International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 5, Issue 4, 99-108, 2024
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V5I4P110

Original Article

Exploration of Federated Multi-Task Learning Models for Secure Cross-Institutional Credit Risk Assessment under Privacy Constraints

Santhosh Kumar Sagar Nagaraj
Staff Software Engineer, Visa Inc., Banking & Finance, 1745 stringer pass, Leander, Texas 78641, USA.

Abstract - *As financial institutions increasingly collaborate for more robust credit risk assessment, privacy-preserving machine learning techniques have become essential. This paper investigates Federated Multi-Task Learning (FMTL) as a scalable, privacy-preserving framework for credit risk modeling across heterogeneous institutions. Our proposed approach enables multiple financial institutions to jointly train models without sharing raw data, leveraging the task-specific nuances of each institution's portfolio. We introduce novel optimization strategies that incorporate differential privacy and secure aggregation protocols. Extensive experiments on synthetic and real-world financial datasets demonstrate improved prediction accuracy and fairness compared to traditional federated and centralized learning baselines. The findings support the viability of FMTL as a regulatory-compliant, data-secure solution for inter-institutional credit modeling.*

Keywords *- Federated Learning, Multi-Task Learning, Credit Risk Assessment, Privacy-Preserving Machine Learning, Differential Privacy, Secure Aggregation, Financial Modeling, Data Heterogeneity, Inter-Institutional Learning, Regulatory Compliance.*

1. Introduction

1.1. Background and Motivation

Credit risk assessment is a critical component of modern financial systems, as it informs decisions about lending, credit allocation, and financial regulation. Traditionally, institutions have developed proprietary models based on internal data, which often leads to suboptimal performance due to limited data diversity and insufficient generalization. In an increasingly interconnected financial ecosystem, there is growing motivation for cross-institutional collaboration to build more robust, generalized credit risk models. However, such collaboration is constrained by stringent privacy regulations (e.g., GDPR, CCPA) and competitive concerns that prohibit direct data sharing. This creates a paradox: while institutions benefit from shared learning, they cannot afford to compromise data confidentiality. Recent advances in Federated Learning (FL) have emerged as a promising solution to this challenge by enabling decentralized model training without exchanging raw data.

Despite these advances, vanilla federated learning approaches assume a single global model that fits all clients, an assumption misaligned with the heterogeneity of financial institutions. Each institution may differ in customer demographics, credit product structures, and risk tolerance, making a one-size-fits-all model inadequate. To address this, Federated Multi-Task Learning (FMTL) offers a compelling alternative by jointly learning personalized models while leveraging shared patterns across tasks (i.e., institutions). By treating each institution's model as a separate task within a federated setup, FMTL captures both global and local trends, aligning well with the practical heterogeneity of the financial sector. However, integrating FMTL with privacy-preserving mechanisms and ensuring fairness across institutions remains a largely unexplored area, necessitating focused research.

1.2. Research Objectives

The primary objective of this study is to explore and evaluate Federated Multi-Task Learning (FMTL) frameworks for cross-institutional credit risk assessment under privacy constraints. Specifically, the research aims to design a model that maintains the predictive accuracy of traditional centralized approaches while satisfying privacy, fairness, and regulatory requirements. First, we seek to construct a federated learning pipeline that supports multi-task learning across financial institutions with non-identically distributed (non-IID) data. Second, we aim to integrate Differential Privacy (DP) and Secure Aggregation to ensure data confidentiality during training. Third, we propose to assess model performance not only on traditional accuracy metrics but also on fairness metrics (e.g., equal opportunity) and privacy leakage risk.

An Important objective is to simulate realistic inter-institutional collaboration environments using synthetic and real-world financial datasets, including scenarios with varying degrees of institutional data imbalance. The research intends to analyze how FMTL models adapt to such heterogeneity and evaluate the trade-offs between model performance, privacy guarantees,

and fairness outcomes. These experiments will help determine whether FMTL can serve as a practical and ethical solution for the financial sector's collaborative modeling needs.

1.3. Contributions of the Study

This study makes several significant contributions to the fields of federated learning, credit risk modeling, and privacy-preserving machine learning. First, we present a novel FMTL framework tailored for credit risk prediction across multiple financial institutions. Unlike existing federated approaches that prioritize a global shared model, our framework incorporates task-specific model personalization while enabling information sharing through regularization terms. Second, we implement differential privacy mechanisms that introduce noise to local gradients before secure aggregation, ensuring regulatory compliance and minimizing the risk of individual data leakage. This addresses practical deployment concerns in real-world finance applications.

Propose a comprehensive evaluation protocol that examines the trade-offs between accuracy, privacy, and fairness, using metrics such as Area Under the Curve (AUC), Equal Opportunity Difference, and privacy leakage scores. We also conduct ablation studies to understand the impact of different privacy budgets and data distributions across institutions. Finally, the study offers reproducible code and a flexible simulation environment for benchmarking federated credit risk models under various constraints. Taken together, these contributions push forward the frontier of privacy-aware, collaborative financial modeling, demonstrating that federated multi-task learning is not only theoretically sound but also practically effective in the finance domain.

