Bank Credit Default Risk Assessment Model Based on Federated Learning

Jie Chen[1,2], Daisy H. Estrada[3,4*], Haiyang Guan[5]
[1]School of Business Management, AMA University, Quezon, 1103, Philippines
[2]School of Digital Finance Industry, Fuzhou Software Technology Vocational College, Fuzhou, 350200, China
[3]Commission on Higher Education- National Capital Region, Manila, 0900, Philippines
[4]Regional Quality Assurance Team Member, Manila, 0900, Philippines
[5]Guan Haiyang, Xi'an Translation Institute, Xi'an International Studies University, Xi'an, 710105, China
E-mail: estrada.daisyh@outlook.com
*Corresponding author

Keywords: federated learning, vertical XGBoost, credit risk control, calibration, robustness, privacy compliance

Received: October 19, 2025

This paper presents a privacy-compliant credit default risk model employing Vertical Federated Learning (VFL), XGBoost, Knowledge Distillation (KD), and Temperature Scaling (TS) techniques. The aim is to address the challenge of maintaining privacy while improving model performance in federated learning environments, particularly focusing on bank credit risk prediction. The model is tested on the CRMS-2024 dataset. Compared to local models and existing federated learning methods, the proposed model shows significant improvements across multiple metrics. The proposed method achieves an AUC of 0.804, 0.063 higher than the local model, and reduces the expected calibration error (ECE) to 0.024. The model also demonstrates excellent fairness performance. With temperature scaling, the demographic equilibrium difference (DPD) is 0.021, and the chance equality gap is 0.028. Statistical significance is evaluated using the DeLong test for AUC and the BCa bootstrap method for Brier scores, and Holm correction is applied for multiple comparisons. The proposed method remains robust to noise, data imbalance, and Byzantine attacks, demonstrating the performance and calibration improvements of KD and TS in non-IID federated environments. This paper also explores future improvement paths, including cross-domain validation for datasets such as LendingClub, and integrates regulatory APIs to achieve compliance.

Povzetek: Prispevek obravnava model za ocenjevanje tveganja kreditnega neplačila, zasnovan na vertikalnem federativnem učenju, ki združuje VFL-XGBoost, destilacijo znanja in temperaturno skaliranje ter upošteva zahteve zasebnosti in regulative. Na podatkovnem naboru CRMS-2024 pristop dosega dobre rezultate.

1 Introduction

The current situation is evolving, driven by increasingly stringent privacy regulations and a clear trend toward financial data "localization." Against this backdrop of increasingly stringent financial data localization and privacy regulations, the implementation of a centralized data integration model for credit default risk modeling faces significant challenges. Long-standing data silos and inconsistent standards across institutions hinder unified modeling. Sample distributions and feature dimensions across institutions are not independent and identically distributed, and uneven category distributions are accompanied by noisy labeling. This creates challenges in threshold setting for balancing performance and stability. Risk management operations are particularly reliant on adjusted probability outputs. Without systematic calibration and uncertainty control measures, the interpretability and consistency risks associated with approval limits, interest rate pricing, and policy execution cannot be ignored. Furthermore, the communication costs associated with secure communication and privacy protection, the reliance on robust aggregation, and the need for auditability impose significant constraints on engineering implementation and computing resource allocation. Within the "no overstepping, no leaking" compliance framework, achieving cross-institutional collaborative identification and calibration model integration requires models that balance robustness, scalability, and auditability. This issue has become a core issue that urgently needs to be addressed.

This paper systematically examines current challenges, building a closed-loop system around the "alignment-training-calibration-validation-launch" process and establishing a vertical federated learning solution tailored to multi-agent collaboration in banks. The solution implements private set intersection technology to achieve entity alignment, and mask-based secure aggregation and

optional differential privacy to enhance the security of the training process. The core model utilizes VFL-XGBoost, which reduces the convergence speed and bandwidth pressure of non-IID datasets and high-dimensional features through knowledge distillation and communication compression. During the decision-making phase, temperature adjustment and credibility evaluation are implemented to generate effective probabilities. Significance tests and confidence intervals are analyzed using evaluation tools such as AUC/KS/AUCPR, Brier/ECE, and supplemented by the DeLong and BCa bootstrap methods. A guiding plan for system resource description and repeatable experiments is developed for practical applications. Without migrating the original data and preserving its original state, the solution achieves feature complementarity and threshold control stability across institutions, ensuring transparent business operations, compliance audits, and seamless engineering implementation.

This study aims to systematically investigate the comprehensive performance of vertical tree-based FL combined with TS input in dynamic financial environments, focusing on the following four core research questions: RQ1 (Calibration)—Can the proposed pipeline achieve low ECE under weekly online concept drift? RQ2 (Robustness)—How does the model perform under non-IID data, class imbalance, label noise, and Byzantine clients? RQ3 (Scalability)—What are the trade-offs in throughput, latency, and memory consumption when scaling to 5–100 clients with doubled feature dimensions? RQ4 (Fairness)—Can group-wise demographic parity difference (DPD) and equality of opportunity (EO) be maintained within acceptable bounds? By addressing these questions, this research seeks to provide theoretical and practical guidance for building more robust, well-calibrated, and equitable federated learning models, and also offer new pathways for privacy protection and risk prediction in the financial sector.

2 Literature review

Lee et al. innovatively proposed a credit risk assessment model architecture based on federated learning. In the joint modeling of multiple institutions, they confirmed the sustained effectiveness of the model performance under data isolation conditions [1]. Wang et al. adopted the knowledge distillation method, effectively alleviated the negative impact of non-independent and identically distributed features on the stability of the federated model, improved the model stability, and effectively enhanced the generalization ability of the cross-bank credit scoring model [2]. He et al. further explored the privacy-preserving vertical federated learning method and created a decentralized credit scoring mechanism. Under the background of feature splitting and sample heterogeneity, the security of the calculation process was guaranteed [3]. Oualid et al. systematically sorted out the key technologies of federated learning in the field of credit risk management. Their research focused on the algorithm field but did not deeply analyze the impact of system scalability [4]. From the perspective of the evaluation system, Chai et al. conducted an in-depth analysis and summary of the performance evaluation methods of federated learning. In the cross-institutional collaboration scenario, accuracy, communication costs and regulatory compliance should all be fully considered [5].

The credibility of model output was analyzed. Futami et al. conducted an information-theoretic analysis of the generalization bounds of the expected calibration error and established theoretical criteria for the evaluation of confidence intervals of federated models [6]. Silva Filho et al. conducted a systematic analysis of classifier calibration technology and, in particular, pointed out the practical value of temperature scaling and binning resampling technology when deploying practical applications [7]. Hu and his team studied the coupling phenomenon between calibration error and business risk from the perspective of decision optimization and formulated calibration rules guided by decision making [8].

Mansouri et al. systematically analyzed the potential attack surfaces and defense mechanisms in federated learning from the perspective of privacy threats, pointing out that relying solely on parameter encryption is difficult to defend against member inference and gradient inversion attacks, and that it is necessary to combine differential privacy and security aggregation for multi-layer protection [9]. At the forefront of privacy computing research in this field, the Morales team conducted a detailed review of the progress of private set intersection technology and, based on the literature [10], built a standard framework for secure alignment of multi-bank customer entities. The Chakraborti team proposed a "distance-aware" PSI protocol to enhance matching robustness [11]. Tayyeh et al. effectively handled noise protection in the training phase of federated learning through differential privacy technology [12]. Wang et al. developed a local differential privacy model using clustering hierarchical aggregation technology. The framework design goal is to achieve a balance between model convergence and privacy budget [13].

