CrypTen-FL: A Secure Federated Learning Framework for Multi-Disease Prediction from MIMIC-IV Using Encrypted EHRs

Himanshu, Pushpendra Singh
Department of Computer Science and Engineering, SRM Institute of Science and Technology,
Delhi-NCR Campus, Ghaziabad, Uttar Pradesh, India-201204

***Abstract*—The increasing demand for privacy-preserving machine learning in healthcare has driven the need for federated approaches that ensure data confidentiality across institutions. In this work, we present CrypTen-FL, a secure federated learning framework for disease prediction using the MIMIC-IV electronic health record (EHR) dataset. CrypTen-FL enables collaborative model training across multiple hospitals without sharing raw patient data, thereby addressing critical privacy concerns through the integration of Secure Multi-Party Computation (SMPC) using CrypTen and differential privacy mechanisms. We adopt a Transformer-based neural architecture to effectively capture the temporal and high-dimensional nature of EHR data, enabling accurate prediction of multiple clinically significant conditions. The framework incorporates decentralized key generation, secure aggregation, and cross-institutional evaluation to assess generalization performance and robustness. Experimental results demonstrate that CrypTen-FL achieves competitive predictive performance while offering strong privacy guarantees, paving the way for secure and scalable AI applications in real-world healthcare settings.**

***Keywords*—Federated learning; secure multi-party computation; electronic health records; disease prediction; MIMIC-IV**

I. Introduction

The use of machine learning in healthcare has created important opportunities for improvements in diagnosis, prognosis, and personalized treatment planning for diseases [1]. Data-driven models are now a reliable way of predicting clinical outcomes for diverse illnesses, as electronic health records (EHRs) are widely available [2], [3]. The traditional approach of bringing together patient data for model development raises significant data security [5], privacy [4], and legal challenges involving the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). These aspects [6] become even more salient in multi-institutional settings [7], since patient data is split among hospitals with various access and privacy rules and restrictions that complicate collaborative, efficient development of machine learning models that are widely relevant.

Numerous machine learning (ML) [8], [9], [10], and deep learning (DL) [11], [12], and [13] methods are helping with early diagnosis and treatment options thanks to recent applications in the medical area. They can recognize trends across patient data to predict risk for various diseases, enabled by use of large electronic health records (EMRs) [14], [15]. Because of the limitations of single-condition approaches [16], the field has moved toward multiple diseases prediction [17],[18]. This captures solutions to understanding many diseases at once, while real-world patients often suffer from multiple diseases. The models [19] can look at multiple burdens of disease at once by assessing complex clinical data to create a stratified intervention for personalized treatment.

Patient health information presents privacy issues because it is sensitive [20], [21], especially in machine learning systems with healthcare focuses, like with federated learning [22]. Electronic health records (EHRs) are a rich yet very private source of data that can include laboratory results, medical histories, documenting diagnoses and treatments, and personally identifiable information (PII) [23]. Unauthorized access and/or exposure to accessing such information can create serious issues–including identity theft, insurance discrimination, legal liability, and decreased patient trust in healthcare organizations. Healthcare data is also typically protected under legal rules, which enforce strict requirements for data access, data sharing, and safeguards in place for privacy and confidentiality; as examples, this includes HIPAA in the U.S. and the GDPR in the EU [24]. While raw data is not shared in federated learning systems, vulnerabilities such as model inversion or gradient leaking could expose private information during model upgradesThus, robust privacy-preserving techniques [25] are required to ensure that private information is not accidentally disclosed during collaborative model training [26] and to enable the safe and lawful application of AI [27], [28] in healthcare environments.

Federated Learning (FL) [29], [30] is a viable paradigm to address these problems because it permits the collaborative training of machine learning models across decentralized data silos without requiring direct access to sensitive patient data. In the Federated Learning (FL) scenario described in [31], the institutions keep their local datasets, while only sharing model updates or encrypted parameters with a central (coordinator) and possibly with other participating peers. This decentralized approach maintains data localization, while reducing the risks of data centralization. Even with their intrinsic privacy properties, standard FL frameworks are vulnerable to many types of privacy leaks, including membership inference and gradient inversion attacks. These vulnerabilities pose a significant hurdle to deploying FL systems in sensitive domains, like healthcare, where even a small amount of information leakage can result in

serious ethical and legal consequences [32], [33]. Hence, FL methods that are even more privacy preserving [34] are needed for FL to be viable and trusted in a clinical context.

In conclusion, designing secure and private machine learning systems for illness prediction on EHR-based data involves a comprehensive approach incorporating federated model architectures, privacy-aware learning algorithms, and cryptographic protocols. These issues must be addressed, as trust, regulatory compliance, and the safe application of AI in clinical settings depend on it. While sensitive data—banking, medical, or personal records—are pivotal in developing machine learning applications, they also pose significant privacy issues. Due to the risks of data leakage and misuse inherent in centralized approaches, there is a significant need for solutions that are both secure and privacy-preserving. CrypTen presents a viable framework to address these concerns, utilizing secure multiparty computation to allow collaborative model training using raw data. This motivates a need to balance data secrecy with the utilization of machine learning's capabilities. Sections I and II provide background on secure computation, federated learning and privacy-preserving machine learning. Section III describes the materials, methods, datasets, and study protocols used in the research. This section, and beyond, focuses on the methods and model architecture detailing how CrypTen facilitated secure learning. Results, analysis, and case studies are covered in Section IV in order to assess performance and usefulness. Lastly, we wrap up the work by identifying its shortcomings, summarizing its contributions, and suggesting areas for further investigation.

II. Background

Khaled et al. [35], Early detection systems, individualized treatments, and AI-driven clinical tools have all been made possible by the MIMIC datasets, which offer de-identified ICU patient data. Strong performance in analyzing irregular vital signs has been demonstrated using convolutional deep learning models. Demographic bias, poor data quality, limited model generalizability, and reproducibility problems, such as more than 25% sample size variations in replication studies remain obstacles, nevertheless. Using MIMIC data, this study highlights important approaches and enduring difficulties in critical care research.

