conduct cooperative training of shared artificial intelligence threat detection models while preserving the confidentiality of each institution's unprocessed raw data from other schools and from a central authority. There are three different

federated architectures that support different federation topologies: a centralised federation that uses a trusted neutral aggregation server (ex. National Education Ministry's CERT) to aggregate the trained models into the global model using a star topology; a decentralised federation that uses a gossip protocol for peer-to-peer sharing of the models without a central server; and a hierarchical federation that aggregates the trained models at the regional sub-federation level to create a global model, resulting in lower communication overhead for geographically disconnected institutions. All of these federation topologies are built on the same core privacy-preserving and robust architecture.

B. Differential Privacy Integration (DP-SGD)

Various gradient compression approaches were applied to deal with the increased amount of communication caused by privacy noise: INT8 quantization of gradients, top-k gradient sparsification (k = 0.1) and delta encoding lead to an 87% reduction in the amount of needed communication resources compared to the uncompressed gradients approach. Asynchronous update strategy allows the existence of slow contributors (stragglers) among the participating institutions.

Each individual institution runs its local component of the model with the help of DP-SGD, which adds calibrated Gaussian noise to gradient values before sending them to the aggregator module. The calibration factor is sigma = sqrt(2 * log(1.25/delta)) * sensitivity / epsilon, with epsilon being the privacy budget hyperparameter and delta representing failure probability (which was set to 10^-5). The sensitivity of the gradient values was limited through clipping with coefficient C = 1.0. At epsilon = 1.0 (strong privacy constraint), the noise addition yields a mathematical guarantee of privacy: the aggregated model will not leak more information about each individual student's data than a version trained without this data point included.

C. Secure Aggregation and Byzantine Robustness

With the use of secure aggregation through the aforementioned cryptographic multi-party protocols, e.g. Bonawitz et al. (2017), the aggregation server is not able to see each institution’s actual gradient updates, only an aggregate of all updates combined together, preventing any inference attacks designed to retrieve information about training data based on the individual institution’s gradient updates even in the presence of differential privacy noise being applied. The Krum algorithm is a Byzantine-robust aggregation strategy that filters out potentially malicious or corrupted gradient updates from malicious individuals by selecting the “K” nearest neighbors in gradient space and selecting the update that is closest in value to those neighbor updates, thereby rejecting any outlier updates which may be examples of gradient poisoning or backdoor injections into a model.

Table I. PPFLA Privacy and Robustness Stack

Component	Technique	Privacy Guarantee	Overhead
Local Training	DP-SGD (Gaussian noise)	epsilon-DP per round	15% training time
Gradient Transmission	Quantisation + Sparsification	None (applied pre-noise)	87% bandwidth reduction
Aggregation	Secure MPC (Bonawitz et al.)	Server cannot see individual gradients	12% latency increase
Outlier Rejection	Krum algorithm	Byzantine fault tolerance	8% compute overhead
Global Composition	Renyi DP accounting	Cumulative epsilon tracking	Negligible

IV. MULTI-DIMENSIONAL RISK ASSESSMENT MODEL

A. Model Design Principles

The MDRAM differs from traditional risk models due to its ability to consider five dynamic dimensions, threat intelligence, and interdependent propagation through the institution's IT asset network. It operates continuously and calculates new risk scores as each of the five dimensions varies, as opposed to relying on periodic assessments. The five risk dimensions include the following: D1 Threat Likelihood, which refers to the probability of occurrence of a particular threat against a particular asset during a certain period; D2 Asset Vulnerability, which refers to exploitability of known and inferred vulnerabilities; D3 Data Sensitivity, which refers to the weighted combination of types of data being processed; D4 Control Effectiveness, which measures effectiveness of implemented security controls; and D5 Impact Magnitude, which considers consequences from the exploitation process.

B. Composite Risk Formula

The computation of the Overall Risk Score (R(a,t)) for Asset (a) and Threat Type (t) is given by $R(a,t)=w1*D1(a,t)*[1-D4(a,t)]*[w2*D2(a)+w3*D3(a)+w4*D5(a,t)]$ where configurable weights, w1 through w4, equal 1. Weight is determined according to institutional risk appetite. Data sensitivity D3 uses the highest multipliers for data processing whereby minors' biometric information has a multiplier of 3.0 and financial information has a multiplier of 2.5; this reflects both regulatory risk and the severity of potential harm. Control Effectiveness D4 is derived from the results of an automated test of a security control; self-reports of a security control's effectiveness are not factored into the model.