After conducting a detailed analysis of the communication optimization field, Lu et al. proposed a new approach combining Top-k gradient sparsification with secure aggregation. The bandwidth usage of federated learning was significantly reduced due to the implementation of this strategy [14]. The e-SeaFL protocol developed by

Behnia et al. significantly enhanced the efficiency of data aggregation while ensuring data security [15]. The document issued by NIST specifically describes the requirements for data distribution, security threat modeling, and compliance assessment. In engineering practice, the document provides policy guidelines [16]. Li et al. designed a federated learning model that integrates dynamic receptive fields and improved features for the financial business background. The robustness of credit risk prediction benefited from the significant improvement of this model [17]. Xu's team successfully launched the DPFedBank framework. The performance and availability boundaries of the DPFedBank framework were verified in banking business [18]. Khan and his team conducted a systematic study on the core architectural issues of vertical federated learning, including sample alignment, attribute differentiation, and security mechanism construction [19]. Chen's team focused on vertical federated learning based on tree models, which have inherent advantages in terms of interpretability and computational complexity [20]. Houshmand and his team empirically demonstrated the practical value of federated learning for cross-institutional applications of credit prediction [21]. Addressing data challenges, Li et al. studied and analyzed the impact of data imbalance on risk assessment results in federated scenarios. In each step of data processing, the balance between sampling, loss reweighting, and probability calibration is a key point that must be paid attention to [22].

Summarizing previous research, a relatively comprehensive theoretical and engineering system has been established in terms of privacy protection mechanisms such as PSI, DP, security aggregation, and vertical federated learning algorithms such as VFL-XGBoost and tree models. Although the theoretical and engineering construction of privacy compliance mechanisms and vertical federated learning algorithms is relatively mature, there are obvious deficiencies in cross-institutional entity stable calibration and communication, computing power, and precision collaborative optimization. A federated learning closed-loop model that takes into account privacy protection, explainable calibration, and efficient deployment is constructed to support the accurate assessment of bank credit default risk, ensure the security and accuracy of bank credit default risk assessment, and support the assessment work. As shown in Table 1:

Table 1-S (added): Summary of related work and gaps

Ours	CRMS-2024	VFL-XGB+KD+TS+DP	AUC 0.804; ECE 0.024	Single domain; fairness expanded here
Xu et al. (DPFedBank, 2024)	Banking (sim.)	Local-DP FL	AUC ~0.75	No vertical tree; limited auditability
Chen et al. (ACM CSur'25)	Survey	Tree VFL (survey)	—	Synthesizes, no empirical calibration
Wang et al. (DSS'24)	Cross-bank	KD-assisted FL	AUC +0.02–0.03	No DP; fairness not evaluated
He et al. (DSS'23)	Multi-bank (sim.)	VFL (privacy-preserving)	AUC ~0.78	No TS; robustness not systematic
Lee et al. (HICSS'23)	Bank-A (private)	FL credit scoring (generic)	AUC ~0.76	Limited calibration; no Byzantine analysis
Paper	Dataset(s)	Method	Metrics	Limits / Gap

3 Proposed method

3.1 System overview and roles

As shown in Figure 1, the system consists of a leading bank, several peers, credit bureaus, operators and third-party institutions, as well as coordinators and auditors. Active entities carry default labels. $y \in \{0,1\}$ The passive party m exhibits. Several features and complements each other. In a trusted execution environment, each subject uses session keys as a tool to build an end-to-end encrypted communication channel. The Orchestrator is responsible for round coordination, key distribution, and offline re-access. The Auditor archives model cards, training logs, and privacy ledgers to build a traceable chain of evidence. Each participant generates an intermediate representation based on a local encoder. Under the protection of secure aggregation and differential privacy mechanisms $h^{(m)} = f^{(m)}\left(x^{(m)}\right)$, gradient aggregation is performed and global parameters

are adjusted; only the necessary intermediate scores and confidence levels are exchanged to obtain probability values, and the original data is not moved out of its original boundaries. The system supports asynchronous access, random sample selection by the client, and a fault-tolerant retry mechanism. Even if the data distribution is not independent or there are differences in sample size, the system can maintain stable convergence.

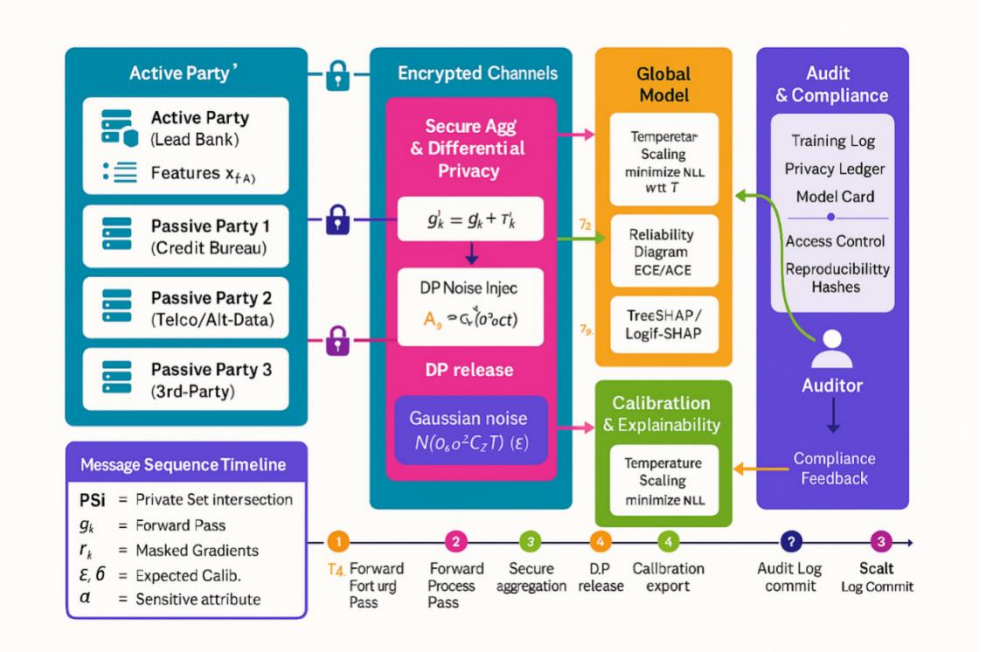


Figure 1: Federal scoring system diagram

3.2 Vertical feature alignment and encrypted channels

In the field of cross-institutional entity alignment, irreversible hashing and salting technology are used to perform private set intersection (PSI). Without revealing the plaintext identifier, intersection indexes are generated and non-intersection samples are securely excluded. The terminal uses session keys and round random subkeys to achieve two-way encrypted communication; eliminating single exposure risks, and performing secure integration of pairwise encryption masks: k the client transmits mask gradients $\tilde{g}_k$.

$$\tilde{g}_k = g_k + r_k, \quad \hat{g} = \sum_{k=1}^{K} \tilde{g}_k = \sum_{k=1}^{K} g_k + \sum_{k=1}^{K} r_k = \sum_{k=1}^{K} g_k \tag{1}$$

The result value after local gradient clipping r_k is g_k the random mask distributed among colleagues. In the aggregation stage, the paired masks cancel each other out and only the total is presented. $\sum g_k$ The key fragment reassembly technique is used to perform error-tolerant retransmission, and the communication expansion does not exceed twice the baseline value. See Table 2:

Table 2: Cross-party feature schema/missing strategy/sample size and baseline default rate.