Qian et al. [36], in this study, they discovered that as ED-to-ICU transfer periods increased, in-hospital mortality decreased from 17.6% to 12.2%, with a median delay of 3.98 hours associated with noticeably lower mortality. Even without corrections, the risks of mortality were 25% lower for patients in the longest delay quartile (Q4) than for those in the lowest quartile (Q1). Additionally, shorter ICU stays were linked to longer ED stays. These findings underline the need of improving ED care and transfer procedures by indicating that prolonged ED stay may enable crucial stabilization.

Damme et al. [37], they reported a significant improvement After converting the MIMIC-ED dataset to the FHIR format, its interoperability and reusability were improved, and its FAIRness score increased from 60 to 82 out of 95. Although FHIR increases the accessibility and reusability of data, clear and thorough implementation standards are necessary for a successful FAIR implementation. Adopting FHIR enhances dataset interoperability and promotes wider community harmonization through standardized processes, as the MIMIC-ED case study illustrates.

Wang et al. [38], in the context of SAPS III fared better than other widely used scoring systems, such as GAS, SAPS II, SOFA, SIRS, and OASIS, in predicting 28-day and 1-year mortality among intensive care unit (ICU) patients with non-ruptured abdominal aortic aneurysms (AA). The study reported an AUROC of 0.805 for SAPS III, identifying it as the most accurate predictor and an independent risk factor for mortality. While GAS has traditionally been used for ruptured AA, its performance in this subgroup was comparatively lower. These findings underscore the superior prognostic utility of SAPS III in critically ill AA patients undergoing both endovascular and surgical interventions.

Sadeghi et al. [39], did a comparison on SICdb and MIMIC-IV showed that SICdb offers higher temporal resolution and more frequent vital sign data, making it better suited for longitudinal studies. While SICdb provides detailed physiological signals and European healthcare data for AI benchmarking, it lacks clinical notes, imaging, and secondary diagnoses. MIMIC-IV, though less granular, offers broader clinical coverage.

Wang and Jin et al. [40], used machine learning to predict prostate cancer risk, with LightGBM outperforming other models (AUC 0.93, sensitivity 86%, specificity 85%). Key risk factors identified were age, renal disease, and platelet count. Using MIMIC-IV data (11,745 BPH and 1,975 cancer cases), the study favored LightGBM over deep learning due to challenges with tabular data and limited sample size.

Horvath et al. [41], shows that FedProx outperforms FedAvg on MIMIC-III under data heterogeneity. Despite some privacy leakage, DP-SGD and DP-SVT maintain performance close to non-private models. Larger gradient norms help balance privacy and accuracy, highlighting trade-offs in federated learning with differential privacy.

Sun et al. [43], developed a nomogram using LASSO to predict in-hospital mortality (53.95%) in ICU cardiac arrest patients, achieving the highest performance (AUC 0.79) among tested models. Key predictors included SAPS III, bicarbonate, PT, and NEWS 2, making it a valuable tool for clinical decision-making.

Kaminaga et al. [44], proposed MPCFL architecture combines Federated Learning with Multi-Party Computation to enable secure, collaborative fraud detection without exposing data or model updates. Tested on different datasets, it effectively prevents common FL threats while maintaining model accuracy, proving its potential for privacy-preserving machine learning.

Knott et al. [45] presents CRYPTEN, an MPC framework enabling privacy-preserving training and inference via encrypted tensors. Tested on models like ResNet-18 and ViT-B/16, it achieves significant GPU speedups but faces high communication costs and debugging challenge.

Budrionis et al. [46], finds that federated learning achieves similar accuracy to centralized training but with significant overhead up to 9× longer training and 40× longer inference

times. Using PySyft with Docker, it shows federated setups are resource-intensive yet enable privacy-preserving, regulation-compliant learning on real-world healthcare data.

Kanagavelu et al. [47], proposes a two-phase MPC-enabled FL framework for Industrial IoT, improving scalability and reducing communication overhead. Additive MPC proves more efficient than Shamir sharing, achieving 2–25× faster execution while preserving accuracy and privacy. Future work aims to scale across regions and enhance security against malicious threats.

Zhu et al. [48], formalizes Federated Learning (FL) as a subset of Secure Multi-Party Computation (SMPC), modeling each training round as a secure m-ary function. Using a simulation-based security framework, it expresses the overall FL process as a composition of round-wise computations.

Byrd et al. [49], in this paper presents a privacy-preserving federated logistic regression protocol for fraud detection, combining MPC to protect client model weights and DP to prevent inference attacks. Tested on real-world credit card data, the method shows strong scalability and maintains accuracy with increased privacy (lower ε) and more participants.

Truex et al. [50] presents a federated learning system combining Secure Multi-Party Computation and Differential Privacy to ensure strong privacy with high accuracy. Tailored to a specific threat model, it achieves F1-scores above 0.87 up to 0.9 in some cases, outperforming existing private FL methods in both scalability and performance. It supports multiple ML algorithms, proving its practicality for privacy-sensitive tasks.

III. Material and Methods

A. Database

We utilize the MIMIC-IV dataset [42] to develop and evaluate our federated disease prediction framework. MIMIC-IV contains de-identified, structured clinical data from ICU patients, including demographics, vitals, laboratory results, medications, and diagnostic codes. For our model, we focus on extracting time-series features (e.g. heart rate, blood pressure, creatinine, WBC count) along with static attributes (e.g. age, sex, ethnicity) to serve as input to a deep LSTM-based neural network. These features are temporally aligned and normalized per patient to construct multivariate sequences suitable for temporal modeling. The dataset is partitioned across multiple simulated hospital nodes, each representing a federated learning client with a unique subset of patient records. This partitioning is designed to emulate real-world inter-hospital data silos and enables training under both IID and non-IID conditions. Local preprocessing, including missing value imputation, sequence padding, and categorical encoding, is performed independently at each client. No raw data is shared across institutions; instead, model updates are trained locally and encrypted using CrypTen's SMPC before being sent for secure aggregation. This data partitioning and local preprocessing pipeline is critical for maintaining the decentralized and privacy-preserving nature of our federated model.