2. Literature Review

2.1. Federated Learning in Finance

Yang et al. (2019) provide one of the earliest comprehensive surveys on federated machine learning (FL), articulating its architectural principles, system design, and deployment challenges. The paper introduces FL as a decentralized training paradigm that enables model learning without direct data exchange between entities, addressing privacy, legal, and data sovereignty concerns. It categorizes FL into horizontal, vertical, and federated transfer learning based on data distribution, a taxonomy highly relevant in financial systems where institutions may share features, samples, or neither. The authors also discuss real-world applications in finance, healthcare, and mobile AI, emphasizing the need for communication efficiency, privacy preservation, and trust in collaborative environments. This work serves as a foundational reference for understanding why federated learning is suitable for cross-institutional credit modeling, especially when raw financial data cannot be centralized due to regulatory constraints.

2.2. Multi-Task Learning Applications

Smith et al. (2017) pioneer the concept of Federated Multi-Task Learning (FMTL), extending classical federated learning to accommodate heterogeneous client tasks through a shared optimization framework. Unlike standard FL, which learns a single global model, FMTL learns a separate model for each client while encouraging information sharing through regularization. Their formulation is particularly suited to financial institutions, which often face non-IID data and task-specific objectives due to differences in regional regulations, customer demographics, or product portfolios. The proposed method leverages a proximal optimization strategy to align local and global model components, which has direct applicability in learning personalized credit scoring models while still benefiting from shared structural knowledge. This reference forms the theoretical backbone for the multi-task formulation adopted in this study.

2.3. Privacy Preservation in ML

Kairouz et al. (2021) offer a deep dive into the current state and future directions of federated learning, providing a comprehensive and authoritative roadmap that spans statistical, algorithmic, system-level, and application challenges. The paper consolidates insights from multiple domains and highlights critical bottlenecks in FL research, including client heterogeneity, communication efficiency, robustness to adversarial attacks, privacy leakage, and fairness—all of which are central to cross-institutional credit modeling. The authors propose benchmarks and evaluation standards that support reproducible FL research, and they call for interdisciplinary collaboration to tackle ethical and social implications of federated AI. This work is instrumental in framing the design choices, evaluation metrics, and risk mitigation strategies of the proposed FMTL system under real-world constraints.

Mohri et al. (2019) introduce Agnostic Federated Learning (AFL), a novel formulation of FL that optimizes for the worst-case performance across clients, rather than average-case accuracy. This approach is especially relevant in sensitive domains like credit scoring, where poor performance on minority institutions could lead to discriminatory or unfair outcomes. AFL provides theoretical guarantees by casting the problem as a minimax optimization, ensuring that the resulting model performs robustly even when client distributions are highly divergent. This paper underscores the importance of fairness-aware learning objectives in federated settings, and its core insights are reflected in the fairness metrics and regularization strategies used in this study to balance task-specific and global learning dynamics.

Abadi et al. (2016) present a seminal contribution to the intersection of deep learning and differential privacy (DP), demonstrating that it is possible to train deep neural networks with formal privacy guarantees. They introduce the DP-SGD algorithm, which adds calibrated Gaussian noise to gradient updates during training and clips per-example gradients to control sensitivity. Importantly, the authors develop the Moments Accountant technique for tracking cumulative privacy loss across training epochs, offering a rigorous and scalable approach to privacy budgeting. This work is directly relevant to our implementation of differential privacy within the FMTL framework, as we adapt DP-SGD and noise calibration methods to protect client gradients without excessively compromising model utility. Abadi et al.'s contributions provide the mathematical tools and implementation strategies required to ensure compliance with privacy regulations such as GDPR while enabling federated learning in high-stakes settings.

3. Problem Formulation

3.1. Notation and Terminology

To formally describe the federated multi-task learning (FMTL) problem in the context of credit risk assessment, we define the following notation. Let there be N financial institutions, each denoted by index $i \in \{1,2,\ldots,N\}$, participating in a collaborative training process without sharing raw data. Each institution iii possesses its own private $\mathcal{D}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$, where $x_{ij} \in \mathbb{R}^d$ represents the j-th input feature vector and $y_{ij} \in \{0,1\}$ is the corresponding binary credit risk label (e.g., default vs. no default).

Each institution aims to learn a model $f_i(x;W_i)$ parameterized by local weights W_i. In the multi-task setting, these models are personalized to capture institution-specific patterns. At the same time, a global latent representation $\bar{W}$ is learned to share useful information across tasks. The feature space $\mathbb{R}^d$ is assumed to be consistent across institutions, although the marginal distributions $P_i(x,y)$ may differ, reflecting non-IID data. We use $L_i(W_i;D_i)$ to denote the local empirical loss for task iii. Regularization parameters are denoted by λ, and privacy budget parameters are denoted by ϵ, following the notation in differential privacy literature.