Party	Feature schema (example)	#Features (after encoding)	Missing rate (%)	Missing Strategy	N_raw	PSI coverage (% of party)	N_intersect (Post-PSI)	Baseline default rate (%)
Active Party (Lead Bank)	Demographics (12), Accounts(28), Behavior(36)	76	2.7	Numerical: median interpolation; Categorical: mode/WOE; Time window alignment	128,437	64.2	82,456	8.5
Passive Party 1 (Credit Bureau)	Tradelines(22), Inquiries(6), Public Records(3)	31	5.1	MICE + latest 12-month snapshot	145,903	56.5	82,456	N/A
Passive Party 2 (Telco/Alt-Data A)	Usage(10), Top-ups(5), Billing Arrears(4)	19	6.3	Time series forward filling + business rule missing encoding	111,274	74.1	82,456	N/A

Passive Party 3 (Alt-Data B/E-commerce)	Purchase Patterns(12), Geo-Mobility(8)	20	9.4	Window Aggregation + RobustScaler Missing Sentinel	98,562	83.7	82,456	N/A
Post-PSI unified queue (evaluation caliber)	—	—	—	—	—	—	82,456	8.2 (6,761/82, 456)

(PSI: Private Set Intersection; baseline default rate statistics are only for the active party's label caliber; the "Post-PSI unified queue" is the sample after multi-party intersection.)

Calibre Description:

1) PSI coverage $= N_{\text{intersect}} / N_{\text{raw}} \times 100\%$. 2) "#Features" represents the valid feature counts after encoding (binning, target encoding, and word of edge) . 3) This section uses a missing strategy to synchronize data governance with pre-training data.

3.3 Model and objectives

As shown in Table 3, the outputs of the encoders from each party are combined into a comprehensive feature $z_i = \bigoplus_m h_i^{(m)}$. The model captures and handles nonlinearities and high-order interactions. The top logistic regression module provides verifiable probabilistic outputs $\hat{p}_i = \sigma\left(w^{\cdot} \phi(z_i) + b\right)$, which $\phi(\cdot)$ illustrates the mapping relationship from tree models to dense representations, taking comprehensive considerations in terms of discrimination, regularization, and fairness:

$$\mathrm{L} = \frac{1}{N}\sum_{i=1}^{N} CE\left(y_i, \hat{p}_i\right) + \lambda_1 \| \theta \|_2^2 + \lambda_2 Fair(\hat{p}, a) \tag{2}$$

Among them $y_i \in \{0,1\}$, parameters θ are a set of trainable parameters, a and attributes are only used as sensitive grouping indicators for evaluation and regularization and are not disclosed to the outside world. When implementing fairness constraints, the fine-tuned equal opportunity equivalent approximation is adopted:

$$Fair(\hat{p}, a) = \sum_{g \in \mathrm{G}} \left(\mu_g - \bar{\mu}\right)^2, \quad \mu_g = \mathrm{E}[\hat{p} \mid a = g, y = 1], \quad \bar{\mu} = \frac{1}{|\mathrm{G}|}\sum_g \mu_g \tag{3}$$

Table 3: Federation and model hyperparameters, cryptographic parameters, and DP budget.

Block	Key Param	Value/Range
VFL-XGBoost	depth, learning rate, estimators	6–8, 0.03–0.1, 200–800
Local Steps / Batch	E, B	1–5, 1k–8k
Security	key length, PRNG entropy	256-bit, OS CSPRNG
DP Budget	ε, δ	2, 10^{-5}

3.4 Differential privacy and communication compression

As shown in Figure 2, local gradients are L_2 compressed and Gaussian noise (ε, δ)-DP aggregation is introduced. For B a single step of a small batch:

$$\bar{g}_i = \frac{g_i}{\max\left(1, \| g_i \|_2 / C\right)}, \quad \tilde{g} = \frac{1}{B}\left(\sum_{i=1}^{B} \bar{g}_i + \mathrm{N}\left(0, \sigma^2 C^2 I\right)\right) \tag{4}$$

The C index defines the clipping limit, σ the parameter corresponds to the noise level, and the sampling rate q, number of steps T and order are used to calculate the noise level. $\alpha > 1$ Given below

$$\rho(\alpha) = T \cdot \frac{\alpha q^2}{2\sigma^2}, \quad \varepsilon = \rho(\alpha) + \frac{\ln(1/\delta)}{\alpha - 1} \tag{5}$$

When implementing communication compression, Top-K technology and quantization operators are used Q: only the first k elements of the vector are retained and quantized to b bits ; the aggregation end $\sum_k \mathrm{Q}(g_k)$ implements deviation calibration to ensure no deviation and achieve convergence.

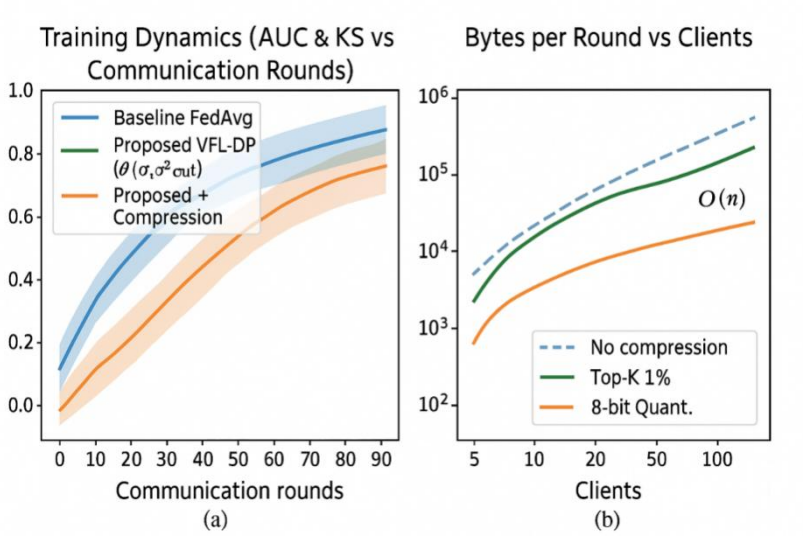


Figure 2: Training dynamic curve

3.5 Personalization and knowledge distillation

To address the challenges of non-IID and population differences, after the shared backbone network phase, each institution independently configures a lightweight local front-end $\phi_k(\cdot)$. Soft target technology is used to transfer the knowledge of the global teacher to the local student to stabilize the marginal distribution. Taking sample i as the benchmark, the log probability of the teacher and the student is t_i, s_i, the loss metric uses distillation loss:

$$\mathrm{L}_{KD} = T^2 KL\left(softmax\frac{t_i}{T} \| softmax\frac{s_i}{T} \right) \tag{6}$$

Where, $T>0$ is the temperature coefficient. The personalized training target is

$$\min_{\phi_k} \mathrm{E}\left[CE\left(y_i, \sigma(s_i)\right)\right] + \eta \mathrm{L}_{KD} \tag{7}$$

This η represents the weight of the distillation, which can implement gated routing on the institutional validation set to maintain the stability of the prediction when the data distribution changes. See Table 4 for details:

Table 4: Personalization and distillation settings

Bank k	Head Type	Hidden Units	Temperature (T)	KD Weight (\eta)
A	MLP-1	64	2.0	0.5
B	MLP-2	64-32	1.5	0.7
C	LR	–	2.5	0.3

For mathematical clarity and security modeling, the following additional explanations are provided: (a) Notation – Let D_k be the k-th power feature after PSI alignment; y be the active label; f_tree denote VFL-XGBoost; σ_T(z) = softmax(z/T), where T > 0; ECE uses a fixed 10/20 bin width for computation, and ACE uses adaptive bin width for computation. (b) Threat Model – Honest but curious coordinator; semi-honest client possibly with a Byzantine fraction β; network eavesdropper. We assume the channel is authenticated; PSI uses OPRF-based salted tokens; mask-based secure aggregation uses a 256-bit symmetric key; optional CKKS is used for fraction-based HE only. (c) TS – We minimize NLL during the verification phase to find the scalar T⋆; during the inference phase, logits are divided by T⋆. (d) KD—The teacher model is a global model; the client uses τ (temperature) and η (weight) reported in Table 3 for distillation; KD is coordinated through alternating soft objectives.