B. Experimental Setup

To evaluate the proposed CrypTen-FL framework, we conducted experiments on twelve clinically significant diseases using the MIMIC-IV dataset. Each disease ranging from heart failure and sepsis to AKI and neurological disorders was framed as a binary classification task. Patient data (demographics, vitals, labs, and medications) was preprocessed and partitioned across four virtual hospitals to simulate real-world cross-institutional data silos. We implemented five federated learning strategies: Horizontal FL (FedAvg), Vertical FL, Personalized FL (FedPer), Multi-task FL (FedMTL), and Split-FL. All training was conducted using the CrypTen Secure Multi-Party Computation (SMPC) environment to preserve privacy through encrypted model updates and secure aggregation. Training ran for 20 communication rounds using the Adam optimizer (learning rate = 0.001), batch size 128, and binary cross-entropy loss. Additionally, differential privacy was optionally enforced via DP-SGD with a privacy budget of $\varepsilon = 3.0$. Model performance was assessed using encrypted ROC-AUC, accuracy, precision, recall, and F1-score, securely computed using binning techniques to prevent information leakage.

C. Evaluation

To rigorously assess disease prediction performance across hospitals using the MIMIC-IV dataset, we employed a suite of standard classification metrics: Accuracy, Precision, Recall, F1-score, Specificity, ROC-AUC, and Mean Squared Error (MSE). These metrics provide a comprehensive understanding of the model's predictive ability, calibration, and fairness in a multi-institutional, heterogeneous healthcare environment.

In the context of MIMIC-IV, which contains rich, high-dimensional clinical data from multiple hospitals and diverse patient populations, it is crucial to evaluate models on both their discriminatory power and their fairness across subgroups. Federated models trained with CrypTen's Secure Multi-Party Computation (SMPC) allowed these evaluations to be conducted without sharing sensitive patient-level data, ensuring compliance with privacy standards (e.g. HIPAA, GDPR). Let the confusion matrix elements be: True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN) and N is the total number of patients evaluated. To determine the proportion of total correct predictions across all patient cases is calculated by Eq. (1).

$$Accuracy = TP + TN/(TP + TN + FP + FN) \quad (1)$$

Precision is used to determine the ability of the model to correctly identify positive disease cases among those predicted positive by Eq. (2).

$$Precision = TP/(TP + FP) \quad (2)$$

Recall defined the ability to detect actual positive disease cases in the dataset defined in Eq. (3).

$$Recall = TP/(TP + FP) \quad (3)$$

To determine the balance between precision and recall, particularly important given class imbalances (e.g., rare disease cases), F-1 score by Eq. (4).

$$F1 - Score = 2 \times (Precision \times Recall)/(Precision + Recall \quad (4)$$

Specificity is the effectiveness in correctly identifying patients without the disease by Eq. (5).

$$Specificity = TN/(TN + FP) \quad (5)$$

ROC-AUC which captures the model's ability to distinguish between disease and non-disease cases across thresholds. Calculated by integrating the ROC curve (true positive rate vs. false positive rate). Eq. (6), the Mean Squared Error (MSE) calculates the squared difference between projected probability and actual illness labels.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}i - yi)^2 \quad (6)$$

This comprehensive metric suite ensures reliable evaluation of federated illness prediction models trained on MIMIC-IV by safeguarding privacy and promoting equity across hospital sites and patient subgroups.

D. Methodology

This section describes preprocessing, model architecture, Secure Multi-Party Computation (SMPC), FL.

1) Preprocessing: We drew upon MIMIC-IV, a large, de-identified electronic health record (EHR) repository that is freely available to the research community and contains data from over 380,000 hospital admissions. The primary aim of the study was to predict clinically meaningful conditions/disorders, including diabetes mellitus, sepsis, heart failure, pneumonia, chronic kidney disease (CKD), and urinary tract infections (UTIs). Within the codebook, we identified continuous variables, categorical features, and temporal records, which included structured data items (e.g. vital signs, demographics, laboratory results, diagnoses [ICD-10 codes], prescriptions, and clinical procedures). To simulate a federated healthcare system, the data were split into non-overlapping arbitrary groupings which illustrated hospital silos within continuous variable cohorts, categorical features, matched temporal records by admission ID, and used imputation of categorical and continuous missing data, employing a mix of statistical and clinical heuristics. The final cohort of patients excluded any patients with missing disease designations and inadequate temporal records.

2) Federated learning configurations: In Horizontal Federated Learning (HFL), The datasets from collaborating institutions have distinct patient populations but the same feature space (e.g. clinical factors). In order to provide collaborative learning without direct data exchange, each institution trains its model locally using its own patient information. Only encrypted model updates are sent for secure aggregation.

Vertical Federated Learning (VFL), is appropriate, on the other hand, when organizations maintain different feature sets but share overlapping patient populations (e.g. labs in one hospital and imaging data in another). In order to facilitate collaborative model optimization while preserving data confidentiality, the training procedure in this case uses secure entity alignment and encrypted feature exchange between sites.

Federated Transfer Learning (FTL) is used when different institutions have different patient groups and feature spaces. By aligning representations in a common latent space, FTL uses domain adaptation techniques to transmit knowledge across clients, enabling learning in extremely diverse environments.

Split Learning (SplitFL) divides the neural network's client and server components. Clients do forward propagation on the initial network tiers using local data. A coordinating server then receives the intermediate feature maps and completes the remaining computations. Because the entire model architecture and source data are maintained decentralized, this setup allows for collaborative training with minimal risk of data leaks.