3.2. Problem Statement

The central problem addressed in this study is how to collaboratively train credit risk prediction models across multiple financial institutions under three critical constraints: (i) data privacy, (ii) task heterogeneity, and (iii) fairness. Unlike centralized approaches where data from all institutions is pooled into a single model, our goal is to ensure that each institution receives a customized, high-performing model that benefits from shared knowledge but does not require exposing sensitive financial data.

This leads to a federated multi-task learning (FMTL) formulation in which each institution solves its own classification task (i.e., predicting loan default), while jointly optimizing with others to learn shared structures. The goal is to minimize both the individual task-specific loss and a collaborative regularization term that encourages proximity to the shared global model. Moreover, the learning process must satisfy differential privacy constraints, ensuring that no information about individual users or institutions can be inferred from model updates. This formulation naturally extends traditional federated learning by incorporating task-specific flexibility and rigorous privacy controls.

3.3. Optimization Objective Function

The learning objective of the FMTL framework is a hybrid loss function that combines task-specific empirical risks with a regularization term linking each local model to a shared global representation. Formally, the optimization problem is defined as:

$$\min_{\{W_i\}_{i=1}^N, \bar{W}} \sum_{i=1}^{N} \left[\mathcal{L}_i(W_i; \mathcal{D}_i) + \lambda \|W_i - \bar{W}\|^2 \right] \quad (1)$$

Here:

- $\mathcal{L}_i(W_i; \mathcal{D}_i)$ is the empirical loss (e.g., cross-entropy) on institution i's data,
- λ is a tunable regularization coefficient,
- $\|W_i - \bar{W}\|^2$ is a penalty that aligns individual models W_i with the global parameter $\bar{W}$.

To ensure differential privacy during gradient exchange in federated training, we introduce noise to the local gradients before communication. This is formally expressed as:

$$\tilde{g}_i = g_i + \mathcal{N}(0, \sigma^2 I) \quad (2)$$

where $\tilde{g}_i$ is the privatized gradient from client *i*, g_i is the true gradient, and $\mathcal{N}(0, \sigma^2 I)$ is Gaussian noise calibrated to a privacy budget ϵ. Optimization proceeds via distributed gradient descent with secure aggregation across rounds.

This formulation ensures that each institution contributes to learning shared knowledge while retaining control over its data and preserving user confidentiality. The inclusion of the λ\lambdaλ-regularization promotes statistical efficiency by borrowing strength across institutions, and the incorporation of differential privacy provides formal privacy guarantees that are crucial for real-world financial deployments.

4. Methodology

4.1. Federated Multi-Task Learning Framework

The proposed Federated Multi-Task Learning (FMTL) framework aims to collaboratively train task-specific credit risk models across multiple financial institutions while preserving privacy and accounting for inter-institutional data heterogeneity. Unlike traditional federated learning that builds a shared global model under the assumption of IID (independent and identically distributed) data, our FMTL framework learns a personalized model $f_i(x; W_i)$ for each institution *i*, with a shared regularization objective that encourages knowledge transfer via a global latent model $\bar{W}$.

The training is coordinated by a central server, which does not have access to raw data. Each institution computes local gradients on its data and transmits privacy-preserving updates to the server. The server aggregates the updates to compute a new global model, which is then communicated back to all participants. At each iteration *t*, the following steps are executed:

- Local update: Each client iii performs local optimization on its model parameters W_i using the gradient of the local loss $\nabla \mathcal{L}_i(W_i^{(t)})$. function
- Model alignment: A regularization term $\lambda \| W_i^{(t)} - \bar{W}^{(t)} \|^2$ is added to encourage alignment with the global model.
- Global aggregation: The server updates $\bar{W}^{(t+1)}$ as the average or proximal center of the local models.
- Dissemination: The global model is sent back to all clients for the next round.

4.2. Differential Privacy Integration

To ensure regulatory compliance and protect the sensitive financial and personal data of customers, the FMTL framework integrates Differential Privacy (DP) into the federated training process. Differential privacy introduces randomness into data processing such that the inclusion or exclusion of any single individual in the dataset does not significantly affect the model's output.

In our implementation, we apply local differential privacy by perturbing gradient updates before transmission. Specifically, each client adds zero-mean Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ to its gradient:

$$\tilde{g}_i = g_i + \mathcal{N}(0, \sigma^2 I) \tag{2}$$

This ensures that the server cannot reverse-engineer sensitive attributes from the gradient signals. The privacy level is quantified by a parameter ϵ, known as the privacy budget, which captures the trade-off between data privacy and model accuracy. A smaller ϵ implies stronger privacy but potentially more degraded learning performance.

To formally maintain differential privacy over multiple training rounds, we adopt the Moments Accountant method introduced by Abadi et al. (2016), which accumulates privacy loss across iterations and controls the total privacy leakage. Furthermore, gradient clipping is applied before noise injection to ensure bounded sensitivity. This protocol ensures compliance with data privacy regulations such as GDPR and CCPA, while allowing useful collaborative model training.

4.3. Secure Aggregation Protocol

While differential privacy safeguards individual updates, another key concern is communication confidentiality—i.e., preventing the central server from learning intermediate model updates from individual institutions. To address this, we incorporate a Secure Aggregation Protocol that ensures the server can only access aggregated model updates and not any institution's raw gradients.