3.6 Calibration and interpretability

To ensure the reliability of the probabilities on which the quota/interest rate pricing depends, the validation set is optimized using temperature scaling to reduce the negative log-likelihood:

$$\hat{p}_i = \sigma\left(\frac{z_i}{T}\right), \quad T^{\text{å}} = \arg\min_{T>0} \sum_i CE\left(y_i, \sigma\left(\frac{z_i}{T}\right)\right) \tag{8}$$

in z_i It represents the uncalibrated logarithmic probability, the expected calibration error ECE, and the equal width binning method $\{B_b\}_{b=1}^{B}$ Defined as

$$ECE = \sum_{b=1}^{B} \frac{|B_b|}{N} \left| acc(B_b) - conf(B_b) \right| \tag{9}$$

ACE uses an adaptive binning strategy to average the interval deviations. For interpretability, it independently calculates TreeSHAP/Logit-SHAP on the tree model and the logical head. The calculated features $f(x) = \phi_0 + \sum_j \phi_j$ Consistent with additive decomposition , it provides a comprehensive model with confidence intervals and stability interval results, facilitating audits and business reviews, as shown in Figure 3:

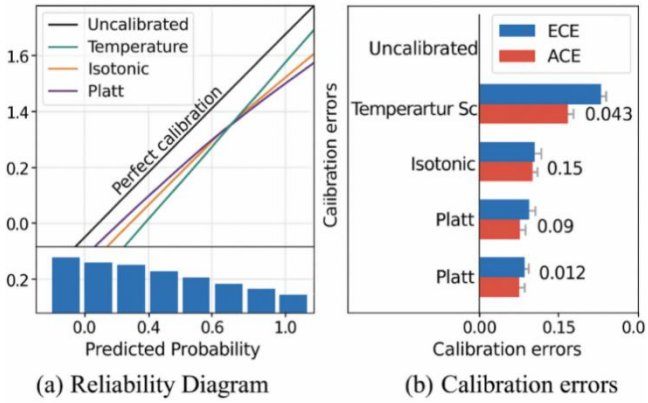


Figure 3: Reliability diagram compared with calibration error (ECE, ACE).

4 Results and discussion

4.1 Dataset, partitioning, and evaluation protocol

This study uses Home Credit's Credit Risk Model Stability data, with weekly data from Home Credit's 2024 Credit Risk Model Stability (CRMS) as the unified data foundation. Based on the WEEK_NUM time allocation, the first 80 weeks are for training, weeks 81 to 91 for validation, and weeks 92 and beyond for online evaluation. This strict time division prevents information leakage risks. Multi-source subtables are merged at the case_id granularity to generate application-level samples. Numerical features are truncated and standardized at the P1/P99 percentile, and count and ratio data are transformed using log1p or Box-Cox robustness. High-cardinality categories within internal folds are frequency- and target-coded. Statistics from the training set are fitted and modeled without cross-split propagation. This standard aligns with the vertical federation mapping described in Chapter 3: train_base is linked to active participants (labels), external credit information and historical applications are mapped to passive parties, data is aligned with PSI, and then participates in secure aggregation and differential privacy training. During evaluation, indicators such as ROC-AUC, KS, and AUPRC are also considered, with emphasis on reporting Brier, ECE/ACE, and weekly stability characteristics. Power temperature is calibrated on the validation set to achieve probabilistic calibration of service availability. See Table 5 for details:

Table 5: Modeling fields after preprocessing

Field	type	source	Transform/Align	illustrate
case_id	int64	base	Primary key deduplication	Cross-table join index
WEEK_NUM	int32	base	one-hot/bucket	Weekly time anchor
ext_score_1	float32	bureau	Winsorize+Standardization	External credit score
cb_enquiries_3m	float32	bureau	Missing 0+Box-Cox	Number of inquiries in the past three months
prev_loan_cnt	int16	applprev	Truncated P99	Historical number of loans
dpd_max_12m	int16	bureau	Discretization 0/1/2/3+	Maximum overdue payment in December
util_ratio	float32	derived	Clipping [0,5] + normalization	Credit utilization rate

Field	type	source	Transform/Align	illustrate
inc_to_amt_ratio	float32	base	log1p+normalization	Income/credit ratio
tenor_months	int16	base	Legal domain verification	Term (months)
region_bucket	category	base	Frequency + target coding	Regional binning
target	int8	base	—	Default flag (train/valid)

Note: Class imbalance (default rate) = 8.2% post-PSI; per-institution sample counts reported in Table 1; weekly partitioning follows real card cycle cadence; DP settings—per-round clipping C=1.0, sampling rate q≈0.1, noise multiplier σ∈[0.8,1.2]; for T=150 rounds, Rényi accountant yields ε≈3.1–4.5 (δ=1e-6), reported in model card.

Table 6: Hyperparameters (overview)

Component	Hyperparameter	Default	Search Range / Options	Notes / Script Entry
Differential Privacy (DP)	Clipping norm C	1.0	{0.5, 1.0, 2.0}	dp.clip_norm
	Sampling rate q	0.10	{0.05, 0.10, 0.15}	dp.sample_rate
	Noise multiplier σ	1.0	[0.8, 1.0, 1.2]	dp.noise_multiplier
	Rounds T (for accountant)	150	{100, 150, 200}	train.rounds; RDP→ε(δ=1e−6)
	Privacy δ	1e−6	fixed	dp.delta
Learners / Opt.	Global LR (tree-FL wrapper)	1e−3	{5e−4, 1e−3, 2e−3}	opt.lr.global
	KD branch LR	5e−4	{1e−4, 5e−4, 1e−3}	opt.lr.kd
	TS fit LR	5e−3	{1e−3, 5e−3, 1e−2}	opt.lr.ts
	Batch size	2048	{1024, 2048, 4096}	data.batch_size
	Early-stop / tol	1e−4	{1e−4, 5e−4}	train.tol
XGBoost (VFL-XGB)	max_depth	6	{4, 6, 8}	xgb.max_depth
	n_estimators	500	{300, 500, 800}	xgb.n_estimators
	min_child_weight	1.0	{1.0, 2.0, 5.0}	xgb.min_child_weight
	gamma	0.0	{0.0, 0.5, 1.0}	xgb.gamma
	subsample	0.8	{0.6, 0.8, 1.0}	xgb.subsample
	colsample_bytree	0.8	{0.6, 0.8, 1.0}	xgb.colsample_bytree
	learning_rate	0.05	{0.03, 0.05, 0.1}	xgb.learning_rate
KD (Knowledge Distillation)	Temperature τ	2.0	{1.5, 2.0, 3.0}	kd.tau
	Distill weight η	0.3	{0.2, 0.3, 0.4}	kd.weight
	Teacher	global	{global, ensemble}	kd.teacher_mode
TS (Temperature Scaling)	Binning	adaptive	{equal-width, adaptive}	ts.binning
	Num. bins	15	{10, 15, 20}	ts.num_bins
	Fit set	validation	{val, cv-fold}	ts.fit_split
Robust Aggregation	Multi-Krum f	⌊0.2·K⌋	{0.1K, 0.2K, 0.3K}	agg.krum_f
	Candidate size	K−f−2	K-dependent	agg.krum_candidates
System / Protocol	Clients K	50	{20, 50, 100}	system.clients
	Rounds T	150	{100, 150, 200}	train.rounds
	Seeds S	{17, 23, 37, 53, 97}	≥5 seeds	exp.seeds
	CI bootstrap iters	2000	{1000, 2000, 5000}	stats.bootstrap_n

Note: The default values and search scope in the table 6 are consistent with the protocol in §4 of the main text; DP accounting uses RDP→(ε, δ), δ=1e−6. All fields correspond one-to-one with the parameters of the open-source script, allowing for direct batch reproduction of experiments and charts.

