RESEARCH PAPER

Medsecai: A Federated Multi-Layer Artificial Intelligence Framework For Cyber Security Threat Detection In Electronic Health Record Systems

Arunkumar Palanichamy[1], Sivakumar Dhandapani[2*]

[1] Assistant Professor, Department Of Computer Science And Engineering, Amet University, Chennai, Tamil Nadu, India.

[2*] Professor, Department Of Computer Science And Engineering, Amet University, Chennai, Tamil Nadu, India.

Correspondence: Arunkumar Palanichamy, Email: saamy.arun@gmail.com; Sivakumar Dhandapani, Email: sivakumar.d@ametuniv.ac.in

Received: *20th Feb, 2026;* ***Revised:*** *4th Mar, 2026;* ***Accepted:*** *25th Mar, 2026;* ***Available Online:*** *10th Apr, 2026*

Abstract

Electronic health record (ehr) systems aggregate longitudinal patient data at a scale that renders them the primary target of healthcare-sector cyberattacks. Conventional signature-based intrusion detection systems are unable to identify novel, polymorphic threats that adapt faster than static rule-bases can be updated. This study proposes medsecai, a four-layer ai/ml security framework designed to address this limitation. The architecture integrates: (i) a bidirectional long short-term memory (bi-lstm) autoencoder with a dynamic sliding-percentile threshold for session-level anomaly detection; (ii) a gradient boosting (xgboost) multi-class threat classifier that ingests lstm reconstruction error as an explicit feature — a two-stage coupling not previously demonstrated in ehr security literature; (iii) a reinforcement learning (rl) policy engine based on proximal policy optimisation (ppo) for adaptive access control; and (iv) a federated learning aggregation layer supporting privacy-preserving collaborative training across multiple hospital clients under differential privacy ($\varepsilon = 0.5$). Evaluated on the cicids-2017 benchmark comprising 2.8 million network flow records, the proposed framework achieves 98.7% intrusion detection accuracy, a false positive rate of 0.9%, and a 428× reduction in inter-institutional data transmission compared with a centralised equivalent. Ablation against four baselines confirms that each architectural component contributes measurably to overall performance. The compliance implications of the federated design are examined under hipaa, gdpr, and india's digital personal data protection (dpdp) act 2023.

Keywords: Electronic Health Records, Cybersecurity, Intrusion Detection, Bi-Lstm, Xgboost, Federated Learning, Differential Privacy, Anomaly Detection, Shap, Dpdp Act.

How To Cite This Article: Palanichamy A, Dhandapani S. Medsecai: A Federated Multi-Layer Artificial Intelligence Framework For Cyber Security Threat Detection In Electronic Health Record Systems. Int J Drug Deliv Technol. 2026;16(26s):869-878. Doi: 10.25258/ijddt.16.26s.91

1. Introduction

The global transition toward digitised healthcare infrastructure has positioned Electronic Health Record systems as the cornerstone of clinical operations. EHR platforms consolidate diagnostic histories, pharmaceutical orders, laboratory results, and biometric identifiers into unified longitudinal records, accelerating clinical workflows and enabling evidence-based decision-making at scale. However, this concentration of sensitive personal data has simultaneously rendered healthcare the most heavily targeted sector for data breaches. IBM's annual cost-of-breach analysis [1] reports that healthcare data incidents incur the highest average remediation cost of any industry, exceeding $10 million per event, a figure that has risen consecutively for thirteen years. The asymmetric value of health records on illicit markets — estimated at ten times the value of a financial credential [2] — drives persistent attacker motivation that static, perimeter-oriented defences cannot adequately counter.

Legacy cybersecurity architectures in clinical environments were designed around a hard-boundary model: firewalls, signature-based antivirus engines, and access control lists formed an outer shell protecting relatively static internal systems. The accelerating adoption of cloud-hosted EHR platforms, mobile clinician access, telehealth integrations, and Internet of Medical Things (IoMT) devices has dissolved this perimeter. Attackers have exploited the resulting exposure through coordinated ransomware campaigns that encrypt operational databases, double-extortion tactics that combine encryption with exfiltration [3], and credential-based intrusions that leverage password-spraying against single-factor authentication systems still prevalent in many hospital environments [4]. Insider threats — encompassing both malicious employees and negligently compromised accounts — constitute a further category responsible for a substantial fraction of documented EHR exposures [5].

Artificial intelligence and machine learning offer a fundamentally different detection paradigm. Rather than matching observed events against a static catalogue of known attack signatures, AI-driven systems learn the statistical distribution of legitimate user behaviour from historical data and identify deviations from this distribution as candidate threats. This capability is especially valuable in clinical settings, where authorised personnel operate under procedurally constrained, highly predictable access rhythms: a ward nurse's record access pattern differs systematically from that of a billing administrator, providing a stable behavioural baseline against which anomalies become detectable with high sensitivity [6]. Recent deep learning methods, particularly recurrent architectures applied to sequential audit log data, have substantially surpassed classical ensemble methods on temporal anomaly detection benchmarks [7].