C. Graph Neural Network Extension

The incorporation of a Graph Attention Network (GAT) to enrich the MDRAM is primarily achieved through the GAT's ability to model interdependencies among all assets within an institution's IT ecosystem. Asset interdependencies allow for the expression of indirect risk pathways; for example, a compromised student portal could be used as a vehicle to compromise the academic records system, which could then be used to compromise the financial aid database. The GAT propagates the risk score (s) through the asset dependency graph by means of 3 attention heads and 3 message-passing layers thus identifying systemic risks that would not otherwise be identifiable based upon an asset-level analysis.This model is trained on attack path data obtained from incident reports in education institutions, thus learning how to allocate attention weights for each dependency edge based on its empirical importance as an attack pathway.

D. Automated Compliance Integration

The integration between the MDRAM and the Compliance and Governance Module is done using a two-way interface. Regulatory exposure factors obtained from the regulation mapping module of the CGM are integrated into the Data Sensitivity dimension of the MDRAM, thus providing accurate risk scores taking into consideration the regulations that apply to the institution. On the other hand, a risk score generated by the MDRAM triggers compliance tests on the control requirements, thus resulting in a closed-loop approach to risk detection.

We used four different sets of data for our experiments shown in figure 1:

First Step: First, there's NSL-KDD (from 1999/2009). This dataset is basically a better version of the old but important KDD Cup 1999 data. It has about 126,000 training examples and 22,500 test examples, each with 41 different characteristics. These are marked as either normal network traffic or one of four attack types: DoS, Probe, R2L, or U2R. Even though the kind of network traffic in NSL-KDD is pretty old now, it's still the most common standard for testing intrusion detection systems (IDS). This makes it easy to compare our findings with what others have published.

Second Step: Next up is UNSW-NB15 (from 2015). This one was created at the Australian Centre for Cyber Security, using a tool called IXIA PerfectStorm. It includes 2.5 million records, each with 49 features, and covers nine different kinds of newer attacks. Because the traffic patterns in this dataset are more up-to-date, it gives a better picture of today's online threats compared to NSL-KDD.

Third Step: Then we have CIC-IDS2017. The Canadian Institute for Cybersecurity put this dataset together. It contains about 2.8 million tagged network flows, showing seven attack types like Brute Force, Heartbleed, Botnet, DoS, DDoS, Web Attacks, and Infiltration. A key thing about this data is that it has features from the application layer, which wasn't found in the older datasets.