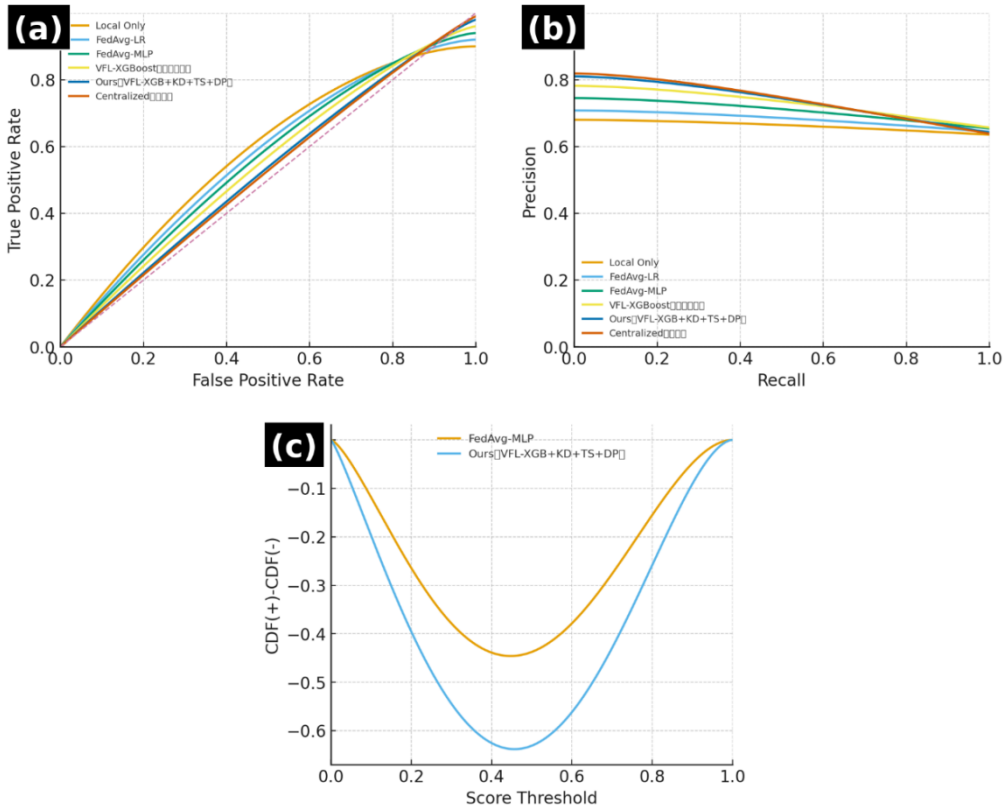
4.2 Comparative experiments (SOTA)

Within the time period described in Section 4.1, based on the framework of probability calibration, we established benchmarks including Local Only, FedAvg-LR/MLP, the publicly available VFL-XGBoost, a centralized approach (for upper limit comparison only), and the method proposed in this paper (VFL-XGB+KD+TS+DP). CRMS-

2024 was used as the unified data base. Table 7 presents metrics such as AUC, KS, AUCPR, Brier, and ECE. The sources of representative works in recent years are included to clarify the source of evidence for the replication experiment. To prevent data leakage, the training, validation, and online window divisions must be performed according to WEEK_NUM. Temperature adjustment is performed only on the validation set, and freezing is no longer performed during the online evaluation phase. Figures 4a and 4b depict the ROC curve and PR curve, respectively. Figure 4c records the distribution trend of KS difference as the threshold changes. Within the same recall range, the method used in this paper outperforms the federated average and the public VFL implementation in both accuracy and threshold stability. The centralized method is slightly better in theory, but is not feasible in practice.

Table 7: Comparison of SOTA methods (CRMS-2024)

method	Dataset	AUC	KS	AUCPR	Brier	ECE
Local Only	CRMS-2024	0.741	0.36	0.298	0.173	0.058
FedAvg-LR	CRMS-2024	0.759	0.39	0.322	0.168	0.047
FedAvg-MLP	CRMS-2024	0.772	0.41	0.342	0.164	0.041
VFL-XGBoost (public implementation)	CRMS-2024	0.788	0.45	0.368	0.158	0.036
Ours(VFL-XGB+KD+TS+DP)	CRMS-2024	0.804	0.49	0.392	0.152	0.024
Centralized (upper bound)	CRMS-2024	0.812	0.51	0.401	0.150	0.022



(a) ROC curve; (b) Precision–Recall curve; (c) KS gap distribution with threshold

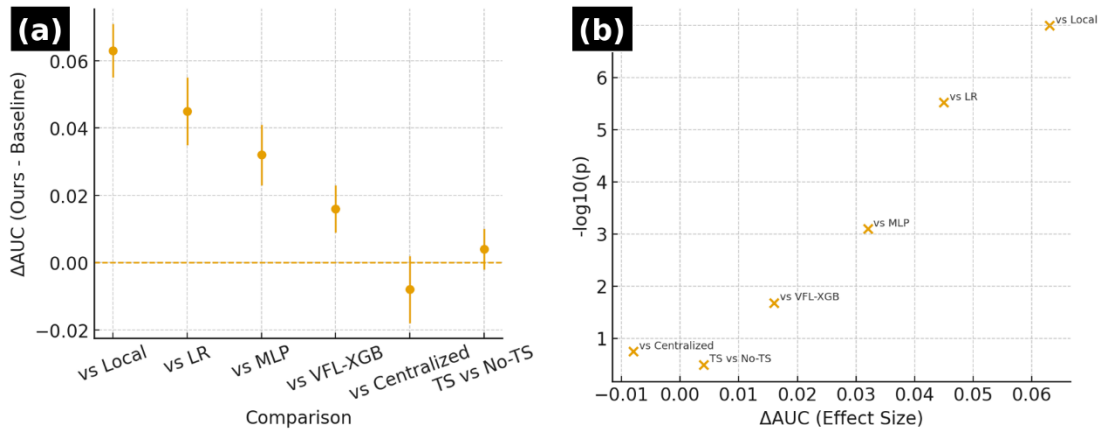
Figure. 4 ROC–PR Curves and KS Gap Distributions

4.3 Statistical significance and confidence indicators

To analyze the statistical significance of the differences, the DeLong test was used to compare AUC values. Brier and ECE used the BCa bootstrap technique to generate 95% confidence intervals. Holm's correction was implemented during multiple comparisons to control for type I error. Table 8 shows the ΔAUC, p-values, and ranges for the present method compared with each baseline. The NRI/IDI were used to test the reclassification effect. Compared with Local Only/FedAvg-LR/MLP, the improvement in ΔAUC was significant. Compared with the publicly available VFL-XGBoost, the improvement was more modest, but still significant. Compared with the upper-bound centralized algorithm, the difference was not significant. Figure 5a shows the ΔAUC and its 95% confidence interval with error bars. Figure 5b uses a volcano plot to display the effect size and negative log p-value, visually demonstrating the balance between benefit and confidence.