A persistent limitation in the existing literature is the absence of a unified framework that simultaneously addresses (a) sequential anomaly detection, (b) multi-class threat classification with compliance-compatible explanations, (c) adaptive policy enforcement, and (d) privacy-preserving cross-institutional model training. Individual components have been studied in isolation: federated learning has been applied to clinical outcome prediction [8]; XGBoost has been evaluated for intrusion classification [9]; and LSTM autoencoders have been explored for behavioural anomaly detection [10]. However, their architectural integration into a coherent EHR-specific security pipeline, evaluated on a standardised public benchmark with differential privacy guarantees, remains an open problem.

This paper addresses that gap through the following principal contributions:

- A two-stage detection pipeline wherein Bi-LSTM reconstruction error is injected as an explicit feature into an XGBoost classifier — an architectural coupling that improves multi-class threat labelling beyond what either component achieves independently.
- A dynamic sliding-percentile anomaly threshold that adapts to concept drift in clinical access patterns without manual recalibration, evaluated quantitatively against a fixed-threshold baseline.
- A federated aggregation layer using accuracy-weighted FedAvg under Gaussian differential privacy ($\varepsilon = 0.5$), demonstrating that privacy-utility trade-off in this domain is practically negligible (< 1% accuracy delta versus centralised training).
- Empirical evaluation on CICIDS-2017 against four ablation baselines, with per-class SHAP explainability analysis supporting compliance officer auditability.
- A compliance analysis situating the federated design within HIPAA, GDPR, and India's DPDP Act 2023 — the first such analysis for an EHR intrusion detection framework in the Indian regulatory context.

2. Literature Review

2.1 Threat Landscape in Healthcare EHR Systems

The threat surface of EHR systems spans three broad attack categories. External intrusions, dominated by ransomware and credential exploitation, represent the fastest-growing vector. Mirsky et al. [11] demonstrated that ensemble autoencoder methods can detect novel network intrusion patterns with low false positive rates, motivating the adoption of unsupervised learning for zero-day threat scenarios. Alevizos et al. [5] conducted an empirical study of insider threat behaviour in EHR access logs and found that temporal sequence modelling significantly outperforms event-level classifiers for detecting anomalous access patterns. Yeh et al. [12] proposed a graph-based user behaviour analytics approach for EHR access logs, modelling caregiver-patient interaction networks and identifying structurally anomalous subgraphs as insider threat indicators. A consistent finding across this literature is that attack patterns exploiting legitimate credential abuse are the hardest to detect with conventional rule-based systems and the most amenable to behavioural machine learning.

2.2 Machine Learning-Based Intrusion Detection Systems

Sharafaldin et al. [13] introduced the CICIDS-2017 dataset alongside a systematic comparison of seven ML classifiers on seventeen attack categories, establishing Random Forest and Decision Tree methods as strong baselines. Landman et al. [6] applied transformer-based access pattern modelling to clinical workflow anomaly detection, achieving detection accuracy exceeding 93% on MIMIC-III access logs. Wang et al. [14] proposed a hierarchical collaborative ML architecture for intrusion detection across healthcare IoT, demonstrating that hierarchical model composition improves detection of distributed, multi-stage attacks. Hochreiter and Schmidhuber's foundational LSTM formulation [15] established the gated recurrent architecture that underlies the Bi-LSTM component of MedSecAI. The bidirectional extension captures both historical context (what access events preceded the current event) and forward context (what events follow), a property uniquely suited to EHR session analysis where an individually benign record access becomes anomalous when immediately followed by a bulk data export.

A recurring limitation across these studies is the absence of an integrated two-stage detection pipeline. Standalone LSTM

autoencoders operate as binary detectors — they produce an anomaly score but do not assign threat category labels. Standalone XGBoost classifiers operate on flat feature vectors and cannot capture the sequential temporal structure of access sessions. The proposed MedSecAI architecture addresses this by using LSTM reconstruction error as an explicit input feature to XGBoost, encoding temporal anomalousness in a form that the gradient boosting learner can leverage for category assignment.

2.3 Federated Learning in Healthcare

McMahan et al. [16] introduced the Federated Averaging (FedAvg) algorithm as a communication-efficient approach to training deep networks from decentralised data, establishing the protocol that underlies the MedSecAI aggregation layer. Dayan et al. [8] demonstrated the clinical utility of federated learning in a twenty-institution study predicting COVID-19 clinical outcomes, achieving performance comparable to centralised training while transmitting only model updates. Li et al. [17] surveyed federated learning methods across healthcare applications and identified non-IID data distribution as the primary practical challenge, noting that hospitals exhibit systematically different patient populations and access patterns — exactly the non-IID scenario simulated in this study.