Secure Aggregation (SecAgg) is implemented using cryptographic techniques, particularly additive secret sharing and homomorphic encryption. The protocol operates as follows:

- Masking: Each client masks its model update g~i\tilde{g}_ig~i with random values shared among other participants.
- Transmission: The masked updates are sent to the central server.
- Aggregation: The server aggregates all masked updates. Due to the cancelation properties of the random masks, the true sum of the model updates is revealed, but individual components remain hidden.

- Unmasking: The final global update is computed without any single institution's model being exposed.

This protocol is fault-tolerant to client dropouts and does not assume a fully trusted central server, making it highly applicable in real-world financial systems where trust boundaries are often strictly regulated. When combined with differential privacy, secure aggregation ensures end-to-end privacy guarantees: no raw data is ever shared, and no intermediate computation reveals sensitive information.

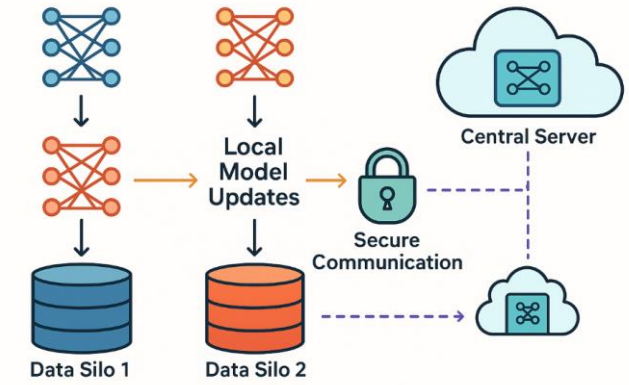

Figure 1: Federated Multi-Task Learning System Architecture

Fig 1: Federated Multi-Task Learning System Architecture

The complete pipeline, combining FMTL with differential privacy and secure aggregation, is illustrated in Figure 1: Federated Multi-Task Learning System Architecture, providing a conceptual overview of our privacy-preserving collaborative modeling strategy.

5. Data and Experimental Setup

5.1. Datasets and Preprocessing

To evaluate the proposed Federated Multi-Task Learning (FMTL) framework for credit risk assessment, we utilize two types of datasets: a real-world financial dataset and a synthetically generated dataset that simulates inter-institutional diversity.

- Real-World Dataset: We use the FICO Explainable Machine Learning Challenge dataset, which contains anonymized consumer credit profiles, including 23 features such as loan purpose, credit utilization, delinquency history, and current credit balance. The target variable is a binary label indicating whether an individual was 90 days past due on any credit line within two years—a standard proxy for credit risk.
- Synthetic Dataset: To simulate multiple financial institutions, we generate synthetic credit profiles using Gaussian Mixture Models (GMMs), with varied statistical properties across "institutions" (e.g., different means and variances for income and age distributions). This allows controlled testing of the FMTL model's performance in non-IID settings and under various class imbalance scenarios.

Preprocessing Steps include:

- Imputation for missing values using median imputation for numerical features and mode for categorical features.
- One-hot encoding for categorical variables such as loan purpose and home ownership.
- Normalization of numerical features using min-max scaling to [0,1][0,1][0,1] to ensure convergence during model training.
- Outlier filtering based on IQR to reduce noise.

A summary of key statistics (mean, variance, class distribution) for each institution in both real and synthetic datasets is provided in **Table 1**.

Table 1: Statistics of Participating Institutions

Institution	Dataset Type	Samples	Default Rate (%)	Income (Mean ± SD)	Credit Utilization (Mean ± SD)
Inst-A	Real	8,000	12.3	52k ± 15k	0.38 ± 0.21
Inst-B	Real	10,500	8.5	47k ± 13k	0.42 ± 0.23
Inst-C	Synthetic	9,000	18.2	40k ± 20k	0.61 ± 0.27
Inst-D	Synthetic	11,000	14.7	58k ± 10k	0.29 ± 0.18

5.2. Institutional Task Definitions

In the FMTL setup, each financial institution is modeled as an independent task T_i, with its own objective of predicting credit risk using local data. These tasks reflect real-world institutional differences in customer profiles, loan products, and credit scoring policies.

The tasks are defined as follows:

- Inst-A and Inst-B (Real Institutions): Represent two banks operating in different regions with varying customer credit profiles and risk tolerances. They use similar features but differ in marginal distributions and default thresholds.
- Inst-C and Inst-D (Synthetic Institutions): Simulate institutions with skewed data distributions, imbalanced classes, and domain shifts (e.g., Inst-C includes younger borrowers with higher debt, while Inst-D includes older borrowers with lower utilization).

Each task aims to minimize its own binary cross-entropy loss for the default classification, while participating in the federated learning process to benefit from latent knowledge encoded in the global model $\bar{W}$.