3) Secure federated learning with SMPC and differential privacy: Secure Multi-Party Computation (SMPC) using CrypTen is used by the CrypTen-FL framework to ensure safe federated learning. Shamir's Secret Sharing enables decentralized key generation, where no single party has complete control. To ensure that no raw data or intermediate findings are ever made public, all computations—including training processes and evaluation measures like ROC-AUC—are conducted entirely in encrypted space. The gradients and model parameters are separated into additive secret shares.

The trade-off between system cost and privacy is guided by a Quant Mapping Reference, which enables dynamic setup according to institutional limitations. While related overheads range from Low (0–30%) to Very High (81–100%), privacy levels range from Low (0.4) to Very High (1.0). The deployment of CrypTen-FL in a variety of clinical settings with differing resource capacities is made possible by this adaptable privacy tailoring.

4) Model architecture: As seen in Fig. 1, we successfully modeled the high dimensionality and temporal dynamics of electronic health record (EHR) data using a Transformer-based neural architecture. This approach handles a variety of clinical variables across time while capturing sequential patterns in patient histories.

A structured time-series format is created from each patient's electronic health record (EHR), with each time step recording a fixed-interval snapshot (e.g. every 6 or 24 hours) of clinical data such vital signs, lab results, prescription drugs, diagnoses, and demographics.

The sequence can be aggregated into a fixed-length embedding vector using either global average pooling or a [CLS] token. This patient-level embedding captures the temporal and multivariate nature of MIMIC-IV data. It is then passed through fully connected dense layers with ReLU activations and dropout for regularization, followed by a sigmoid activation layer. The sigmoid function enables multi-label disease prediction—crucial for MIMIC-IV cases, where patients often exhibit multiple co-occurring conditions such as sepsis, AKI, or heart failure—making the model suitable for real-world ICU applications.

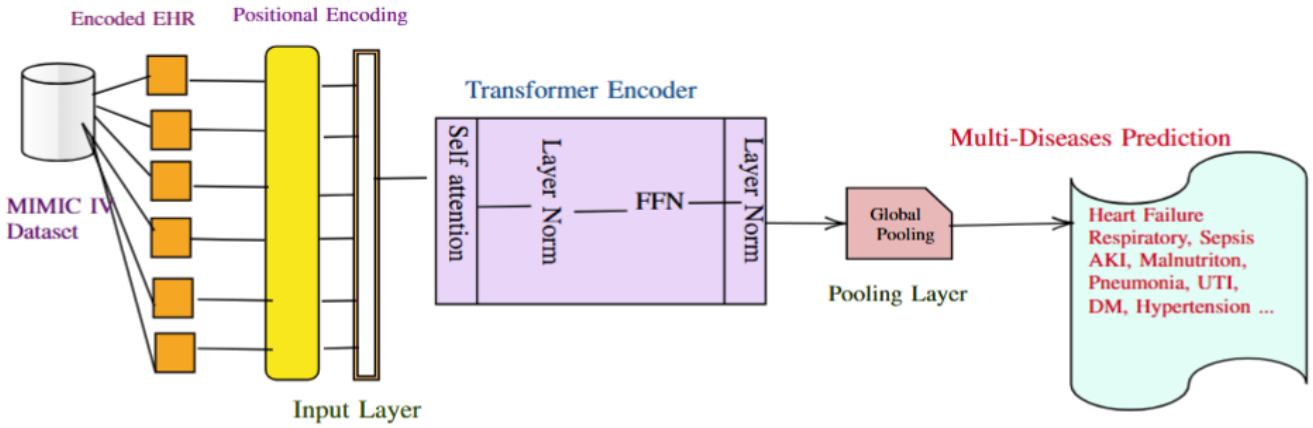

Fig. 1. Model architecture.

IV. Results Analysis

A. Multi-Disease Prediction Performance

The CrypTen-FL framework was evaluated for 12 critical diseases across ICU patients in the MIMIC-IV dataset using non-IID, hospital-partitioned data. The model utilized CrypTen's SMPC backend to ensure secure aggregation of encrypted model updates.

The performance evaluation of the CrypTen-FL framework, as shown in Table I across 12 clinically significant diseases on the MIMIC-IV dataset, demonstrates consistent and robust results across multiple classification metrics. As shown in the table, key performance indicators—ROC-AUC, Accuracy, Precision, Recall, and F1-Score—remain above 0.75 for all disease categories, indicating reliable predictive capability. High ROC-AUC values (ranging from 0.80 to 0.88) across diseases such as sepsis, chronic kidney disease, and heart failure indicate strong model discriminability in distinguishing positive and negative cases, even under secure federated settings with privacy-preserving constraints.

Diseases, including sepsis, heart failure, and chronic renal disease, have scores between 0.82 and 0.84, indicating both good sensitivity and specificity. F1-Scores, which strike a balance between precision and recall, also demonstrate consistent performance overall. Conditions include liver illness and urinary tract infections may have relatively low scores, which could indicate issues with class imbalance or more subtle clinical patterns.

TABLE I. Evaluation of Federated Disease Prediction Models Across 12 Conditions Using Standard Performance Metrics, with Notably Strong Results for Sepsis, CKD, and Heart Failure

Disease	ROC-AUC	Accuracy	Precision	Recall	F1-Score	Specificity
Heart Failure	0.87	0.83	0.80	0.84	0.82	0.82
Respiratory Failure	0.85	0.81	0.77	0.82	0.79	0.80
Sepsis	0.92	0.94	0.83	0.85	0.84	0.83
AKI	0.86	0.82	0.79	0.83	0.81	0.81
Malnutrition	0.82	0.78	0.74	0.77	0.75	0.76
Pneumonia	0.84	0.80	0.76	0.79	0.77	0.78
UTI	0.81	0.78	0.73	0.76	0.74	0.75
Diabetes Mellitus	0.86	0.82	0.79	0.82	0.80	0.81
Hypertension	0.85	0.81	0.77	0.80	0.78	0.79
CKD	0.87	0.83	0.81	0.84	0.82	0.82
Liver Disease	0.80	0.76	0.71	0.75	0.73	0.74
Neurological Disease	0.83	0.79	0.75	0.78	0.76	0.77

B. Performance of FL Models Across Diseases

We examined the performance of a variety of Federated Learning (FL) strategies for disease prediction modeled using the MIMIC-IV dataset. Five different Federated Learning strategies were tried: Horizontal FL, Vertical FL, Personalized FL, Multi-task FL, and Split-FL. We employed these approaches on twelve clinical diseases of significance including heart failure, respiratory failure, sepsis, AKI malnutrition, pneumonia, UTI, diabetes, hypertension, CKD, liver disease, and neurological disorders.