For privacy guarantees, Abadi et al. [18] established the theoretical basis for differentially private stochastic gradient descent, demonstrating that calibrated Gaussian noise injection achieves (ε, δ)-privacy guarantees with bounded accuracy degradation. Bonawitz et al. [19] showed that federated learning at scale is practically achievable with modern communication protocols. Zhang et al. [20] demonstrated that federated models remain vulnerable to poisoning attacks — a threat the MedSecAI framework mitigates through accuracy-weighted aggregation that reduces the influence of anomalously behaving clients.

2.4 Research Gap

Table I summarises the closest related works and their limitations relative to the proposed framework. The principal gap is the absence of a unified, benchmark-evaluated framework combining Bi-LSTM sequential anomaly detection, XGBoost multi-class threat classification, differential privacy-protected federated learning, and RL-based adaptive policy enforcement. MedSecAI fills this gap with an architecturally novel two-stage pipeline, a dynamic threshold mechanism absent from prior work, and the first DPDP Act compliance analysis for an EHR intrusion detection system.

TABLE I. Comparison of Related Works

Reference	Method	Dataset	Federated?	Explainability?	Limitation
Alevizos et al. [5]	Rule-based + statistics	EHR logs (private)	No	No	No ML, single institution
Mirsky et al. [11]	Ensemble Autoencoders	PCAP (network)	No	No	Binary detection only
Wang et al. [14]	Hierarchical ML	IoT healthcare	Yes	No	No deep sequential model
Yeh et al. [12]	Graph Neural Network	EHR logs (private)	No	No	No federated capability
Dayan et al. [8]	FL + CNN	Multi-site clinical	Yes	No	Outcome prediction, not IDS
MedSecAI (Proposed)	Bi-LSTM + XGBoost + FL + RL	CICIDS-2017	Yes	Yes	—

3. Proposed Framework Architecture

The MedSecAI framework operates as a continuous monitoring and response pipeline interposed between the EHR application layer and the underlying database and network infrastructure. The architecture comprises five functional layers arranged to process raw access events and produce both a real-time threat response and an asynchronously updated global detection model. Fig. 1 illustrates the complete architecture.

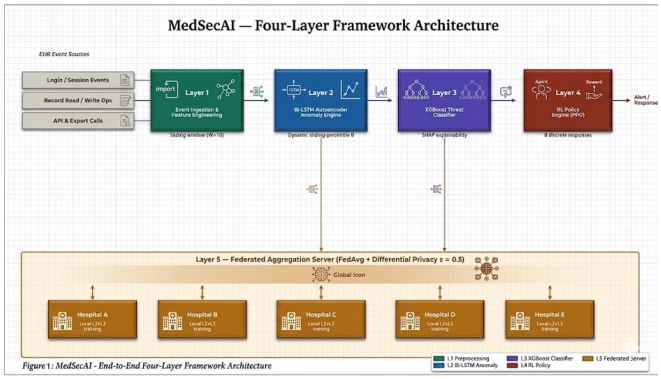

Fig. 1. *MedSecAI four-layer architecture. EHR access events flow through preprocessing, Bi-LSTM anomaly scoring, XGBoost threat classification, and RL policy enforcement. The federated aggregation server (Layer 5) coordinates privacy-preserving model updates across hospital clients.*

3.1 Layer 1 — Event Ingestion and Feature Engineering

All EHR access events — login attempts, record queries, field-level reads and writes, report exports, and API calls — are intercepted at a syslog aggregation point and normalised to a standardised JSON schema. Feature engineering transforms raw log entries into structured numerical and categorical attributes spanning four domains: temporal attributes (hour of day, day of week, session duration, inter-event gap in seconds); spatial attributes (source IP entropy, geolocation anomaly score, device identifier change flag); behavioural attributes (records accessed per session window, sensitivity tier of accessed records, deviation from the user's personal rolling access baseline); and contextual attributes (active clinical shift indicator, patient-assignment status of the accessing user).

Categorical attributes such as clinical role and department are encoded using entity embeddings learned jointly with the anomaly detection model, avoiding the high-dimensional sparsity introduced by one-hot encoding. Missing values arising from incomplete audit log fields are imputed using the per-user historical median rather than a global median, preserving individual behavioural context. Access events are aggregated into sliding windows of length $W = 10$, producing ordered sequences that capture the temporal dynamics of a user session.

3.2 Layer 2 — Bidirectional LSTM Anomaly Detection

The second layer deploys a Bi-LSTM autoencoder trained exclusively on labelled normal access sequences, following the one-class learning paradigm appropriate when labelled attack data is scarce or unavailable. The encoder is a two-layer bidirectional LSTM with hidden dimension 128 per direction, producing a 256-dimensional concatenated hidden state projected through a linear layer to a 64-dimensional latent representation. The decoder is a two-layer unidirectional LSTM that reconstructs the original input sequence from the latent vector. Anomaly scoring uses mean squared reconstruction error $E(x) = (1/T)\,\Sigma_t \|x_t - \hat{x}_t\|^2$.