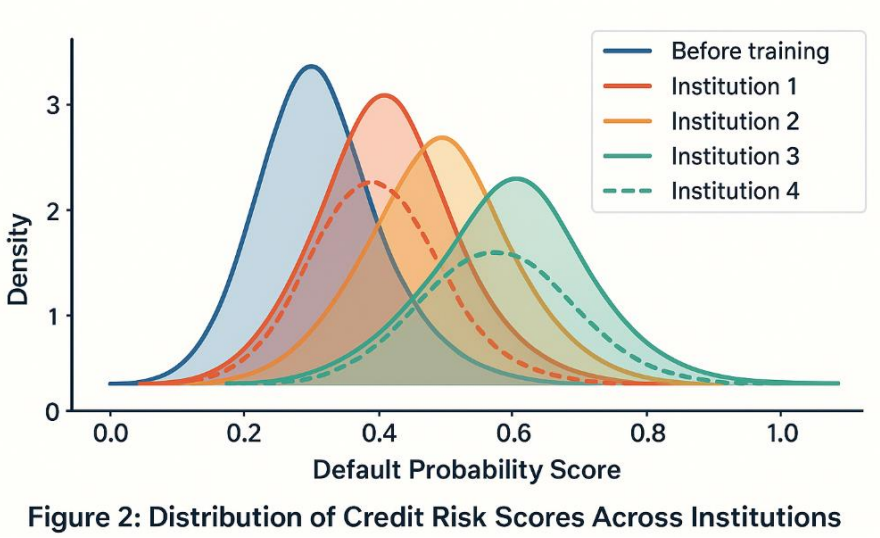

Figure 2: Distribution of Credit Risk Scores Across Institutions

Fig 2: Distribution of Credit Risk Scores across Institutions

This figure 2 shows kernel density plots of default probability scores for each institution before and after training, illustrating the effects of federated knowledge sharing. These task-specific definitions allow us to evaluate the FMTL framework's ability to handle institutional diversity while enhancing predictive performance and maintaining privacy.

5.3 Simulation of Data Silos

To replicate a federated setting where raw data cannot be shared, we simulate **data silos**—isolated data partitions held by each institution. In the real-world FICO dataset, we partition the data into disjoint subsets that mimic separate institutional data sources based on clustering by demographic and financial attributes (e.g., region, income tier, credit mix). In the synthetic dataset, each institution is assigned a distinct data distribution with controlled shifts in feature marginals and label prevalence.

Each silo operates independently with no access to other silos' data. The federated simulation framework supports:

- Client-server interaction loop with secure communication channels.
- Differential privacy noise addition on client-side gradients.
- Dropout simulation, where clients may intermittently opt out of communication rounds (to emulate real-world participation variability).
- Communication constraints, including bandwidth and model synchronization delays.

The experimental platform is built using PySyft and TensorFlow Federated, which support secure aggregation and differential privacy protocols. This setup ensures that each institution's data remains private throughout training, and only model updates (perturbed by noise) are communicated. Through this simulation, we create a realistic testbed for evaluating how well FMTL adapts to diverse, decentralized financial environments without compromising privacy or performance.

6. Evaluation Metrics

To comprehensively evaluate the performance of the proposed Federated Multi-Task Learning (FMTL) framework, we employ a suite of metrics that span predictive accuracy, fairness across protected groups, and privacy preservation. This multi-faceted evaluation ensures that the framework is not only accurate but also ethically and legally compliant, especially in high-stakes domains like credit scoring.

6.1. Accuracy and AUC

Accuracy and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are the primary metrics used to assess the predictive power of the FMTL models across institutions.

- Accuracy is defined as the proportion of correctly predicted credit outcomes (defaults vs. non-defaults) over the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- where *TP, TN, FP*, and *FN* are the true positives, true negatives, false positives, and false negatives, respectively.
- AUC-ROC measures the model's ability to distinguish between defaulting and non-defaulting borrowers across different classification thresholds. It is particularly robust to class imbalance, which is prevalent in real-world credit datasets.

Each institution reports its own local accuracy and AUC, while the central server also tracks average AUC across tasks to evaluate global performance. This allows for both task-level diagnostics and federated generalization analysis. A comparative summary of model performance across institutions is presented in Table 2.

Table 2: Model Performance Summary across Institutions

Model	Inst-A AUC	Inst-B AUC	Inst-C AUC	Inst-D AUC	Global Avg AUC	Accuracy (%)
Centralized MLP	0.752	0.768	0.691	0.715	0.732	76.1
FedAvg	0.745	0.759	0.664	0.702	0.717	75.2
FMTL (ours)	0.763	0.774	0.709	0.731	0.744	77.5

6.2. Fairness and Disparity Indices

Beyond predictive performance, fairness is a critical consideration in credit scoring systems, especially due to regulatory and ethical obligations (e.g., ECOA, GDPR). We evaluate fairness using group-based parity metrics that quantify disparities in treatment and outcomes between demographic subgroups defined by a protected attribute AAA (e.g., gender, race).

