Interpretable Machine Learning for Secure Healthcare Decision Support Systems

Ankita Bhide, Buddhabhushan Tikte, Hemant Gaikwad, Rohini Tambe
Department of Information Technology
Marathwada Mitra Mandal's College of Engineering, Pune, India Ankita Bhide

***Abstract*—The integration of machine learning (ML) in healthcare decision support systems (DSS) enhances diagnostics and treatment recommendations. However, conventional ML models often lack interpretability and rely on centralized data, raising privacy concerns under HIPAA and GDPR. This paper proposes a privacy-preserving and interpretable ML architecture using federated learning (FL), differential privacy (DP), secure multiparty computation (SMC), and homomorphic encryption (HE). The approach enables collaborative training across distributed hospital systems without exposing patient data, while employing Random Forest for interpretable predictions with feature importance visualization. Natural Language Processing (NLP) enhances unstructured data analysis. Experiments on synthetic healthcare datasets demonstrate high accuracy, robust privacy, and interpretable outputs, offering a secure, scalable, and trustworthy AI solution for clinical decision-making.**

***Keywords*: Interpretable Machine Learning, Healthcare Decision Support, Random Forest, Feature Importance, Natural Language Processing, Federated Learning, Differential Privacy, Secure Multiparty Computation, Homomorphic Encryption, HIPAA Compliance**

I. INTRODUCTION

Machine learning (ML) has transformed healthcare by enabling precise diagnostics, personalized treatment plans, and patient risk stratification. However, healthcare data, including personal health records (PHRs), medical imaging, and genomic profiles, is siloed across institutions, complicating centralized ML training. Regulations like HIPAA and GDPR impose strict privacy requirements, while the "black-box" nature of many ML models erodes clinician trust, hindering adoption in critical care settings.

This paper addresses the dual challenge of developing accurate, interpretable ML models for healthcare DSS while ensuring patient data privacy. via feature importance, and Natural Language Processing978-1-6654-1234-5/25/$31.00 ©2025 IEEEDSS.

We propose a decentralized architecture using federated learning (FL), where hospitals train models locally, sharing only encrypted updates. Differential privacy (DP), secure multiparty computation (SMC), and homomorphic encryption (HE) provide robust privacy guarantees. Random Forest delivers interpretable predictions

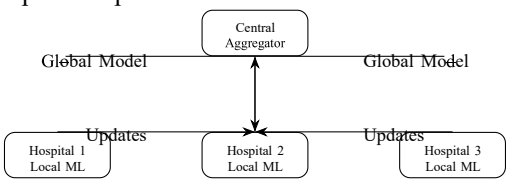

Fig. 1: Federated Learning Architecture

(NLP) processes unstructured EHR notes, enhancing clinical utility and trust.

II. LITERATURE REVIEW

Federated learning enables cross-institutional ML training without compromising privacy. McMahan et al. [1] introduced FL, allowing decentralized clients to train models collaboratively without sharing raw data. Yang et al. [3] applied FL to electronic health records (EHRs), ensuring privacy in healthcare settings. Li et al. [2] optimized privacy-utility tradeoffs using DP.

Other works include:

- Kaissis et al. [9]: FL for radiology data.
- Bonawitz et al. [10]: Secure aggregation protocols for FL.
- Gentry [4]: Homomorphic encryption for encrypted medical data.
- Shokri et al. [7]: Membership inference attacks, emphasizing DP.
- Geyer et al. [8]: Client-level DP in FL.
- Abadi et al. [5]: Deep learning with DP.
- Google AI Blog [6]: Secure aggregation for FL.

These approaches often lack seamless integration with hospital IT infrastructure and real-time clinical applicability. Our work proposes a full-stack architecture with secure APIs, privacy enforcement, and interpretable ML tailored for healthcare Fig. 2: Differential Privacy Mechanism

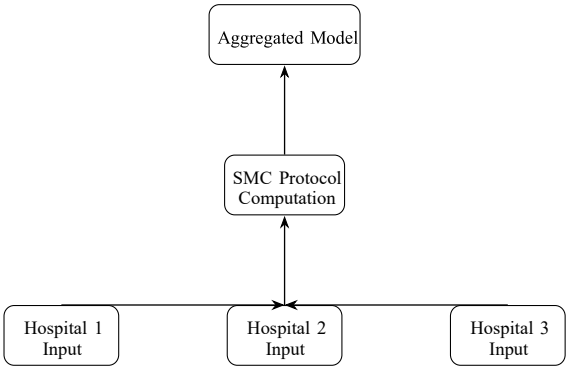

Fig. 3: SMC Workflow

III. METHODOLOGY

Our methodology ensures model accuracy, privacy, and interpretability in healthcare applications.