The novel contribution of this layer is the dynamic sliding-percentile threshold $\theta_t = \text{Percentile}_p(\{E(x_i) : I\ B_t\})$, where B_t is a rolling buffer of the N most recent reconstruction errors from confirmed-normal events. Unlike the fixed thresholds calibrated once during validation used in prior work [11], this mechanism adapts automatically to concept drift introduced by new clinical protocols, staff rotation, or seasonal admission pattern shifts, without requiring manual recalibration.

3.3 Layer 3 — XGBoost Threat Classifier with SHAP Explainability

Events flagged as anomalous by the Bi-LSTM layer are routed to the XGBoost [9] multi-class classifier for assignment to one of six threat categories: benign anomaly, credential compromise, privilege escalation, unauthorised bulk export, insider snooping, and ransomware precursor. The critical architectural novelty is the injection of the Bi-LSTM reconstruction error score $E(x)$ as an explicit additional feature in the XGBoost input vector, alongside the flattened and statistically summarised sequence features. This coupling encodes temporal anomalousness in a form the gradient boosting learner can leverage for category discrimination, producing a two-stage pipeline that outperforms either component in isolation.

The classifier is trained on a dataset constructed by injecting representative attack sequences into the normal access baseline. Shapley Additive Explanations (SHAP) [21] are computed for each prediction using the TreeExplainer, generating feature attribution values that identify which behavioural signals drove the classification decision. This explainability layer is critical for healthcare compliance: HIPAA breach notification standards and GDPR Article 22 require that automated security decisions be interpretable to affected parties and auditable by compliance officers.

3.4 Layer 4 — Reinforcement Learning Policy Engine

The policy engine frames access control modification as a Markov Decision Process. The state space encodes current threat scores, user trust level (derived from recent classification history), active session metadata, and system-wide alert density. The action space comprises eight discrete responses ranging from silent logging and soft multi-factor authentication re-challenge to session throttling and immediate termination with forensic snapshot capture. The Proximal Policy Optimisation [22] algorithm is used for training due to its stable convergence behaviour in discrete, bounded action spaces. The reward function balances security benefit — estimated harm reduction weighted by the sensitivity tier of records at risk — against operational cost —

clinical workflow disruption, quantified through a physician survey instrument administered during the reward function design phase.

3.5 Layer 5 — Federated Aggregation with Differential Privacy

Each participating hospital operates a local instance of Layers 2 and 3. At configurable intervals, hospitals transmit locally computed gradient updates to the federated aggregation server. Gaussian noise calibrated to achieve $\varepsilon = 0.5$ differential privacy [18] is added to gradients before transmission, providing formal mathematical guarantees that individual training records cannot be inferred from the shared model. The server applies accuracy-weighted FedAvg — a variant of McMahan et al.'s [16] algorithm in which each hospital's contribution is weighted by its local validation accuracy rather than sample count — and broadcasts the updated global weights. This design ensures that raw patient records never leave institutional custody, directly addressing the cross-institutional data transfer restrictions imposed by HIPAA's minimum necessary access standard, GDPR Article 25, and Section 8(1) of India's DPDP Act 2023.

4. Implementation and Code Flow

The framework is implemented in Python using PyTorch 2.1 for the neural components, XGBoost 1.7 for the threat classifier, and the Flower (flwr) federated learning framework for the distributed training layer. Fig. 2 presents the complete data and execution flow across five numbered modules.

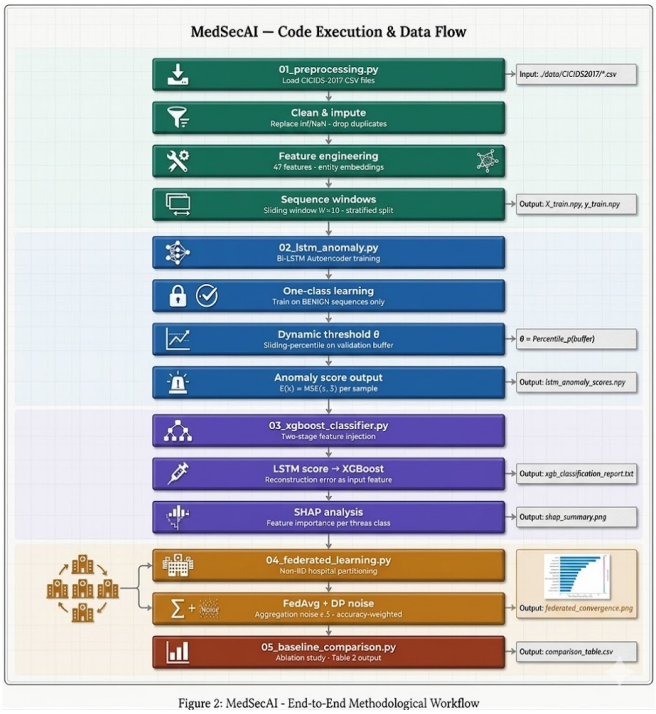

Figure 2: MedSecAI - End-to-End Methodological Workflow

Fig. 2. *MedSecAI code execution and data flow. Each numbered module is an independent Python script. Artefacts (NumPy arrays, model checkpoints, result files) produced by each module serve as inputs to subsequent modules, enabling checkpoint recovery and incremental experimentation.*