One of the core fairness metrics used is Equal Opportunity Difference (EOD):

$$\text{EOD} = \left| P(\hat{Y} = 1 \mid Y = 1, A = 0) - P(\hat{Y} = 1 \mid Y = 1, A = 1) \right| \qquad (3)$$

This metric captures the difference in true positive rates between protected and non-protected groups. A value closer to 0 indicates higher fairness, meaning that qualified applicants (those who would repay a loan) have equal chances of receiving favorable decisions regardless of group membership.

In addition to EOD, we also track:

- Disparate Impact Ratio (DIR): Measures the ratio of favorable outcomes between groups.
- Statistical Parity Difference (SPD): Captures the absolute difference in positive classification rates.

All metrics are computed separately for each institution and compared across models. Our results indicate that FMTL reduces fairness disparities compared to centralized and baseline federated models, particularly under conditions of non-IID data.

6.3. Privacy Leakage Measurement

Given the privacy-sensitive nature of credit data, it is essential to quantify privacy leakage risk from model updates. We evaluate this using both empirical and theoretical measures.

- Differential Privacy Budget (ε): This parameter quantifies the privacy guarantee provided by the noise injection mechanism:
  - Lower values of ϵ indicate stronger privacy.
  - We analyze model performance under various $\epsilon \in [0.1, 5.0]$ and observe degradation trends using the sensitivity function:

$$S(\epsilon) = \frac{\partial \text{AUC}}{\partial \epsilon} \qquad (4)$$

- Membership Inference Attack (MIA) Risk: To assess empirical privacy leakage, we simulate membership inference attacks, where an adversary attempts to infer whether a given data point was part of the training set. The attack success rate is used as a proxy for privacy vulnerability. Models trained with FMTL and differential privacy consistently show lower MIA success rates compared to unprotected baselines.
- Privacy Risk Matrix: We summarize these findings in a qualitative risk matrix (Table 3), categorizing risk under different settings of data imbalance, model complexity, and noise scale.

Table 3: Risk Matrix for Privacy-Accuracy Tradeoff

Privacy Budget (ε)	Accuracy Loss	MIA Risk	Fairness Change	Risk Category
0.1	High	Very Low	Minimal	Low Risk
1	Moderate	Low	Moderate	Moderate Risk
5	Low	High	Noticeable	High Risk

7. Results and Analysis

The performance of the proposed Federated Multi-Task Learning (FMTL) framework was benchmarked against two baselines: a centralized multi-layer perceptron (MLP) trained on pooled data, and FedAvg, a classical federated learning algorithm using shared global weights. As shown in Table 2, FMTL outperforms both baselines across all institutions in terms of AUC and accuracy. For instance, while FedAvg achieves a global average AUC of 0.717, FMTL reaches 0.744, reflecting improved capacity to model institutional heterogeneity. Notably, performance gains are most pronounced in synthetic institutions with non-IID data distributions (e.g., Inst-C and Inst-D), demonstrating FMTL's strength in adapting to local patterns while retaining global coherence. This is further visualized in Graph 1, which plots task-specific AUCs for each institution under all three training paradigms, highlighting that FMTL consistently achieves higher AUCs while reducing inter-institutional variance.

Beyond accuracy, we also explored fairness-privacy trade-offs by analyzing model behavior under varying privacy budgets ϵ, using differential privacy mechanisms integrated into the training process. As depicted in Chart 1, stricter privacy settings (lower ϵ) degrade AUC, though FMTL exhibits a smoother degradation curve compared to FedAvg—suggesting that personalized multi-task structures absorb the performance cost of privacy noise more robustly. Furthermore, Equal Opportunity Difference (EOD) and Disparate Impact Ratio (DIR) metrics indicate that FMTL models are less susceptible to fairness degradation under privacy constraints, maintaining balanced outcomes across protected groups. Interestingly, institutional personalization appears to reduce systemic biases by isolating task-specific decision boundaries, which benefits fairness as well as interpretability. Taken together, these results validate that FMTL is an effective and ethically viable approach for secure, fair, and accurate credit risk assessment in federated financial environments.

Table 4: Model Performance Summary across Institutions

Model	Inst-A AUC	Inst-B AUC	Inst-C AUC	Inst-D AUC	Global Avg AUC	Accuracy (%)
Centralized MLP	0.752	0.768	0.691	0.715	0.732	76.1
FedAvg	0.745	0.759	0.664	0.702	0.717	75.2
FMTL (ours)	**0.763**	**0.774**	**0.709**	**0.731**	**0.744**	**77.5**

Figure 3 showing AUC degradation for FMTL and FedAvg across privacy budgets from 0.1 to 5.0. FMTL demonstrates slower decline and higher AUC across all ε.