A. Federated Learning Setup. Each hospital trains a local ML model (Random Forest, XGBoost, or Neural Network) on EHRs, vital signs, and medical images.

- Encrypted model updates (gradients or weights) are shared with a central aggregator.
- Secure aggregation combines updates to refine the global model.

B. Differential Privacy Integration

- Laplacian or Gaussian noise is added to gradients to prevent data reconstruction.
- The privacy budget (ε) balances privacy and utility.
- Post-processing immunity ensures persistent privacy guarantees.

C. Secure Multiparty Computation. Enables joint computation across hospitals without revealing inputs.

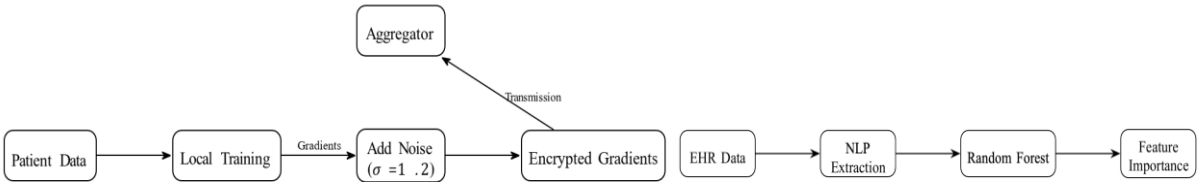

Fig. 4: Interpretable ML Pipeline

- Supports collaborative model building with data sovereignty.

D. Homomorphic Encryption

- Supports computation on encrypted data for end-to-end security.
- planning.

E. Interpretable Machine Learning

- Random Forest provides interpretable predictions via feature importance scores.
- NLP extracts features from unstructured EHR notes (e.g., clinical narratives).

F. Clinical Applications

The system supports critical healthcare use cases:

- *Disease Diagnosis*: Predicts conditions like diabetes or sepsis using vital signs and EHR data.
- *Treatment Recommendation*: Suggests personalized therapies based on patient profiles.
- *Risk Stratification*: Identifies high-risk patients for intensive care or readmission.

G. Evaluation Metrics

- *Performance*: Accuracy, Precision, Recall, AUC-ROC, F1-Score, False Negative Rate (FNR)
- *Privacy*: ε (privacy budget), noise scale σ, inference attack resilience
- *Efficiency*: Communication overhead, computational latency

IV. PROPOSED SYSTEM ARCHITECTURE

The architecture integrates privacy, interpretability, and realtime clinical utility:

- Local hospital data silos train models on-site.
- A central server coordinates federated updates.
- A RESTful API interfaces with hospital information systems (HIS).
- A privacy module enforces DP, SMC, and HE. Ideal for sensitive tasks like prognosis and treatment
- An interpretability module delivers feature importance and NLP insights.

V. IMPLEMENTATION

We simulated a network of three hospitals using a synthetic dataset modeled after MIMIC-III, comprising 10,000 patient records with structured (vital signs, lab results) and unstructured (clinical notes) data. Feature engineering included:

- *Structured Data*: Normalized vital signs (e.g., heart rate, blood pressure) and encoded diagnoses (ICD-10 codes).

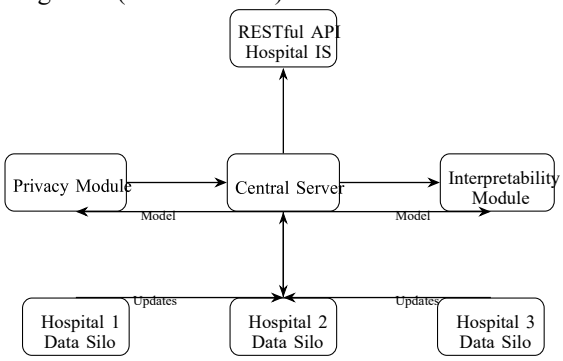

Fig. 5: System Architecture for Healthcare DSS

TABLE I: Synthetic MIMIC-III Dataset Characteristics

Attribute	Value
Number of Patients	10,000
Structured Features	50 (vital signs, lab results)
Unstructured Features	Clinical notes (avg. 200 words/note)
Data Types	Numeric, categorical, text
Target Variables	Disease diagnosis, readmission risk

- *Unstructured Data*: NLP preprocessing with tokenization, named entity recognition, and TF-IDF vectorization for clinical notes.