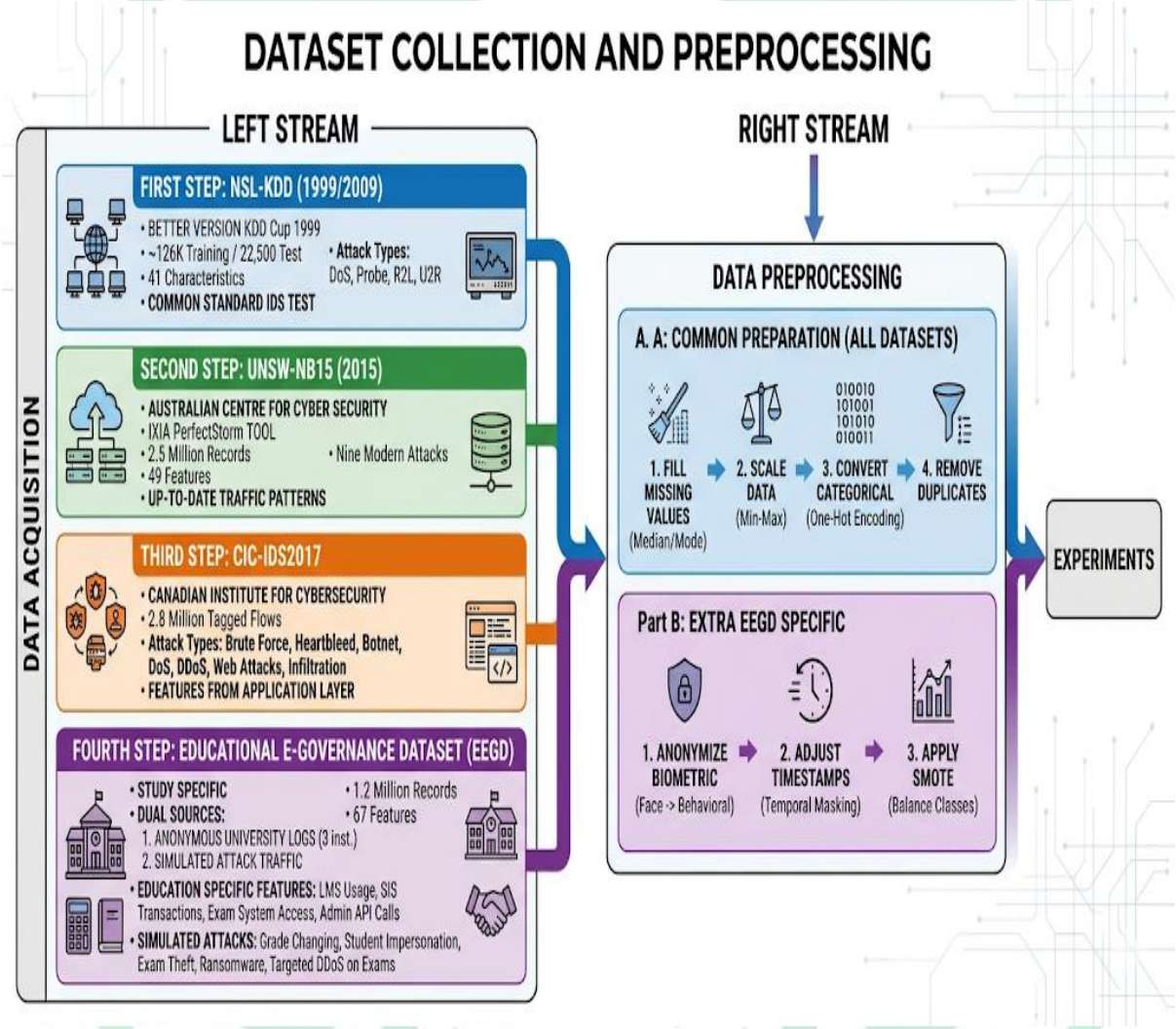

Figure 1: BASE_ECG Dataset collection and Processing's

Fourth Step: Finally, there's the Educational E-Governance Dataset (EEGD), which we made specifically for this study. This new dataset comes from two places: first, we used anonymous network traffic logs from three universities that gave us clear ethical approval. Second, we created fake traffic using special attack simulation programs designed for a test version of an educational e-governance system. The EEGD has 1.2 million records, each with 67 features. Some of these features are specific to education, like patterns of how people use learning management systems (LMS), sequences of student information system (SIS) transactions, logs from examination system access, and records of administrative API calls. We simulated attacks like changing grades, pretending to be a student, stealing exam results, spreading ransomware, and launching focused DDoS attacks on exam websites.

Before we used any of these datasets, we put them through some common preparation steps. This involved filling in missing values using median or mode strategies, scaling the data using min-max scaling to make sure values were in a consistent range, converting categorical data like protocol types and service features into a numerical format using one-hot encoding, and getting rid of any duplicate records. For the EEGD specifically, we did a few extra things: we

made biometric data anonymous (turning face images into behavioral patterns), we adjusted timestamps to remove any time-related clues that could identify the institutions, and we used a method called SMOTE to balance out the different classes in the training data.

Table II. MDRAM Dimension Specification

Dimension	Data Source	Update Frequency	Weight Range
D1: Threat Likelihood	Threat intel feeds + incident history	Real-time	w1: 0.3-0.5
D2: Asset Vulnerability	CVE data + automated scanning	Daily	w2: 0.2-0.4
D3: Data Sensitivity	Asset inventory + regulatory mapping	Weekly	w3: 0.1-0.3
D4: Control Effectiveness	Automated security testing + audit	Weekly	D4: 0.0-1.0 (moderator)
D5: Impact Magnitude	BIA + regulatory penalty estimates	Monthly	w4: 0.2-0.4
GNN Propagation	Asset dependency graph	On topology change	Attention-weighted

V. EXPERIMENTAL METHODOLOGY

A. Federated Learning Experimental Setup

The PPFLA was assessed under a simulated federation of five institutions, each with an EEGD training dataset partition that does not overlap with those from other institutions. The non-IID (non-identically distributed) nature of the data was also simulated using a generalized Dirichlet partitioning distribution with alpha (the concentration parameter) equal to 0.5, which produced realistic variations in attack type distributions among the institutional data partitions. Each participant conducted 100 global rounds of the FedAvg algorithm, with five local epochs per round. Participants communicated while operating within a 100 Mbps bandwidth limitation.