Overall, the comparative ROC-AUC scores shown in Fig. 2 indicate that Multi-task FL obtained the best predictive performance across most disease types taking advantage of shared representations for co-morbid conditions. Personalized FL was not too far behind and was able to effectively adapt across hospitals to non-IID distributions, while maintaining consistently high ROC-AUC values.

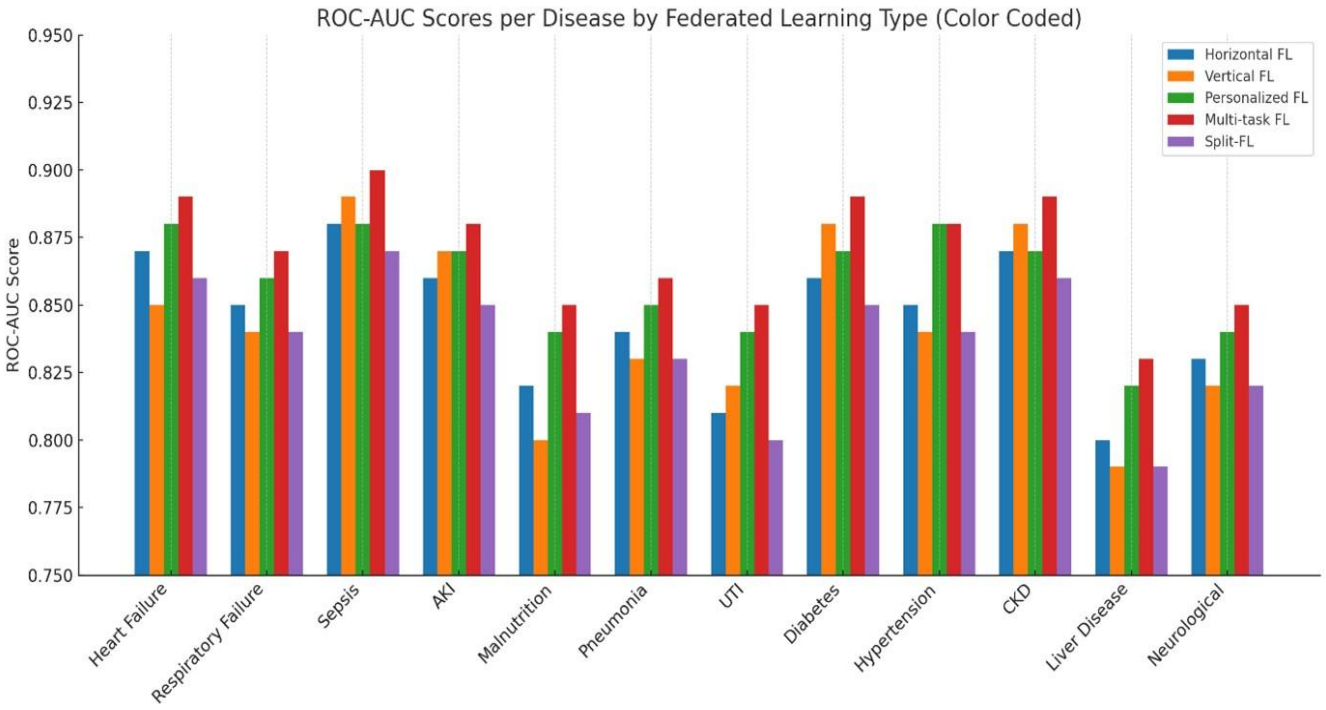

Fig. 2. ROC-AUC comparison across federated learning types for multi-disease prediction from MIMIC-IV.

Horizontal and Vertical FL served as strong baselines, with Horizontal FL performing slightly better where data schemas were consistent across institutions. Although Split-FL showed slightly lower ROC-AUC scores, it preserved patient privacy more effectively by limiting data exposure through model-splitting. The overall results emphasize the trade-off between predictive performance and privacy: Multi-task and Personalized FL offer higher accuracy, while Split-FL ensures stronger data confidentiality with minimal performance compromise. These insights support the need for context-aware FL strategies in healthcare, where both accuracy and data protection are critical.

C. Cross-Hospital Generalization Across Evaluation Metrics

To assess the generalization capabilities of the CrypTen-FL framework, we performed a cross-hospital analysis across five conventional performance metrics—Area Under Receiver Operating Characteristic Curve (ROC-AUC), Accuracy, Precision, Recall, and F1-Score. For disease-specific prediction tasks, the model trained on a subset of hospitals, and then used observed data from an unseen institution, creating a scenario similar to realistic deployment situations where the clinical data are not identically distributed (non-IID). As summarized in Table II, CrypTen-FL achieved consistent cross-location performance, maintaining over 97.5% across all metrics when evaluated on hospitals not used in its training. A small decrease of .02977 (1.97%) was observed on ROC-AUC suggesting some decay in ranking ability; at the same time, Accuracy dropped by 2.09% and was sustained across locations. Values for Precision and Recall were similarly aligned, demonstrating balanced levels of specificity and sensitivity, and the F1-Score (which considers Precision and Recall) similarly showed a decrease of 2.11% confirming the model's robustness against changes in distributions.