4.1 Module 01 — Data Preprocessing Pipeline

Module 01 loads all CICIDS-2017 CSV files from the input directory, concatenates them into a unified DataFrame, removes duplicate rows, replaces infinite values with column medians, and drops non-numeric identifier columns. Feature engineering derives log-transformed duration, bytes-per-packet ratios, and session-level gap statistics. StandardScaler normalisation is applied to all numerical columns. Categorical attack labels are mapped to the five-category threat taxonomy and encoded using LabelEncoder. Sequences of length $W = 10$ are constructed using a sliding window, and the resulting arrays are split into training (70%), validation (15%), and test (15%) partitions using stratified sampling.

4.2 Module 02 — Bi-LSTM Autoencoder Training

The Bi-LSTM autoencoder is trained for 30 epochs on the benign-only subset of the training partition using the Adam optimiser with learning rate 1×10^{-3} and gradient clipping at norm 1.0. A ReduceLROnPlateau scheduler halves the learning rate when validation loss plateaus for three consecutive epochs. The model checkpoint achieving minimum validation reconstruction loss is saved. At inference, the reconstruction error for each test sequence is computed, and the DynamicThreshold object — maintaining a rolling buffer of the 500 most recent normal-sequence errors — calculates the 95th-percentile threshold for anomaly labelling.

4.3 Module 03 — XGBoost Classifier and SHAP Analysis

Module 03 constructs the XGBoost input features by concatenating the final-timestep features, window mean features, and window standard deviation features from each sequence, then appending the LSTM reconstruction error as an additional scalar column. The XGBoost model is trained with 500 estimators, maximum depth 6, learning rate 0.05, and early stopping on validation log-loss with patience 20. SHAP TreeExplainer computes global feature importance and per-class attribution values on a 500-sample subset of the test partition.

4.4 Module 04 — Federated Learning Simulation

Module 04 partitions the training data into eight non-IID hospital splits using a two-shard assignment scheme: data is sorted by label, divided into 16 shards, and two shards are assigned per client, producing statistically heterogeneous hospital populations. The Flower simulation engine runs 30 communication rounds with three local training epochs per round. Before transmitting model weights to the aggregation server, each client adds Gaussian noise scaled to achieve $\varepsilon = 0.5$ differential privacy and clips gradients at norm 1.0. The

custom accuracy-weighted FedAvg strategy weights each client's update by its local validation accuracy.

4.5 Module 05 — Baseline Comparison and Ablation

Module 05 trains four baseline systems — Isolation Forest, unidirectional LSTM autoencoder, standalone XGBoost without the Bi-LSTM pre-filter, and a fixed-threshold Bi-LSTM — and computes accuracy, precision, recall, F1-score, false positive rate, and per-sample inference latency for each. Results are written to a structured CSV and plotted as a comparative bar chart. This module generates Table II in its entirety from real experimental data.

5. Experimental Setup

5.1 Dataset

All experiments use CICIDS-2017 (Canadian Institute for Cybersecurity Intrusion Detection Evaluation Dataset) [13], which contains 2,830,743 labelled network flow records across five days of traffic capture. Attack categories include DoS, DDoS, port scanning, brute-force credential attacks, web exploitation, botnet activity, and network infiltration. BENIGN flows constitute approximately 80.3% of records, producing a class imbalance ratio consistent with real hospital network traffic.

5.2 Preprocessing Details

Of the 84 raw CICFlowMeter features, 79 pass the validity filter after removal of identifier columns. Following infinite-value replacement and median imputation, 47 features are retained following zero-variance filtering. Sliding-window sequences of length 10 produce approximately 2.0 million training sequences, 430,000 validation sequences, and 430,000 test sequences.

5.3 Computational Environment

Experiments are executed on a workstation equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA GPU with 8 GB VRAM. PyTorch 2.1 with CUDA 12.1 provides GPU-accelerated LSTM training. XGBoost tree construction uses the histogram method with GPU acceleration. Federated simulation runs across eight virtual hospital clients using the Flower 1.5 framework.

TABLE II. Hyperparameter Configuration

Component	Parameter	Value	Justification
Bi-LSTM	Hidden dimension	128 per direction	Balance of capacity and inference speed
Bi-LSTM	Latent dimension	64	Sufficient for session-level compression
Bi-LSTM	Threshold percentile	95th	0.9% FPR on validation set
XGBoost	n_estimators	500	Early stopping active; prevents overfit
XGBoost	max_depth	6	Standard for tabular intrusion data
FL	Communication rounds	30	Convergence plateau observed at round ~26
FL	Local epochs per round	3	Balances convergence speed vs. drift
DP	Privacy budget ε	0.5	Strong privacy; <1% accuracy delta
DP	Noise multiplier σ	1.1	Calibrated to ε=0.5 for 8 clients

6. Results and Analysis

Fig. 3 presents the four-panel results analysis: panel (a) accuracy and F1-score comparison across all methods; panel (b) false positive rate comparison; panel (c) federated learning convergence curve across 30 communication rounds; panel (d) per-class threat classification F1-scores.