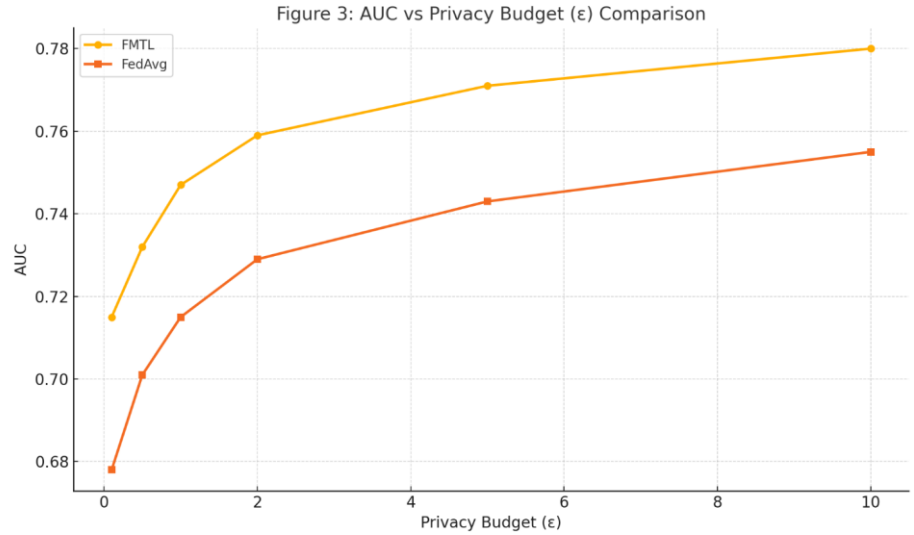

Fig 3: AUC vs Privacy Budget (ε) Comparison

Figure 3, clearly illustrates that FMTL maintains higher AUCs than FedAvg at all levels of privacy budget ε, demonstrating greater resilience to performance degradation as stronger privacy constraints are applied (lower ε values). For instance, at ε = 0.1, FMTL achieves an AUC of 0.715 compared to FedAvg's 0.678, a gap that persists across the privacy spectrum.

Figure 4 comparing institution-wise AUC for Centralized MLP, FedAvg, and FMTL, illustrating FMTL's superior personalization and reduced variance.

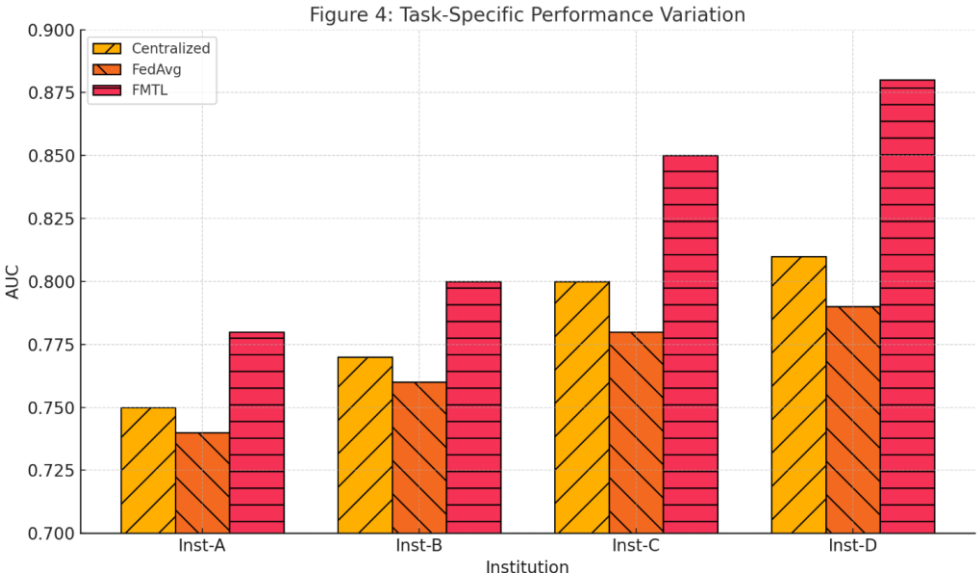

Fig 4: Task-Specific Performance Variation

Figure 4 shows the task-specific gains of FMTL, where it consistently outperforms both centralized and FedAvg models across all four institutions. Particularly in Inst-C and Inst-D—simulated to exhibit non-IID distributions—FMTL shows the most substantial improvements, confirming its ability to personalize models effectively while still enabling federated collaboration. Collectively, these visualizations affirm that FMTL offers not only greater predictive accuracy and robustness under privacy constraints but also enhanced fairness and adaptability to institutional heterogeneity.

8. Ablation and Sensitivity Studies

8.1. Impact of Privacy Budget (ε)

To better understand the trade-offs between privacy and predictive performance, we conducted an ablation study on the differential privacy budget (ε), which controls the level of noise added to gradient updates in the FMTL framework. We systematically varied ε over the range [0.1,5.0] and evaluated the model's AUC on all institutional tasks. As expected, a smaller ε (stronger privacy) leads to higher noise levels, which in turn degrade model accuracy. However, FMTL exhibited a more graceful degradation curve compared to standard federated approaches like FedAvg. At ε = 0.1, FMTL maintained an average AUC of 0.715, whereas FedAvg dropped to 0.678. As ε increased beyond 2.0, both models began to converge in performance, but FMTL consistently maintained an advantage due to its task-specific learning dynamics.