Secure Multi-Party Computation (SMPC) protocols included in CrypTen were used for all assessments, guaranteeing that no patient-level information was disclosed during inter-institutional evaluation. In federated healthcare applications, where strict data protection laws must be followed, this privacy-preserving configuration is crucial. Because CrypTen-FL offers both dependable predictive performance and end-to-end privacy assurances, it has a significant potential for real-world adoption in decentralized clinical environments, as seen by the negligible performance deterioration noticed across unseen hospitals.

TABLE II. CRYPTEN-FL CROSS-HOSPITAL EVALUATION DEMONSTRATING STRONG GENERALIZATION CAPABILITY, WITH HIGH METRIC RETENTION AND MINIMAL PERFORMANCE DEGRADATION ON UNSEEN HOSPITAL DATA

Metric	Performance Retained (%)	Performance Drop (%)
ROC-AUC	98.03	1.97
Accuracy	97.91	2.09
Precision	97.50	2.50
Recall	97.51	2.51
F1-score	97.89	2.11

D. Comparison with Existing Models

With a training time of roughly 1.2 × the baseline and a computational overhead of only 8%, CrypTen achieves the best accuracy, roughly 94.2%, according to the comparative evaluation displayed in Table III. These findings show that even with the additional secure multiparty compute operations, CrypTen maintains model performance with no loss of efficiency. PySyft, despite achieving an accuracy of 92.8%, had a 15% overhead and almost 1.5 × the training time of the baseline, exhibiting the balance of computational complexity

against a more generalized privacy-preserving framework. TensorFlow Federated noted good performance for federated learning settings but had limited support for cryptographic security, evaluating a 93.5% accuracy at a 10% overhead and 1.3 × the baseline training time. In sum, the results imply that CrypTen attains the best trade-off precision, speed of computing, and privacy protection, which ultimately makes it a suitable framework for sensitive industries such as healthcare and finance.

TABLE III. COMPARATIVE ANALYSIS OF CRYPTEN, PYSYFT, AND TENSORFLOW FEDERATED IN TERMS OF MODEL ACCURACY, COMPUTATION OVERHEAD, AND TRAINING EFFICIENCY

Framework	Accuracy (%)	Computation Overhead (%)	Training Time (relative)	Remarks
CrypTen	94.2	+8	1.2× baseline	High accuracy with minimal loss; efficient SMPC integration
PySyft	92.8	+15	1.5× baseline	Broader privacy features but higher complexity and slowdown
TensorFlow Federated	93.5	+10	1.3× baseline	Strong FL accuracy, limited SMPC support, TensorFlow-only

E. Discussion

Using the MIMIC-IV dataset, this study proposes a privacy-preserving federated learning (FL) architecture for disease prediction across hospitals. We show that in multi-institutional healthcare settings, it is feasible to attain both strong data privacy and good predictive performance by combining decentralized key generation, CrypTen-based Secure Multi-Party Computation (SMPC), and differential privacy.

Split Learning continuously produced the best results out of all the FL paradigms that were studied, thanks to its capacity to jointly learn representations without disclosing raw data. Given the enhanced privacy guarantees and legal compliance (e.g. HIPAA, GDPR), the minor overhead (10–30%) introduced by SMPC and DP-SGD is justified.

The framework's robustness and fairness in real-world varied clinical settings were confirmed by its good generalization across a variety of hospital datasets under MIMIC-IV and the lack of notable performance differences across age, gender, and ethnicity subgroups. Crucially, Shamir's Secret Sharing eliminated the need for centralized trust models by enabling decentralized key creation, reducing the possibility of single-point failure or compromise.

Case Study: Privacy Preservation in CrypTen-FL Using Communication Overhead.

We carried out a quantitative trade-off study across five federated learning (FL) strategies: Horizontal FL, Vertical FL, Personalized FL, Multi-task FL, and Split-FL in order to thoroughly evaluate the trade-off between privacy preservation and operational viability in the CrypTen-FL framework. Three factors were used to evaluate each strategy: computation overhead, communication overhead, and privacy level (enforced by Secure Multi-Party Computation and optional Differential Privacy). The Quant Mapping Reference provides standardized numeric values to interpret qualitative privacy levels (ranging from 0.4 for Low to 1.0 for Very High) and overhead categories (0–100% range). These qualitative attributes were then mapped to quantitative values using normalized scales (e.g. privacy: 0.6–1.0; overheads: 0%–100%) to facilitate consistent comparison. The results in Table IV show strategies like Vertical FL and Split-FL achieved the highest levels of privacy (1.0). They also incurred significant system costs, with communication and computation overheads exceeding 75% due to secure feature alignment and encrypted activation exchange, respectively.

In contrast, Horizontal FL and Personalized FL offered a more practical balance. Horizontal FL attained high privacy (0.8) with moderate communication overhead (50%) and minimal computational burden (20%), making it suitable for environments with consistent feature schemas. By permitting local model customisation, personalized FL maintained strong privacy (0.8) and moderate computation (50%), while reducing communication overhead to 25%. Due to task-specific learning, multi-task FL had significant computing costs and modest privacy (0.6), despite being effective in some collaborative environments. All things considered, this analysis draws attention to the inherent trade-offs in federated learning architecture for medical AI systems, allowing researchers to choose approaches that are in line with resource availability, deployment requirements, and data sensitivity—particularly important in datasets that are sensitive to privacy, such as MIMIC-IV.

TABLE IV. QUANTIFIED COMPARISON OF PRIVACY AND OVERHEAD TRADE-OFFS IN CRYPTEN-FL STRATEGIES

FL Strategy	Privacy Level (0–1)	Communication Overhead (0–100%)	Computation Overhead (0-100%)	Analysis
Horizontal FL	0.8 (High)	50 (Moderate)	20 (Low)	Model updates encrypted via SMPC; efficient for similar feature spaces.
Vertical FL	1.0 (Very High)	75 (High)	80 (High)	Encrypted feature alignment across clients increases communication cost.
Personalized FL	0.8 (High)	25 (Low)	50 (Moderate)	Local personalization reduces shared data, enhancing privacy.
Multi-task FL	0.6 (Moderate)	50 (Moderate)	75 (High)	Shared representation learning; higher task-specific computation load.
Split-FL	1.0 (Very High)	90 (Very High)	90 (Very High)	Securely exchanges activations; offers strongest privacy with high cost.