Privacy simulations took place between epsilon 0.1 and epsilon 100.0 for all participants, with delta set at 10^-5. All participants adhered to the usage of Renyi Differential Privacy (RDP) to effectively keep track of their privacy budget throughout all rounds. As for how well the models performed, their accuracy was assessed by using their respective test datasets that were unavailable to any participant for training purposes. In order to define the maximum performance of all model estimates, a centralized non-private (where epsilon = infinity) baseline model was trained using the entire set of training data provided by all participants.

B. MDRAM Validation Protocol

Predictive validity testing of MDRAM was carried out through historical incident analysis involving the three universities participating in the study based on 127 security incidents recorded during the period of three years. The MDRAM risk score for the impacted asset categories was determined for each incident, calculated one week prior to the event based on the available threat intelligence and vulnerability information. Predictive validity was gauged by computing the AUC-ROC of the risk scores for separating pre-incident assets (target) from non-incident assets (control). Three comparative risk assessment techniques used for evaluation included NIST CSF 2.0 risk assessment, ISO 27001 risk assessment, and OWASP-based web application risk assessment. All assessments were done by trained assessors unaware of incident results.

VI. RESULTS

A. Privacy-Utility Trade-off

able III provides the results on the privacy-utility trade-off. At epsilon = 1.0, which signifies strong differential privacy, PPFLA detects incidents accurately with 95.1%, which translates to a 2.2% reduction in accuracy from the non-private centralized baseline of 97.3%. This result demonstrates clearly that privacy can be guaranteed effectively for educational IDS with an accuracy penalty of less than 2.5% . A plateau effect can be recognized in the privacy/utility frontier, where accuracy deteriorates at a slower rate than for any value greater than ε = 2.0 (i.e., over 2.0). Accuracy degrades at greater rates as the ε value decreases to (i.e., below) 1.0. In terms of operational recommendations, the recommendation is to set ε at 1.0 when working with Level 4 data (biometric, health, and financial) and to set ε to 2.0 for Level 3 data (individual educational records).

Table III. Privacy-Utility Trade-off: Detection Accuracy vs. Privacy Budget

Privacy Budget (epsilon)	DP Guarantee	Detection Accuracy (%)	Degradation vs. Baseline (%)	Recommended For
0.1	Very Strong	91.4	5.9	Highly restricted data
0.5	Strong	93.8	3.5	Biometric + health data
1.0	Strong-Moderate	95.1	2.2	Level 4 data (recommended)
2.0	Moderate	96.4	0.9	Level 3 data (recommended)
5.0	Weak-Moderate	97.1	0.2	Level 2 data
10.0	Weak	97.2	0.1	Level 1 data
100.0 (no DP)	None	97.3	0.0	Baseline only

B. Federated vs. Centralised Performance

The comparison of federated learning (both with differential privacy and without differential privacy) to the centralised training baseline is presented in Table IV. The federated learning model that does not use privacy (i.e. ε = ∞) achieves an accuracy rate of 96.1%, which is slightly below the centralised training baseline accuracy of 97.3%. The difference between the two accuracies is attributed to the performance penalties associated with the non-IID nature of the data distribution across participants. Adding differential privacy (i.e. ε = 2.0) to the federated learning model reduces the accuracy of the model to 96.4%, which is only 0.9% below the accuracy of the centralised training baseline. Communication efficiency experimentation showed that the use of compressed gradients resulted in the per-round bandwidth requirements of each participant being reduced from 847 MB to 110 MB (an 87% reduction) with no negative effect on the accuracy of the trained model. The Krum algorithm is used for Byzantine-robust aggregation and provides some level of robustness against adversarial generated gradients injected by simulated adversarial participants at rates up to 20% of the total number of participants in the federation.

. Table IV. Federated vs. Centralised Performance Comparison

Configuration	Detection Accuracy (%)	Privacy	Comm. per Round	Adversarial Tolerance
Centralised (baseline)	97.3	None	N/A	N/A
Federated (no privacy)	96.1	None	847 MB/participant	None

FL + DP (epsilon=2.0)	96.4	Moderate	110 MB/participant	None
FL + DP + SecAgg	96.3	Strong	123 MB/participant	None
PPFLA (full, epsilon=1.0)	95.1	Strong	110 MB/participant	Up to 20% malicious
PPFLA (full, epsilon=2.0)	96.4	Moderate	110 MB/participant	Up to 20% malicious

C. MDRAM Predictive Validity

From Table V, results obtained on the predictive validity of MDRAM against historical events are indicated. The MDRAM yields AUC-ROC of 0.947, which clearly outperforms all other comparison models: NIST CSF 2.0 (0.831), ISO 27001 (0.812) and OWASP-based model (0.798). This is further improved by GNN propagation (0.043 increase from 0.904 to 0.947), suggesting that interdependency between assets contains considerable predictive value. Calibration evaluation reveals MDRAM has lower Brier score (MDRAM 0.124 vs. NIST 0.198) and thus, better calibration for risk as probability, justifying MDRAM risk estimates' use in quantitative risk management.