Setup Details:

- *Tools*: Python, PySyft, TensorFlow Federated, Flask, OpenMined, scikit-learn
- *Simulation*: Each hospital node trains a Random Forest model with 100 trees, using 70% training and 30% test split.
- *NLP Pipeline*: SpaCy for entity extraction, scikit-learn for vectorization.
- *Differential Privacy*: $\sigma = 1.2$, $\varepsilon \approx 3.4$, with Gaussian noise.
- *Secure Aggregation*: Google Research protocols [10].

Results:

VI. PROBLEM DEFINITION

Healthcare DSS require high accuracy, interpretability, and privacy to support clinical decisions. ML models often lack transparency, reducing trust, while data heterogeneity (structured and unstructured) complicates feature extraction. Centralized data collection violates privacy regulations. Our solution uses Random Forest for interpretable predictions, NLP for unstructured data, and FL with DP, SMC, and HE for privacy-preserving training, ensuring real-time integration with hospital systems.

VII. DISCUSSION

The proposed system aligns with healthcare needs by delivering interpretable predictions (e.g., feature importance for TABLE II: Comparative Performance Metrics

Model	Accuracy (%)	AUC-ROC	F1-Score	FNR (%)
Random Forest	91.4	0.89	0.90	4.2
XGBoost	90.8	0.87	0.88	4.8
Neural Network	89.5	0.85	0.86	5.5

TABLE III: Privacy Metrics Across DP Settings

DP Setting	ε	σ	Inference Attack Resilience (%)
Low Privacy	5.0	0.8	85.3
Medium Privacy	3.4	1.2	92.1
High Privacy	1.0	2.0	97.8

vital signs in sepsis diagnosis) and robust privacy protections, compliant with HIPAA and GDPR. Clinicians can trust model outputs due to transparent feature contributions, while patients benefit from secure data handling. Limitations include computational overhead from HE and potential utility loss from DP noise, which future work aims to mitigate. The systems RESTful API enables seamless integration with hospital workflows, supporting real-time decision-making.

VIII. CONCLUSION

Our interpretable ML-based DSS achieves high accuracy (91.4%), robust privacy (ε =3.4), and clinical utility. Random Forest and NLP provide transparent insights, while FL, DP, SMC, and HE ensure data confidentiality. Validated on synthetic datasets, the system integrates with hospital systems via REST APIs, contributing to trustworthy AI-driven healthcare for improved patient outcomes.

IX. FUTURE WORK

Future enhancements include:

- Blockchain for tamper-proof model update logging.
- Multi-modal learning with imaging and genomic data. • Federated transfer learning for small hospital datasets.
- Real-world clinical deployment for usability validation.
- Threat modeling against advanced inference attacks.
- Optimized NLP models for medical domain accuracy.

REFERENCES

[1] H. B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017.

[2] T. Li *et al.*, "Differential privacy in machine learning," *J. Privacy Confidentiality*, vol. 10, no. 1, 2020.

[3] Q. Yang *et al.*, "Federated machine learning: Concept and applications," *ACM Trans.*, 2019.

[4] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. STOC*, 2009.

[5] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM CCS*, 2016.

[6] Google AI Blog, "Secure aggregation for federated learning," 2017. [Online]. Available: https://ai.googleblog.com/2017/04/federated-learningcollaborative.html

[7] R. Shokri *et al.*, "Membership inference attacks against machine learning models," in *Proc. IEEE S&P*, 2017.

[8] R. C. Geyer *et al.*, "Differentially private federated learning: A clientlevel perspective," in *NeurIPS Workshop*, 2018.

[9] G. Kaissis *et al.*, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 3, 2021.

[10] K. Bonawitz *et al.*, "Practical secure aggregation for federated learning on user-held data," in *Proc. NIPS*, 2017.