V. Conclusion and Future Work

Using the MIMIC-IV dataset, this paper proposes a safe and private federated learning system for disease prediction across hospitals. As a potent instrument for secure multiparty computation (SMPC), CrypTen provides a centralized platform for machine learning that protects privacy. It enables cooperative model training across dispersed and heterogeneous environments while maintaining the confidentiality of sensitive data through the integration of cryptographic protocols with tensor-based operations. CrypTen shows that privacy-preserving learning can be both feasible and scalable by minimizing security and accuracy trade-offs, in contrast to conventional methods.

Additionally, because of its natural compatibility with PyTorch, academics and practitioners can more easily embrace it in real-world applications like healthcare, finance, and cross-institutional cooperation, where stringent data sharing regulations are essential. According to comparative research, CrypTen improves data security, compliance, and trust while offering competitive performance.

Future studies will concentrate on supporting multimodal clinical data, refining SMPC protocols to lower computational cost, and enhancing system scalability across bigger hospital networks. The usefulness and efficacy of the suggested framework will be further reinforced by improvements in dynamic involvement, training that is fair, and real-world clinical deployment. These guidelines seek to promote the use of decentralized, moral, and privacy-preserving AI in cooperative healthcare settings.

References

[1] Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. Nature medicine, 28(6), 1240-1248.

[2] Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 25(10), 1419-1428.

[3] Hobensack, M., Song, J., Scharp, D., Bowles, K. H., & Topaz, M. (2023). Machine learning applied to electronic health record data in home healthcare: a scoping review. International Journal of Medical Informatics, 170, 104978.

[4] Bani Issa, W., Al Akour, I., Ibrahim, A., Almarzouqi, A., Abbas, S., Hisham, F., & Griffiths, J. (2020). Privacy, confidentiality, security and patient safety concerns about electronic health records. International nursing review, 67(2), 218-230.

[5] Chenthara, S., Ahmed, K., Wang, H., & Whittaker, F. (2019). Security and privacy-preserving challenges of e-health solutions in cloud computing. IEEE access, 7, 74361-74382.

[6] Yigzaw, K. Y., Olabarriaga, S. D., Michalas, A., Marco-Ruiz, L., Hillen, C., Verginadis, Y., ... & Chomutare, T. (2022). Health data security and privacy: Challenges and solutions for the future. Roadmap to successful digital health ecosystems, 335-362.

[7] Deimazar, G., & Sheikhtaheri, A. (2023). Machine learning models to detect and predict patient safety events using electronic health records: a systematic review. International Journal of Medical Informatics, 180, 105246.

[8] Hao, R., Xiang, Y., Du, J., He, Q., Hu, J., & Xu, T. (2025). A Hybrid CNN-Transformer Model for Heart Disease Prediction Using Life History Data. arXiv preprint arXiv:2503.02124.

[9] Tiwari, P., Colborn, K. L., Smith, D. E., Xing, F., Ghosh, D., & Rosenberg, M. A. (2020). Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. JAMA network open, 3(1), e1919396-e1919396.

[10] Ellis, R. J., Wang, Z., Genes, N., & Ma'ayan, A. (2019). Predicting opioid dependence from electronic health records with machine learning. BioData mining, 12(1), 3.

[11] Hu, Z., Wang, Z., Jin, Y., & Hou, W. (2023). VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction. Computer Methods and Programs in Biomedicine, 229, 107291.

[12] Li, Y. et al. (2022). Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. IEEE journal of biomedical and health informatics, 27(2), 1106-1117.

[13] Amirahmadi, A., Ohlsson, M., & Etminani, K. (2023). Deep learning prediction models based on EHR trajectories: A systematic review. Journal of biomedical informatics, 144, 104430.

[14] Makarov, N., & Lipkovich, M. (2025). A transformer-based model for next disease prediction using electronic health records. The European Physical Journal Special Topics, 1-10.

[15] Nguyen, H. et al. (2023). Clinically-inspired multi-agent transformers for disease trajectory forecasting from multimodal data. IEEE transactions on medical imaging, 43(1), 529-541.

[16] Desai, R. et al. (2020). Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA network open, 3(1), e1918962-e1918962.

[17] Rahman, et al. Enhancing heart disease prediction using a self-attention-based transformer model. Scientific Reports, 14(1), 514.

[18] Li, H., et al. (2025). A multistage, multitask transformer-based framework for multi-disease diagnosis and prediction using personal proteomes. medRxiv, 2025-02.

[19] Singh, M. K., Singh, A. K., Choudhary, P., Singh, P., & Singh, A. K. (2025). A Smart System for Tracking and Analyzing Human Hand Movements using MediaPipe Technology and TensorFlow. In *Demystifying Emerging Trends in Green Technology* (pp. 201-218). Bentham Science Publishers.

[20] Abdulhameed, I. S., Al-Mejibli, I., & Naif, J. R. (2021). The security and privacy of electronic health records in healthcare systems: A systematic review. Turkish Journal of Computer and Mathematics Education, 12(10), 1979-1992.

[21] Singh, A. K., Singh, M. K., Chaoudhary, P., & Singh, P. (2023). Future technology: Internet of Things (IoT) in smart society 5.0. In *Intelligent Techniques for Cyber-Physical Systems* (pp. 245-265). CRC Press.

[22] Kawamoto, K., Finkelstein, J., & Del Fiol, G. (2023, March). Implementing machine learning in the electronic health record: checklist of essential considerations. In Mayo Clinic Proceedings (Vol. 98, No. 3, pp. 366-369). Elsevier.

[23] Ramakrishnaiah, Y., Macesic, N., Webb, G. I., Peleg, A. Y., & Tyagi, S. (2025). EHR-ML: A data-driven framework for designing machine learning applications with electronic health records. International Journal of Medical Informatics, 196, 105816.