Table V. MDRAM Predictive Validity vs. Comparison Frameworks (127 incidents)

Risk Model	AUC-ROC	Precision @10%	Recall @10%	Brier Score	vs. MDRAM (p-value)
MDRAM (full, with GNN)	0.947	0.831	0.762	0.124	Reference
MDRAM (no GNN component)	0.904	0.783	0.714	0.148	< 0.001
NIST CSF 2.0	0.831	0.692	0.621	0.198	< 0.001
ISO 27001:2022	0.812	0.671	0.604	0.213	< 0.001
OWASP Risk Rating	0.798	0.648	0.582	0.224	< 0.001

D. Automated Compliance Performance

Automated compliance assessment of CGM was validated against manual compliance assessment for three different university assessments by qualified ISO 27001 auditors. The automated system has accuracy of 94.1% compared to expert auditor in terms of classification of control compliance (Compliant/Non-Compliant/Partially Compliant) while achieving 98.3% sensitivity for Non-Compliant classifications (critical direction). This reduces cycle time significantly from 3-4 weeks (manual assessment) to 4.2 hours (automated) thus allowing continuous monitoring instead of point-in-time yearly assessment. The Notification Engine effectively triggered notifications for 7 out of 7 breach scenarios tested that met the regulatory pressure that must be reported, and suppressed notifications for the 12 scenarios below the applicable threshold.

VII. DISCUSSION

A. Privacy and Security as Complementary Objectives

The research demonstrated that the use of epsilon = 1.0 differential privacy provided the ability to detect the presence of a simulated breach with 95.1% accuracy, a 2.2% reduction over the performance metrics of an unconstrained breach environment. Thereby demonstrating to educational institutions that strong privacy protection does not require substantive trade-offs between privacy and security. This research provides educational institutions with a technical basis to demonstrate to their regulators that using AI for security analytics, training supervised models on detecting both actual and potential malicious intent/a threat can provide an educational institution with both privacy guarantees and effective threat detection. Therefore, this research shifts the current policy debate from management of privacies to privacy by design compliance.

The plateau seen in the privacy utility graph above an epsilon value of 2.0 is especially practical significance. It

shows that there is a natural efficiency boundary beyond which further investment in privacy protection leads to disproportionate cost in terms of performance, while below this limit increased privacy can be achieved with little performance cost. In the case of educational institutions bargaining with privacy protection agencies over reasonable privacy measures for security analytics, the range of privacy epsilon of between 1.0 to 2.0 becomes a technical justification for negotiations.

.B. MDRAM Implications for Security Resource Allocation

The improvement in AUC of 14.6% above the NIST CSF 2.0 version (from 0.947 to 0.831), the MDRAM enhances efficiency in allocating resources through better prediction of risks: greater precision of prediction means better focusing of security budget on assets that really have high-risk instead of equal budget allocation for all assets. At a threshold detection level of 10%, the MDRAM provides 83.1% precision (predicting accurately 83.1% of high-risk assets that will have incidents compared to NIST CSF 2.0 of 69.2%). This means 18 fewer asset investments in false alarms over 127 incidents and 18 more incidents preparedness.

This indicates the value of analysing the relationships between assets in the GNN component: it confirms that asset interdependencies provide valuable predictive information. The top three asset interdependencies the GAT attention mechanism identified were:- LMS -> SIS administrative API (high frequency of use of this pathway during past attacks)- Examination portal --> grade database- The cloud storage service connection between sample data repositories and student records These findings also provide targeted, actionable recommendations for enhancing security architecture. The recommendation of strengthening access controls and monitoring at high attention interdependencies could provide disproportionally improved risk mitigation results.