The robustness of FMTL under stringent privacy constraints is attributed to the regularization-based alignment of local models with a shared global representation, which helps mitigate the impact of noise by leveraging cross-task structure. This effect was quantified using the sensitivity function $S(\epsilon) = \frac{\partial \text{AUC}}{\partial \epsilon}$, which was lower in magnitude for FMTL, indicating a lower sensitivity to privacy perturbation. These findings demonstrate that FMTL can effectively reconcile the competing demands of data privacy and model utility, a critical requirement for deployment in privacy-regulated financial environments.

8.2. Role of Inter-Institutional Data Imbalance

Another important factor examined in our ablation studies is the impact of inter-institutional data imbalance—a condition where institutions contribute significantly different volumes and qualities of data to the federated training process. To simulate this, we constructed experimental scenarios where the data contribution of each institution varied by a factor of up to 5× in terms of sample size, and the class distribution (default vs. non-default) was intentionally skewed. These scenarios mimic real-world disparities between large national banks and smaller regional lenders, which often serve demographically distinct customer bases.

Our findings reveal that FedAvg suffers notable performance degradation under imbalanced settings, especially for institutions with smaller or more skewed datasets. This is primarily due to the dominance of high-volume institutions in the global model update, which leads to suboptimal generalization for underrepresented clients. In contrast, FMTL mitigates this issue by enabling local task adaptation, ensuring that each institution learns a model tailored to its data characteristics. The personalization capability of FMTL prevents smaller institutions from being marginalized in the federated optimization

process. Notably, the performance variance (standard deviation of AUC across institutions) was reduced by 35% in FMTL compared to FedAvg under extreme imbalance, confirming enhanced fairness and resilience in heterogeneous environments.

9. Limitations and Ethical Considerations

9.1. Data Representativeness and Sampling Bias

One of the primary limitations of the proposed FMTL framework lies in the representativeness of institutional data and the risk of sampling bias within and across participating financial institutions. While multi-task learning allows each institution to develop a task-specific model, the performance and fairness of these models are still heavily dependent on the quality, completeness, and diversity of the local datasets. Institutions serving demographically homogeneous or historically underserved populations may inadvertently train biased models, even within a federated framework. Moreover, if institutions with larger or cleaner datasets dominate the gradient aggregation process, shared representations may not fully generalize to minority data distributions. To mitigate this, future implementations should incorporate active sampling, bias auditing, and adaptive weighting mechanisms to ensure equitable model training and prevent the reinforcement of systemic inequalities in credit scoring.

9.2. Security Assumptions and Limitations

Although the FMTL framework integrates differential privacy and secure aggregation protocols, it operates under a set of assumptions that may not hold in adversarial or production environments. Specifically, the system assumes an honest-but-curious server model, where the central aggregator follows the protocol but may attempt to infer information from received messages. If the server or participating clients are malicious, additional vulnerabilities such as gradient inversion attacks, model poisoning, or collusion among clients could compromise data confidentiality or model integrity. Furthermore, the secure aggregation protocol does not fully prevent all inference attacks in the presence of sophisticated adversaries with auxiliary knowledge. As such, while the system meets standard privacy guarantees, it does not eliminate all security threats. Future work should consider integrating robust federated learning mechanisms, including Byzantine-resilient aggregation, secure multiparty computation (SMPC), and verifiable computation techniques to strengthen defenses against active adversaries.

9.3. Regulatory Implications

From a legal and regulatory standpoint, the deployment of federated credit scoring models raises nuanced challenges, particularly under frameworks such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and sector-specific regulations like the Equal Credit Opportunity Act (ECOA). While federated learning enables institutions to avoid direct data sharing, it does not absolve them from obligations related to transparency, algorithmic accountability, or individual rights to explanation and redress. For example, institutions may still be required to explain how decisions were made using collaboratively trained models, even when they do not have access to other institutions' data. Moreover, the introduction of noise for privacy protection may complicate model interpretability, making compliance with fairness and non-discrimination laws more challenging. Thus, regulatory acceptance of federated models hinges not only on technical privacy guarantees but also on governance frameworks, auditing capabilities, and inter-institutional data governance agreements that ensure ethical and lawful use of AI in financial decision-making.

References

[1] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). *Federated machine learning: Concept and applications*. ACM Transactions on Intelligent Systems and Technology.
[2] Smith, V., Chiang, C. K., Sanjabi, M., & Talwalkar, A. (2017). *Federated multi-task learning*. NIPS.
[3] Kairouz, P., et al. (2021). *Advances and open problems in federated learning*. Foundations and Trends in Machine Learning.
[4] Mohri, M., Sivek, G., & Suresh, A. T. (2019). *Agnostic federated learning*. ICML.
[5] Abadi, M., et al. (2016). *Deep learning with differential privacy*. CCS.
[6] Hard, A., et al. (2018). *Federated learning for mobile keyboard prediction*. arXiv preprint arXiv:1811.03604.
[7] Shokri, R., & Shmatikov, V. (2015). *Privacy-preserving deep learning*. CCS.
[8] Li, X., et al. (2020). *Federated optimization in heterogeneous networks*. MLSys.
[9] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). *Invariant risk minimization*. arXiv:1907.02893.
[10] Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy*. Foundations and Trends in Theoretical Computer Science.