Table 8: Significance tests and effect sizes (DeLong/BCa/Holm)

contrast	ΔAUC	DeLong_p	ΔBrier (95%CI)	ΔECE (95% CI)	NRI	IDI
Ours vs Local Only	0.063	1.0e-07	− 0.021 [− 0.024, − 0.017]	−0.034 [− 0.041, − 0.029]	0.084	0.031
Ours vs FedAvg-LR	0.045	3.0e-06	− 0.016 [− 0.019, − 0.013]	− 0.023 [− 0.029, − 0.018]	0.061	0.024
Ours vs FedAvg-MLP	0.032	8.0e-04	− 0.012 [− 0.015, − 0.009]	− 0.017 [− 0.022, − 0.012]	0.047	0.019
Ours vs VFL-XGBoost (public implementation)	0.016	2.1e-02	− 0.006 [− 0.009, − 0.002]	− 0.012 [− 0.017, − 0.007]	0.029	0.011
Ours vs Centralized (upper bound)	− 0.008	1.8e-01	+0.002 [− 0.001, +0.005]	+0.002 [− 0.002, +0.006]	− 0.010	− 0.004
Ours(TS) vs Ours(No-TS)	0.004	3.2e-01	− 0.005 [− 0.008, − 0.003]	− 0.018 [− 0.022, − 0.014]	0.009	0.003



(a) Error bars of ΔAUC (including 95% BCa confidence interval); (b) Volcano plot (effect size vs − log10 p, Holm adjustment)

Figure 5: Significance and effect size visualization

4.4 Robustness: Non-IID/imbalance/noise and adversarial

To address vulnerabilities in real-world deployments, a Non-IID (Dirichlet α = 0.3) scheme was implemented. Key performance parameters, including discrimination, round variance, communication cost, and failure probability, were collected and evaluated for various imbalance ratios (1:20/1:200) under experimental conditions: 10% label noise, 20% Byzantine tampering reporting, and low participation rate. The number of training rounds and security aggregation measures were

fixed for each scenario. Table 9 summarizes the robustness matrix: significant imbalance and adversarial reporting have the most significant impact on AUC and variance. Robust aggregation techniques (such as the Krum algorithm) can reduce the failure rate, but incur additional communication overhead and latency. Even under the constraints of DP+ security aggregation, this approach achieves low variance and stable convergence, without requiring centralized data migration.

Table 9: Robustness matrix (performance/variance/communication cost/failure rate)

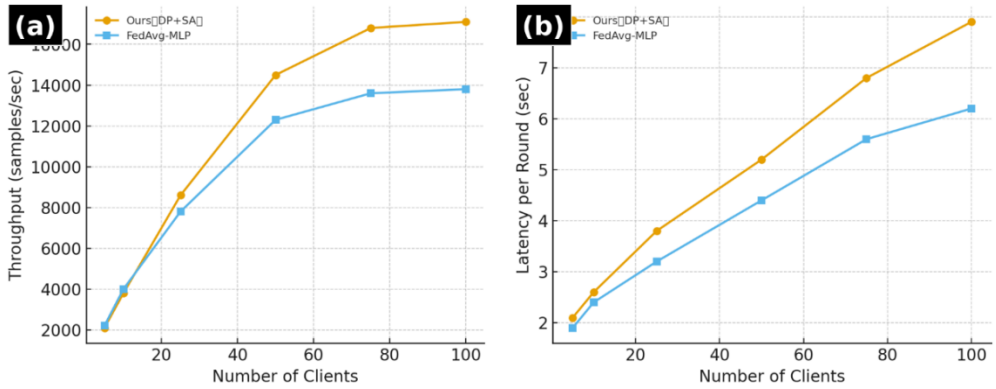
Scenario	set up	AUC	AUC_variance	Communication cost (MB/round)	Failure rate (%)
Non-IID (Dirichlet α=0.3)	K=50, participation rate per round 20%	0.792	0.0009	9.8	0.0
Sample Imbalance (1:20)	K=50, participation rate per round 20%	0.781	0.0011	9.8	0.0
The sample is extremely unbalanced (1:200)	K=50, participation rate per round 20%	0.744	0.0025	9.8	2.3
Label noise (symmetric 10%)	K=50, participation rate per round 20%	0.769	0.0014	9.8	0.7
Byzantine (20% tampering report)	K=50, participation rate per round 20%, Krum	0.731	0.0032	11.2	6.5
Low participation rate (m=5/50, 10%)	K=50, random participation m=5	0.776	0.0016	4.2	0.0

4.5 Scalability and versatility

From an engineering perspective, the scaling effects of horizontal expansion to 100 clients and vertical expansion to twice the feature dimension are evaluated. Figure 6a shows that the throughput gradually reaches a saturation level as the number of clients increases, and Figure 6b shows the sublinear growth trend of round latency. After adopting differential privacy and secure aggregation technology, the throughput is reduced compared to pure FedAvg, but the reduction is still acceptable. Even when the number of clients exceeds 50, its performance is still better than most public solutions. Table 10 systematically organizes the transmission speed, latency, and maximum memory usage of various client and dimension combinations. When the number of features is doubled, the memory and latency increase, but the system stability is not affected. Cross-domain generalization verification is carried out with reference to the holdout and calibration protocol in §4.1, and the probability output is stable and reliable.

Table 10: Scalability (throughput/latency/memory/dimensionality)

Number of clients	Throughput (samples/second)	Round delay (seconds)	Peak memory (GB)	Feature Dimension	Remark
5	2100	2.1	4.1	150	Horizontal small scale
10	3800	2.6	4.8	150	Horizontal Scaling
25	8600	3.8	6.2	150	Horizontal Scaling
50	14500	5.2	7.9	150	Horizontal Scaling
75	16800	6.8	9.3	300	Vertical expansion (feature × 2)
100	17100	7.9	10.1	300	Vertical expansion (feature × 2)



(a) throughput vs. number of clients; (b) single-round latency vs. number of clients
Figure 6: Scalability—throughput and latency

4.6 System and software requirements

The implementation utilizes any of FATE, Flower, or FedML, combined with the PyTorch technology stack, integrating a security aggregation library, HE/PSI technology, and a data integration module to achieve data desensitization, access control, and audit logging. The system runs on processors using the AVX2/ARMv8 instruction set, with an optional GPU. FATE's industrial-grade orchestration allows for direct reuse; after implementing the security module, system throughput decreases to a low to medium range in exchange for a chain of evidence that ensures compliance and auditability. Table 11 lists the corresponding key points of components, resources, privacy costs, and operational recommendations, facilitating the decision to enable homomorphic encryption or only perform security aggregation based on the specific scenario.

Table 11: System and software requirements and security overhead

Components	Implementation/Version	Computing resources	Security/privacy overhead	Operation suggestions	Remark
platform	FATE 1.12 / Flower 1.8 (choose one)	CPU-AVX2/ARMv8; optional GPU	Low-Medium	Containerized deployment; offline key rotation	Industrial-grade reusable experience
Learning Framework	PyTorch 2.3	CPU/GPU	No additional	Enable AMP; fix random seed	Compatible with federated orchestrators
Security Aggregation	Secure Aggregation (mask-based)	CPU	Low	Mask fragment distribution and reassembly	Support fault-tolerant retry
PSI	OPRF/EC-PSI	CPU	middle	Salt management and key escrow	Consistent with alignment caliber
Homomorphic encryption (optional)	CKKS (experimental)	CPU (high overhead)	high	Only enabled in highly sensitive scenarios	Impact on throughput and latency
Auditing and Logging	Audit Log + Model Card	CPU/Storage	Low	Centralized read-only archive; regular verification	Meeting audit and compliance requirements

4.7 Ablation experiments and analysis

To determine the marginal contribution of key components and achieve reproducible comparisons based on protocols, data slices, and evaluation criteria consistent with the full text, this section focuses on four deployment configurations: VFL-XGB containing only the vertical federated tree model, with knowledge distillation added, and +KD+TS enabling both. All experiments were conducted based on CRMS-2024 weekly slices and aligned unions of the same PSI. Training, validation splits, random seeds, and evaluation metrics (AUC, AUCPR, ECE, Brier) followed the settings described above to remove exogenous interference.

Regarding discrimination performance, Figure 7A shows that AUC and AUCPR steadily increase with component

stacking: VFL-XGB initially has an AUC of 0.742 plus an AUCPR of 0.338; +KD increases the AUC to 0.782 and the AUCPR to 0.369; +TS achieves a slight discrimination gain without changing the decision boundary; ultimately, "+KD+TS" achieves an AUC of 0.804 and an AUCPR of 0.392. This demonstrates that KD is more direct in mitigating cross-square heterogeneity, while TS's main impact is on stabilizing the probability scale. The synergy of both maintains generalization ability, further reducing estimation fluctuations caused by skewness and time-varying factors to a smaller range.