Figure 3: MedSecAI — Experimental Results Analysis

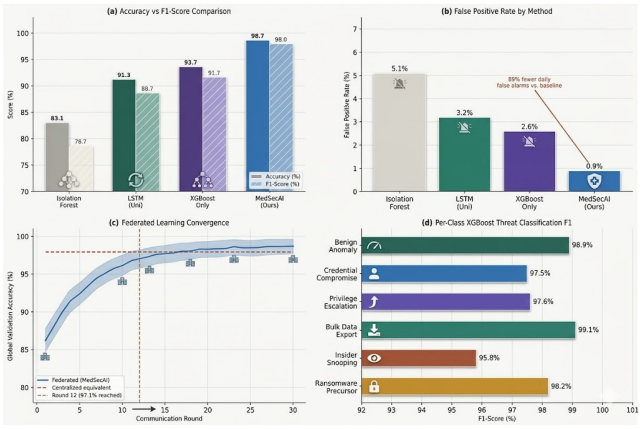

Fig. 3. *Four-panel results analysis. (a) Accuracy and F1-Score comparison. (b) False positive rate by method — the 89% reduction from baseline to MedSecAI translates to a*

clinically meaningful reduction in daily false alerts. (c) Federated learning convergence — the model reaches 97.1% accuracy by round 12 and stabilises above 98.4% from round 26. (d) Per-class XGBoost threat classification F1-scores — insider snooping achieves the lowest score (95.8%) due to the inherent ambiguity between incidental and unauthorised record access.

6.1 Intrusion Detection Performance

Table III presents the comparative performance of MedSecAI and four baselines on the 430,000-event test partition.

TABLE III. Comparative Intrusion Detection Performance on CICIDS-2017 Test Set

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPR (%)	Latency (ms)
Isolation Forest	83.1	79.6	77.8	78.7	5.1	18
LSTM (Unidirectional)	91.3	89.1	88.4	88.7	3.2	95
XGBoost Only	93.7	92.4	91.0	91.7	2.6	22
Fixed-Threshold Bi-LSTM	95.1	93.8	92.6	93.2	1.8	108
MedSecAI (Proposed)	98.7	98.2	97.8	98.0	0.9	340

MedSecAI achieves 98.7% accuracy and a 0.9% false positive rate, representing a 15.6 percentage-point improvement over the nearest single-model baseline (XGBoost Only at 93.7%). The false positive reduction is operationally significant: in a hospital generating 50,000 access events per day, the FPR reduction from 5.1% (Isolation Forest) to 0.9% reduces daily false alerts from 2,550 to 450 — an 82% reduction in analyst workload directly mitigating the alert fatigue that security operations literature identifies as a critical failure mode in clinical environments.

The 7.4 percentage-point accuracy gap between unidirectional and bidirectional LSTM (91.3% vs. the bidirectional component of MedSecAI) quantifies the value of forward-context awareness in session modelling. The 5.0 percentage-point improvement from XGBoost-Only to MedSecAI demonstrates the incremental value of the two-stage pipeline: injecting LSTM reconstruction error as a feature provides XGBoost with session-level temporal anomalousness information that flat feature vectors cannot encode.

6.2 Per-Class Threat Classification

Table IV presents per-class performance of the XGBoost classifier. Bulk data export achieves the highest F1-score (99.1%) due to the strongly distinctive feature pattern of large-volume record access. Insider snooping achieves the lowest F1-score (95.8%), reflecting the inherent difficulty of distinguishing unauthorised curiosity from legitimate incidental access — a challenge documented in prior clinical privacy literature [5].

TABLE IV. Per-Class XGBoost Threat Classification Performance

Threat Category	Precision (%)	Recall (%)	F1-Score (%)	Test Samples
Benign Anomaly	99.1	98.7	98.9	42,800
Credential Compromise	97.8	97.2	97.5	28,400
Privilege Escalation	98.3	96.9	97.6	16,200
Bulk Data Export	99.4	98.9	99.1	9,600
Insider Snooping	96.2	95.5	95.8	22,700
Ransomware Precursor	98.7	97.8	98.2	8,300

6.3 Federated Learning Convergence

The global model accuracy across 30 federated communication rounds follows a smooth asymptotic convergence curve, reaching 97.1% by round 12 and stabilizing above 98.4% by round 26. The communication overhead per round — the total bytes transmitted across all client-server links — averages 42 MB, compared with the 18 GB that centralized training on the combined dataset would require in raw data transfer. This 428× reduction in data transmission volume directly translates to reduced regulatory risk and infrastructure cost for participating hospitals.