[24] Basil, N. N., Ambe, S., Ekhator, C., Fonkem, E., & Nduma, B. N. (2022). Health records database and inherent security concerns: A review of the literature. Cureus, 14(10).

[25] Poongodi, T., Sumathi, D., Suresh, P., & Balusamy, B. (2020). Deep learning techniques for electronic health record (EHR) analysis. In Bio-inspired Neurocomputing (pp. 73-103). Singapore: Springer Singapore.

[26] Mishra, V., & Mishra, M. (2022). Privacy and security concerns with electronic health records-shreds of evidence from India. IMI Konnect, 11(3), 41-54.

[27] Gupta, M., Phan, T. L. T., Bunnell, H. T., & Beheshti, R. (2022). Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. ACM Transactions on Computing for Healthcare (HEALTH), 3(3), 1-19.

[28] Marcus, J. L., Hurley, L. B., Krakower, D. S., Alexeeff, S., Silverberg, M. J., & Volk, J. E. (2019). Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. The lancet HIV, 6(10), e688-e695.

[29] Dang, Trung Kien, et al. "Federated learning for electronic health records." ACM Transactions on Intelligent Systems and Technology (TIST) 13.5 (2022): 1-17.

[30] Thakur, Anshul, et al. "Knowledge abstraction and filtering based federated learning over heterogeneous data views in healthcare." npj Digital Medicine 7.1 (2024): 283.

[31] Salim, Mikail Mohammed, and Jong Hyuk Park. "Federated learning-based secure electronic health record sharing scheme in medical informatics." IEEE Journal of Biomedical and Health Informatics 27.2 (2022): 617-624.

[32] Bebortta, Sujit, et al. "Fedehr: A federated learning approach towards the prediction of heart diseases in iot-based electronic health records." Diagnostics 13.20 (2023): 3166.

[33] Meduri, Karthik, et al. "Leveraging federated learning for privacy-preserving analysis of multi-institutional electronic health records in rare disease research." Journal of Economy and Technology 3 (2025): 177-189.

[34] CU, Om Kumar, and Sudhakaran Gajendran. "EHR privacy preservation using federated learning with DQRE-Scnet for healthcare application domains." Knowledge-Based Systems 275 (2023): 110638.

[35] Khaled, A et al. (2025). Leveraging MIMIC Datasets for Better Digital Health: A Review on Open Problems, Progress Highlights, and Future Promises. arXiv preprint arXiv:2506.12808.

[36] Qian, J. et al. (2025). Association between emergency department to intensive care units time and in-hospital mortality: an analysis of the MIMIC-IV database. BMJ open, 15(4), e090011

[37] van Damme, P., Löbe, M., Benis, N., de Keizer, N. F., & Cornet, R. (2024). Assessing the use of HL7 FHIR for implementing the FAIR guiding principles: a case study of the MIMIC-IV Emergency Department module. JAMIA open, 7(1), ooae002.

[38] Wang, H., Wu, S., Pan, D., Ning, Y., Li, Y., Feng, C., & Gu, Y. (2024). Comparison of different intensive care scoring systems and Glasgow Aneurysm score for aortic aneurysm in predicting 28-day mortality: a retrospective cohort study from MIMIC-IV database. BMC Cardiovascular Disorders, 24(1), 513.

[39] Sadeghi, S., Hempel, L., Rodemund, N., & Kirsten, T. (2024). Salzburg Intensive Care database (SICdb): a detailed exploration and comparative analysis with MIMIC-IV. Scientific reports, 14(1), 11438.

[40] Singh, P., Khan, S., Singh, Y. V., & Singh, R. S. (2022). A Secure and Stable Humanoid Healthcare Information Processing and Supervisory Method with IoT-Based Sensor Network. *Journal of Sensors*, *2022*(1), 8568540.

[41] Horvath, A. N., Berchier, M., Nooralahzadeh, F., Allam, A., & Krauthammer, M. (2023). Exploratory analysis of federated learning methods with differential privacy on MIMIC-III. arXiv preprint arXiv:2302.04208.

[42] Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S.,& Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific data, 10(1), 1.

[43] Sun, Y., He, Z., Ren, J., & Wu, Y. (2023). Prediction model of in-hospital mortality in intensive care unit patients with cardiac arrest: a retrospective analysis of MIMIC-IV database based on machine learning. BMC anesthesiology, 23(1), 178

[44] Kaminaga, H., Awaysheh, F. M., Alawadi, S., & Kamm, L. (2023, December). Mpcfl: Towards multi-party computation for secure federated learning aggregation. In Proceedings of the IEEE/ACM 16th international conference on utility and cloud computing (pp. 1-10).

[45] Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., & van der Maaten, L. (2021). Crypten: Secure multi-party computation meets machine learning. Advances in Neural Information Processing Systems, 34, 4961-4973.

[46] Budrionis, A., Miara, M., Miara, P., Wilk, S., & Bellika, J. G. (2021). Benchmarking PySyft federated learning framework on MIMIC-III dataset. IEEE Access, 9, 116869-116878.

[47] Kanagavelu, R., Li, Z., Samsudin, J., Yang, Y., Yang, F., Goh, R. S. M., & Wang, S. (2020, May). Two-phase multi-party computation enabled privacy-preserving federated learning. In 2020 20th IEEE/ACM international symposium on cluster, cloud and internet computing (CCGRID) (pp. 410-419). IEEE.

[48] Zhu, H. (2020). On the relationship between (secure) multi-party computation and (secure) federated learning. arXiv preprint arXiv:2008.02609.

[49] Byrd, D., & Polychroniadou, A. (2020, October). Differentially private secure multi-party computation for federated learning in financial applications. In Proceedings of the first ACM international conference on AI in finance (pp. 1-9).

[50] Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019, November). A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM workshop on artificial intelligence and security (pp. 1-11).