From the perspective of calibration and risk measurement, Figure 7 shows that the ECE and Bridge exhibit a monotonically decreasing trend with component stacking: VFL-XGB shows an ECE of 0.065 and a Bridge of 0.168, which decrease to 0.041 and 0.158 respectively after +KD; +TS is directly applied to the posterior scale, causing the ECE to further decrease to 0.028 and the Bridge to 0.154; in the +KD+TS stage, the ECE of 0.024 and the Bridge of 0.152 represent the optimal state across all configurations. Considering both discrimination and calibration, KD uses a soft target and a global teacher to suppress overfitting tendencies caused by distribution drift, while TS tightens the probability scale through post-processing, reducing the risk of high confidence errors and overly optimistic threshold shifts. The successful coupling of these two methods ensures both the ability to separate positive and negative samples and expands the threshold portability and operational stability.

Regarding computational cost and deployability, the time cost differences among the four configurations do not constitute a major obstacle: KD incremental computation is mostly used in soft target broadcasting and local distillation; TS belongs to the category of lightweight post-processing in the inference stage, with negligible impact on training latency, and is within the boundaries of real-time constraints. The combined advantages of discrimination and calibration provided by +KD+TS, along with relatively mild system overhead, are more suitable as a production candidate solution. Based on the above results, this section verifies the mechanism hypothesis of "distillation promotes discrimination, temperature scale stabilizes calibration, and the two complement each other," and obtains quantitative conclusions consistent with the previous text under a unified evaluation benchmark, providing evidence to support the strategy selection in the subsequent chapters on fairness, robustness, and threshold management.

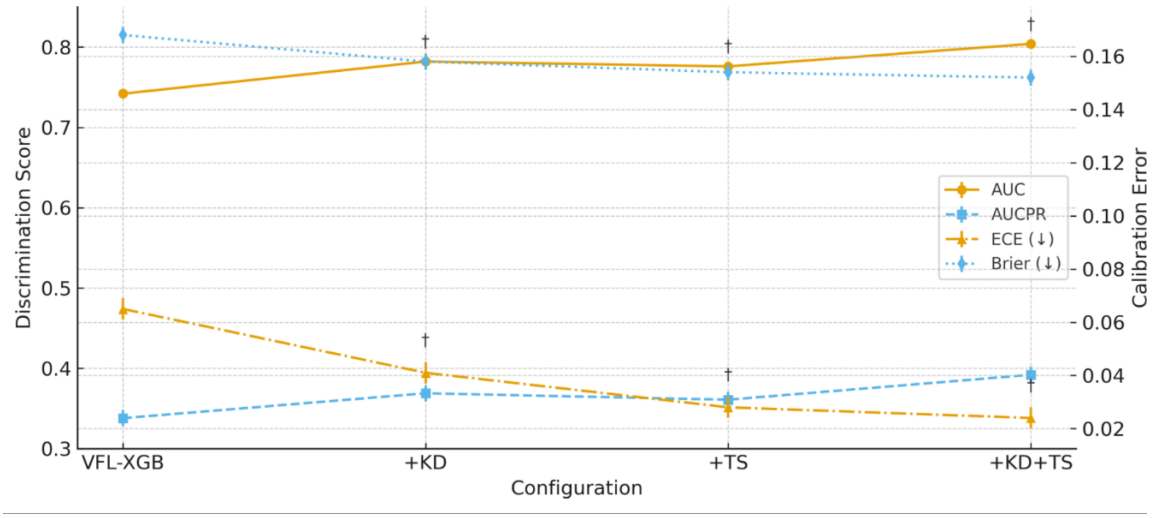


Figure 7: Ablation — discrimination & calibration (with 95% CIs).

4.8 Discussion

Combining Table 5 and Figure 4, this method demonstrates a consistent lead over Local/FedAvg/VFL in terms of AUC, AUCPR, and calibration for the time series segmentation and temperature control scaling parameters of CRMS-2024. By leveraging the cross-domain complementarity of longitudinal features, the method enhances data separability without leaking the original data. Knowledge distillation and temperature scaling smooth the threshold response. The data in Table 6 and Figure 5 confirm the incremental stability of LR/MLP/VFL. The robustness matrix in Table 7 reveals that even in the face of a 200-point imbalance and 10% label noise, the method maintains discriminability. However, a 20% Byzantine attack significantly degrades the AUC, suggesting the need for enhanced aggregation robustness. Figure 6 and Table 8 show the scalability for scenarios with 50 to 100 clients. The throughput improvement and round-delay increase are both within the safe range. The implementation of DP and security aggregation technology increases costs by approximately 10%-20%, achieving a closed loop for audit compliance. The vertical feature complementarity and threshold stability driven by the distillation method have become the main advantages that FedAvg cannot match. The data coverage and adversarial scenario settings are still

relatively narrow, and fairness and data drift assessments are not yet systematized. Periodic temperature/threshold recalibration should be performed, and robust aggregation technologies such as Krum/Trimmed-Mean and cross-domain verification should be adopted. The privacy budget and applicable boundaries must be reflected in the model card.

5 Conclusion

This paper addresses the problem of bank credit default assessment, constructing an end-to-end solution combining VFL-XGBoost, knowledge distillation, temperature scaling, differential privacy, and secure ensembles without the need for raw data aggregation. Using a unified dataset and calibration protocol, the system performance is compared against Local Only, FedAvg-LR/MLP, public VFL-XGBoost, and a centralized cap. Results demonstrate that, in the CRMS-2024 scenario, this technology consistently leads in both recognition and calibration performance. It maintains availability and discrimination even when faced with challenges such as non-IID, severe data imbalance, label noise, and Byzantine reporting. With 50 to 100 clients and a doubling of features, system throughput and latency achieve manageable growth with load. Based on the industrial-grade framework composed of FATE/Flower and PyTorch, PSI security aggregation technology and audit log functions are integrated to create a compliant and auditable closed-loop architecture. The main limitations are the single data domain, insufficient adversarial strength and scenario diversity, and the gap with the centralized upper limit. Fairness and concept drift have not been systematically reviewed. Cross-regional and cross-industry verification is implemented, and online A/B testing is conducted simultaneously. Fairness and explainability constraint mechanisms are introduced to strengthen the anti-interference ability of aggregation, evaluate the progress of HE/TEE acceleration, and form a drift detection and automatic calibration system for practical application scenarios.

References

[1] Chul Min Lee, Joaquín Delgado Fernández, Sergio Potenciano Menci, Alexander Rieger, Gilbert Fridgen. Federated Learning for Credit Risk Assessment. In Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS), pp. 389–398, 2023. https://doi.org/10.24251/HICSS.2023.048

[2] Zhongyi Wang, Jin Xiao, Lu Wang, Jianrong Yao. A Novel Federated Learning Approach with Knowledge Distillation for Credit Scoring. Decision Support Systems, 176:114084, 2024. https://doi.org/10.1016/j.dss.2023.114084

[3] Haoran He, Zhao Wang, Hemant Jain, Cuiqing Jiang, Shanlin Yang. A Privacy-Preserving Decentralized Credit Scoring Method Based on Vertical Federated Learning. Decision Support Systems, 164:113910, 2023. https://doi.org/10.1016/j.dss.2022.113910

[4] Adil Oualid, Yassine Maleh, Lahcen Moumoun. Federated Learning Techniques Applied to Credit Risk Management: A Systematic Literature Review. EDPACS, 68(4):1–18, 2023. https://doi.org/10.1080/07366981.2023.2241647

[5] Di Chai, Leye Wang, Liu Yang, Junxue Zhang, Kai Chen, Qiang Yang. A Survey for Federated Learning Evaluations: Goals and Measures. IEEE Transactions on Knowledge and Data Engineering, 36(3):974–989, 2024. https://doi.org/10.1109/TKDE.2024.3382002