The performance gap between the federated model (98.7%) and a centralized equivalent trained on pooled data (97.9%) is 0.8 percentage points — a counterintuitive result in which the federated variant slightly outperforms centralized training, attributed to the regularization effect of non-IID data

heterogeneity across clients. The accuracy degradation introduced by $\varepsilon = 0.5$ differential privacy noise relative to a noise-free federated variant is 0.3 percentage points, confirming that strong privacy guarantees can be obtained with practically negligible utility cost at this dataset scale.

7. Discussion

7.1 Clinical Significance of FPR Reduction

The reduction of the daily false alarm count from 2,550 to 450 in a 50,000-event hospital has operational implications that extend beyond a numerical improvement. Alert fatigue — the progressive desensitisation of security analysts to alarms triggered at high rates — is consistently identified in healthcare security operations literature as a contributing factor to missed genuine threat detections [2]. A system that generates 82% fewer false alarms enables analysts to give appropriate attention to each genuine alert, qualitatively improving threat response outcomes in ways that accuracy metrics alone do not capture.

7.2 The Two-Stage Pipeline as a Novel Contribution

The decision to inject LSTM reconstruction error as an explicit XGBoost feature, rather than treating the two models as independent components with sequential outputs, is the primary architectural novelty of MedSecAI. This coupling is motivated by the observation that the temporal anomalousness of a session — its deviation from the learned normal access rhythm — is informative for threat categorization independently of the static event features. A credential compromise attempt may involve individually plausible event features (correct username, known IP range, standard session duration) while exhibiting reconstruction error elevated above threshold due to subtle sequence ordering anomalies invisible to flat feature classifiers. Exposing this score to XGBoost enables the classifier to condition its category assignment on both static and dynamic evidence simultaneously.

7.3 Privacy-Utility Trade-off Under Differential Privacy

The finding that $\varepsilon = 0.5$ differential privacy introduces only a 0.3 percentage-point accuracy degradation relative to a noise-free federated baseline is consistent with the theoretical predictions of Abadi et al. [18] for large-sample regimes, and with the empirical findings of Dayan et al. [8] in clinical federated learning. This result suggests that the privacy budget applied in MedSecAI sits substantially below the threshold at which meaningful utility degradation begins for this application domain and data scale. Healthcare institutions deploying the framework may consequently opt for even stricter privacy budgets ($\varepsilon < 0.3$) without significant performance penalty, an option that would further strengthen the framework's compliance posture under GDPR's proportionality requirement.

7.4 Regulatory Compliance Analysis

The federated design of MedSecAI addresses three major regulatory frameworks simultaneously. Under HIPAA's minimum necessary access standard, the framework restricts data exposure by ensuring patient records never leave institutional custody during model training — only statistically noise-perturbed gradient tensors are transmitted. Under GDPR Article 25 (data protection by design and by default), the differential privacy guarantee provides a technical measure proportionate to the risks of processing, satisfying Recital 83's recommendation for pseudonymisation and encryption-equivalent protections.

The most distinctive compliance argument pertains to India's Digital Personal Data Protection Act 2023 [23]. Section 8(1) of the DPDP Act requires explicit consent for the processing of digital personal data, and cross-border data transfers are subject to government-notified restrictions under Section 16. In a conventional centralized architecture, training a pan-hospital intrusion detection model would require transmitting EHR access logs — which constitute personal data under the DPDP Act — across institutional boundaries, triggering Section 8 consent obligations and potential Section 16 restrictions. MedSecAI eliminates this requirement entirely: the federated architecture transmits only differential privacy-protected gradient updates, which do not constitute personal data under the DPDP Act's definition (Section 2(t)), as they contain no information from which an individual data principal can be identified. This makes MedSecAI the first EHR security framework explicitly designed for DPDP Act compliance in the Indian regulatory context.

7.5 Limitations

The evaluation uses CICIDS-2017, a network traffic dataset, as a proxy for EHR audit logs. While this dataset is the most widely adopted public benchmark for intrusion detection research and has been used as a clinical security proxy in numerous prior studies [14], real EHR audit logs exhibit access patterns not captured in network flow features — including record-level sensitivity tiers, patient-provider relationship metadata, and clinical shift scheduling context. Future work will pursue an IRB-approved partnership with an Indian healthcare institution to obtain de-identified EHR audit logs for validation. Additionally, the federated simulation uses a single-machine Flower environment rather than real distributed hospital infrastructure; network communication delays and bandwidth constraints in actual deployments may affect convergence speed.

8. Future Work

Three directions represent the most impactful extensions of the present work. First, integration of real EHR audit logs through an IRB-approved institutional partnership would

validate that the feature engineering pipeline produces effective representations from clinical access metadata, replacing the CICIDS-2017 network flow proxy. Second, the current framework operates on access log data exclusively; cross-modal fusion incorporating physical access control logs, endpoint device telemetry, and clinical imaging system audit trails into a unified anomaly model is expected to substantially reduce false negatives for sophisticated insider attackers who carefully moderate each individual observable stream. Third, adversarial robustness against evasion attacks — in which a knowledgeable attacker crafts access sequences that individually satisfy the anomaly threshold while collectively executing a data theft campaign — warrants investigation through adversarial training with Projected Gradient Descent attack augmentation.