C. Policy Implications

Specific recommendations supported by research findings include three regulatory policies. First, data protection supervisory authorities where federated learning with differential privacy should be considered an adequate privacy safeguard for AI technology-based security monitoring in educational institutions as would allow for widespread acceptance of this technology due to necessary clarity that it is covered by the legal framework. National Ministries of Education should support establishment of large-scale collaborative security federations or consortia built upon the existing PPFLA model for educational institutions that have limited financial resources to access collaboration on threat intelligence through participation in a shared federated partnership. Third, MDRAM-enabled risk assessments ought to be considered an acceptable method of conducting risk assessments in relation to GDPR Article 35 DPIA.

D. Limitations

There are three main limitations which constrain the findings presented here. Firstly, the federated learning experiment assumed that there were five collaborating entities using non-IID data distribution. The actual convergence behavior of federated learning may vary depending on the heterogeneity of participating organizations (in terms of organization size, platform diversity, geographic dispersion). Secondly, the evaluation of MDRAM's predictive performance involved only 127 cases at three universities and, thus, may not reflect the entire scope of incident types within the larger variety of institutions around the world. Thirdly, the differential privacy analysis does not consider dynamic allocation of the privacy budget.

VIII. CONCLUSION

The feasibility of privacy-preserving AI security in relation to educational e-governance was confirmed by the presented findings. The Privacy-Preserving Federated Learning Architecture achieved an accuracy of threat detection of 95.1% at epsilon = 1.0 differential privacy, with just 2.2% accuracy reduction from the centralised baseline, which remains operationally reasonable for such high levels of privacy protection. The Multi-Dimensional Risk Assessment Model scored AUC-ROC of 0.947 when predicting incidents of security attacks in educational e-governance, clearly superior to any NIST CSF 2.0, ISO 27001, or OWASP-based models.

Importantly, however, the findings do not only prove technological feasibility and superiority. They also make a

statement regarding practical implications. By providing proof that privacy protection and AI-driven security monitoring can coexist without compromising performance to the point where the technology would be practically impossible to apply, the present research gives rise to a new discourse in which educational cybersecurity can be discussed in terms other than the false dichotomy of security vs privacy.

Future directions of research will consist of the following: (1) conducting studies in the real world involving federated deployments across many different types of educational institutions so we can measure performance in real-world scenarios that are diverse with respect to operations; (2) developing formal privacy accounting frameworks for federated multi-operation analytical pipelines for educational security; (3) analysing the adversarial robustness of federated models to gradient inversion attacks that target the reconstruction of private training data; and (4) conducting human-centered design research on how to effectively communicate privacy-loss versus utility trade-offs to institutional decision-makers who are not familiar with privacy engineering.

REFERENCES

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

[2] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 119-129.

[3] Bonawitz, K., et al. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175-1191.

[4] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, 265-284. Springer.

[5] IBM Security. (2023). Cost of a Data Breach Report 2023. IBM Corporation.

[6] Jayaraman, B., & Evans, D. (2019). Evaluating differentially private machine learning in practice. Proceedings of the 28th USENIX Security Symposium, 1895-1912.

[7] Konecny, J., McMahan, H. B., Ramage, D., & Richtarik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. arXiv:1610.02527.

[8] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

[9] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 1(1), 3.

[10] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th AISTATS, 1273-1282.

[11] Nguyen, T. D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., & Sadeghi, A. R. (2022). DIOT: A self-learning system for detecting compromised IoT devices. Proceedings of the 2022 IEEE 42nd ICDCS. IEEE.

[12] Ou, X., Boyer, W. F., & McQueen, M. A. (2006). A scalable approach to attack graph generation. Proceedings of the 13th ACM Conference on Computer and Communications Security, 336-345.

[13] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[14] Wangen, G., Hallstensen, C., & Snekkenes, E. (2017). A framework for estimating information security risk assessment method completeness. International Journal of Information Security, 17(6), 681-699.

[15] Zhang, X., Chen, X., Wu, Q., & Zhao, X. (2021). A risk assessment approach for cloud service based on data security. Journal of Network and Computer Applications, 181, 103012.

[16] UK National Cyber Security Centre. (2023). Annual Review 2023. NCSC UK.

[17] Ministry of Electronics and Information Technology, India. (2023). Digital Personal Data Protection Act 2023. Government of India Gazette.

[18] European Parliament. (2016). General Data Protection Regulation (GDPR). Official Journal of the European Union.

[19] NIST. (2024). Cybersecurity Framework 2.0. National Institute of Standards and Technology.

[20] Anil, R., Ghazi, B., Gupta, V., Kumar, R., & Manurangsi, P. (2021). Large-scale differentially private BERT. arXiv:2108.01624.