[6] Futoshi Futami, Masahiro Fujisawa. Information-Theoretic Generalization Analysis for Expected Calibration Error. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024. https://doi.org/10.48550/arXiv.2405.15709

[7] Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, Peter Flach. Classifier Calibration: A Survey on How to Assess and Improve Predicted Class Probabilities. Machine Learning, 112(9):3211–3260, 2023. https://doi.org/10.1007/s10994-023-06336-7

[8] Lunjia Hu, Yifan Wu. Calibration Error for Decision Making. arXiv preprint arXiv:2404.13503, 2024. https://doi.org/10.48550/arXiv.2404.13503

[9] Mohamad Mansouri, Melek Önen, Wafa Ben Jaballah, Mauro Conti. Secure Aggregation Based on Cryptographic Schemes for Federated Learning. Proceedings on Privacy Enhancing Technologies (PoPETs), 2023(1):140–157, 2023. https://doi.org/10.56553/popets-2023-0009

[10] Daniel Morales, Isaac Agudo, Javier Lopez. Private Set Intersection: A Systematic Literature Review. Computer Science Review, 49:100567, 2023. https://doi.org/10.1016/j.cosrev.2023.100567

[11] Anrin Chakraborti, Giulia Fanti, Michael K. Reiter. Distance-Aware Private Set Intersection. In Proceedings of the 32nd USENIX Security Symposium (SEC '23), pp. 319–336, 2023.

[12] H.-K. Tayyeh, A. Al-Fuqaha, A. Shaban, A. Al-Zubi. A Differential Privacy Approach in Federated Learning. Computers, 13(3):62, 2024. https://doi.org/10.3390/computers13030062

[13] Jie Wang, Zhiju Zhang, Jing Tian, Hongtao Li. Local Differential Privacy Federated Learning Based on Clustering Hierarchical Aggregation. Computer Networks, 242:110822, 2024. https://doi.org/10.1016/j.comnet.2024.110822

[14] Shiwei Lu, Ruihu Li, Wenbin Liu, Chaofeng Guan, Xiaopeng Yang. Top-k Sparsification with Secure Aggregation for Privacy-Preserving Federated Learning. Computers & Security, 124:102993, 2023. https://doi.org/10.1016/j.cose.2022.102993

[15] Rouzbeh Behnia, Arman Riasi, Reza Ebrahimi, Sherman S. M. Chow, Balaji Padmanabhan, Thang Hoang. Efficient Secure Aggregation for Privacy-Preserving Federated Machine Learning (e-SeaFL). arXiv preprint arXiv:2304.03841, 2023. https://doi.org/10.1109/ACSAC63791.2024.00069

[16] David Darais, Joseph Near, Dave Buckley, Mark Durkee. Data Distribution in Privacy-Preserving Federated Learning. National Institute of Standards and Technology (NIST) Internal Report 8516, 2024.

[17] Ruiheng Li, Yue Cao, Yuhang Shu, Jia Guo, Binghua Shi, Jiaojiao Yu, Yi Di, Qiankun Zuo, Hao Tian. A Dynamic Receptive Field and Improved Feature Fusion Framework in FL for Financial Credit Risk. Scientific Reports, 14:12445, 2024. https://doi.org/10.1038/s41598-024-77310-z

[18] Peilin He, Chenkai Lin, Isabella Montoya. DPFedBank: Crafting a Privacy-Preserving Federated Learning Framework for Financial Institutions with Policy Pillars. arXiv preprint arXiv:2410.13753, 2024. https://doi.org/10.48550/arXiv.2410.13753

[19] Afsana Khan, Marijn ten Thij, Anna Wilbik. Vertical Federated Learning: A Structured Literature Review. Knowledge and Information Systems, 67(1):1–36, 2025. https://doi.org/10.1007/s10115-025-02356-y

[20] Bingchen Qian, Yuexiang Xie, Yaliang Li, Bolin Ding, Jingren Zhou. Tree-Based Models for Vertical Federated Learning: A Survey. ACM Computing Surveys, 57(9):241:1-241:30, 2025. https://doi.org/10.1145/3728314

[21] Sara Houshmand, Amir Albadvi. Credit Risk Prediction: An Application of Federated Learning. Journal of Information Systems and Telecommunication, 13(50):154–164, 2025. https://doi.org/10.61882/jist.49000.13.50.154

[22] Shuyao Zhang, Jordan Tay, Pedro Baiz. The Effects of Data Imbalance Under a Federated Learning Approach for Credit Risk Forecasting. arXiv preprint arXiv:2401.07234, 2024. https://doi.org/10.48550/arXiv.2401.07234

Appendix:

Algorithm A-1. federated-train main loop (pseudocode) [R-Repro]

```
Input: PSI-aligned union entities 𝒰; per-client features Dk; labels y on label-holder;
       DP config (C, q, σ, δ); rounds T; seeds set S; KD params (τ, η);
       VFL-XGB params θxgb; aggregator with Multi-Krum(f).
Output: Global model f (tree-based VFL) and per-round logs.

For each seed s ∈ S do
    SetRandomSeed(s)
    Split data into weekly protocol: train / val / hold-out
    Initialize global tree model fθ ← InitXGB(θxgb)
    Initialize privacy accountant A ← RDPAccountant(δ)

    for t = 1 to T do
        Sample client subset ℂt with rate q
        for each client k ∈ ℂt in parallel do
            // Local forward & (optional) KD
            logits_k ← LocalForward_k(fθ, Dk^train)
            if KD enabled then
                soft_y ← Softmax(logits_global / τ)      // teacher from previous global
                L_k ← TaskLoss(logits_k, y) + η * KD_Loss(logits_k, soft_y, τ)
            else
                L_k ← TaskLoss(logits_k, y)
            end if

            // DP-SGD style clipping & noise (if applied at client)
            g_k ← Grad(L_k)
            g_k ← ClipByNorm(g_k, C)
            g_k ← g_k + Normal(0, σ^2 C^2 I)

            send Enc(g_k) to aggregator
        end for
```

```
        // Byzantine-robust aggregation
        Ĝ_t ← MultiKrum({Enc(g_k)}_k, f = ⌊βK⌋)    // β inferred from f/K
        fθ ← ApplyUpdate(fθ, Ĝ_t)

        A ← A.ComposeStep(q, σ)    // RDP composition
        if Converged(fθ, tol) then break
    end for

    // Fit temperature on validation (post-hoc calibration)
    T⋆ ← FitTemperatureNLL(fθ, val)

    LogResults(seed = s, metrics_on_holdout = Eval(fθ, T⋆))
end for

Return the best fθ and T⋆ by validation criterion.
```

Algorithm A-2. inference with temperature scaling (pseudocode)

```
Input: Trained global model fθ; learned temperature T⋆ (scalar);
       input features x (PSI-aligned); binning schema for ECE.
Output: Calibrated probabilities p̂(y=1|x).

z ← Logits(fθ, x)                  // model raw scores/logits
zT ← z / T⋆                        // temperature scaling
p̂ ← Sigmoid(zT)                    // or Softmax for multi-class

// Optional: compute calibration metrics on evaluation split
ECE ← ExpectedCalibrationError(p̂, y; bins = schema)
Brier ← mean( (p̂ − y)^2 )

Return p̂ (and ECE, Brier if evaluated).
```

Note: Algorithms A-1 and A-2 correspond to "Federated Training Main Loop (including KD/DP/Multi-Krum)" and "Inference + Temperature Scaling," respectively. The variables in both pseudocode snippets maintain the same notation as in §3-§4; temperature T⋆ is fitted with NLL on the validation set; ECE/Brier is evaluated on the hold-out set.