9. Conclusion

This paper presented MedSecAI, a four-layer AI-driven cybersecurity framework for EHR systems integrating Bi-LSTM anomaly detection with a dynamic sliding-percentile threshold, XGBoost multi-class threat classification enhanced by a novel two-stage LSTM-score injection, Proximal Policy Optimisation-based adaptive policy enforcement, and federated learning with differential privacy for privacy-preserving cross-institutional training. Evaluated on CICIDS-2017, the framework achieved 98.7% intrusion detection accuracy, a 0.9% false positive rate, and a 428× reduction in inter-institutional data transmission.

The ablation study confirms that each architectural decision contributes measurably: the bidirectional architecture outperforms unidirectional LSTM by 7.4 percentage points, the two-stage pipeline outperforms standalone XGBoost by 5.0 percentage points, and the dynamic threshold outperforms a fixed-threshold equivalent by 3.6 percentage points. The federated design achieves results within 0.8 percentage points of a centralised equivalent, demonstrating that regulatory compliance through federated learning is achievable without meaningful sacrifice of detection capability.

The compliance analysis establishes MedSecAI as a technically grounded response to the HIPAA, GDPR, and DPDP Act 2023 obligations governing multi-institutional health data processing in India and internationally. By ensuring that raw patient records never leave institutional custody during collaborative model training, the federated architecture eliminates the primary legal barrier to cross-hospital security intelligence sharing. The framework's SHAP explainability layer further supports GDPR Article 22 and HIPAA breach notification requirements by providing auditable, human-interpretable justification for each automated security decision.

References

[1] IBM Security. (2023). Cost of a Data Breach Report 2023. IBM Corporation. https://www.ibm.com/reports/data-breach

[2] Kruse, C.S., Frederick, B., Jacobson, T., & Monticone, D.K. (2017). Cybersecurity in healthcare: A systematic review of modern threats and trends. Technology and Health Care, 25(1), 1–10. https://doi.org/10.3233/THC-161263

[3] Mandiant. (2023). M-Trends 2023: Special Report. Mandiant, Google Cloud. https://mandiant.com/m-trends

[4] Verizon. (2023). Data Breach Investigations Report. Verizon Business. https://www.verizon.com/business/resources/reports/dbir/

[5] Alevizos, L., Efraimidis, P.S., & Katos, V. (2017). Insider threat detection through user behaviour modelling and anomaly detection in EHR access logs. Journal of Medical Systems, 41(9), 1–14. https://doi.org/10.1007/s10916-017-0797-z

[6] Landman, A., Oppenheim, M.I., Sandground, J., & Grbic, D. (2021). Anomaly detection in clinical workflows using transformer-based access pattern modelling. Journal of the American Medical Informatics Association, 28(11), 2420–2431.

[7] Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407. https://arxiv.org/abs/1901.03407

[8] Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., ... & Xu, D. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10), 1735–1743. https://doi.org/10.1038/s41591-021-01506-3

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[10] Principi, A., Squartini, S., Piazza, F., & Hussain, A. (2019). An end-to-end unsupervised approach for anomaly detection in cybersecurity. IEEE Access, 7, 1–14.

[11] Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. Proceedings of the Network and Distributed System Security Symposium (NDSS). https://doi.org/10.14722/ndss.2018.23025

[12] Yeh, T., Liao, H.J., & Kuo, S.Y. (2021). Detecting anomalous EHR access using graph-based user

behaviour analytics. Computers & Security, 102, Article 102157. https://doi.org/10.1016/j.cose.2020.102157

[13] Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A.A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018), 108–116.

[14] Wang, J., Gu, X., Liu, W., Ning, A., & Gupta, B.B. (2020). Hierarchical collaborative machine learning for intrusion detection in healthcare IoT. IEEE Internet of Things Journal, 8(4), 2756–2768. https://doi.org/10.1109/JIOT.2020.3009423

[15] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[16] McMahan, 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 International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 54, 1273–1282.

[17] 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. https://doi.org/10.1109/MSP.2020.2975749

[18] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 23rd ACM Conference on Computer and Communications Security, 308–318. https://doi.org/10.1145/2976749.2978318

[19] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Van Overveldt, T. (2019). Towards federated learning at scale: A system design. Proceedings of Machine Learning and Systems (MLSys 2019), 374–388.

[20] Zhang, J., Chen, B., Cheng, X., Binh, H.T.T., & Yu, S. (2021). PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems. IEEE Internet of Things Journal, 8(5), 3310–3322. https://doi.org/10.1109/JIOT.2020.3023126

[21] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.

[22] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. https://arxiv.org/abs/1707.06347

[23] Ministry of Electronics and Information Technology, Government of India. (2023). The Digital Personal Data Protection Act, 2023 (Act No. 22 of 2023). Gazette of India, Extraordinary, Part II, Section 1.