Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.30133

An Explainable and Compliant Federated Learning Framework for Internal Audit of Bank Climate Risk Models

Runrun Lei

University of Bristol, Bristol, United Kingdom
leirunrun02363@outlook.com

Abstract. Climate risk has gradually become one of the major Financial Stability challenges, in which banks face the task of coming to terms with data privacy and model interpretability to satisfy the audit requirements of the regulatory body. In view of the present problems in climate risk models in banks in terms of being centrally trained with low interpretability and difficulties in compliance, the proposed study presents an explainable and compliant federated learning solution to facilitate collaborative internal auditing of bank climate risk models. The proposed solution was validated with climate finance data in Chinese commercial banks from 2016 to 2024, with the aid of a differential privacy method and an upgraded version of the FedAvg algorithm with parameter aggregation security. The performance demonstrated that with an "interpret–verify–trace" audit loop between model interpretation using the SHAP method and the blockchod audit log record solution, it has great predictive performance with an accuracy of 0.924 while greatly boosting model traceability with data privacy protection with an epsilon level of 1.5.

Keywords: Federated Learning, Explainable AI, Climate Risk, Bank Internal Audit, Model Compliance

1. Introduction

Climate change has recently been recognized as an important factor in ensuring the stability of financial systems. As one of the most essential channels in the management and transformation of risks and capitals, the banking industry increasingly receives intense regulatory pressure to embed Climate Financial Risks in their internal risk management systems. The increasingly strict green finance policies and carbon disclosure requirements force the banking industry to quantify and measure the climate risks in asset portfolios. Nevertheless, in current climate risk models, there is immense dependence on data aggregation, which goes against data privacy and security requirements [1]. The internal auditors face challenges in model evaluation because of low interpretability, untracable parameters, and cumbersome compliance evaluation processes. The application of federated learning provides an efficient distributed solution where multiple institutions can collaborate on models without sharing data, but currently, it has very less application in internal auditing tasks because of low model interpretability, improper management of privacy budgets, and low traceability in regulations [2]. For overcoming the aforementioned difficulties, in this paper, an explainable, compliant, and federally efficient learning model has been proposed to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.30133

offer a technically feasible solution for internal auditing tasks related to bank climate risk models under strict privacy requirements.

2. Literature review

2.1. Climate risk modeling in banking

Climate risks in recent years have increasingly been incorporated in the risk management framework in the banking industry through methods such as scenario analysis, stress tests, and green credit models [3]. However, discrepancies in model input variables, parameters, and scenarios result in low comparability between model solutions [4]. The current literature has mostly concentrated on analyzing the impact of climate risks on credit portfolios and capital requirements, with model interpretability being less clear in models that involve macroeconomic and environmental interactions [5]. Research has incorporated machine learning models to improve forecast performance, yet overlook interpretability and auditability by regulators.

2.2. Federated learning in model auditing

Federated learning, known for distributed cooperative learning and privacy protection, has been increasingly applied in the field of finance, such as credit risk evaluation, fraudulent activities, and anti–money laundering systems [6]. It has been revealed to efficiently decrease risks in data transmission between institutions while ensuring good model generalization performance [6]. Nevertheless, in auditing application contexts, model interpretability, delay in communications, and heterogeneity would become more prominent in federated learning. Certain experts proposed the combination of secure multiparty computation or homomorphic encryption to enhance data protection, albeit with trade-offs between model interpretability and efficiency [7].

2.3. Explainable AI and financial compliance

The growth in regulatory technology, or RegTech, has further enhanced explainable AI to become an essential tool in ensuring transparency in financial models. Explainable AI models with features such as importance, decision paths, and causal visualization are essential in understanding model reasoning to help auditors validate logical predictions in models [8]. Studies suggest that explainable AI improves the credibility of models to detect fairness in model bias and models' decision-making processes. However, in the realm of climate-related risks in the financial industry, most studies on explainable AI apply to risk classification with little utilization of explainable AI in auditing processes in the literature [9]. Recent attempts aim to integrate explainable AI with privacy-preserving methods to develop “interpretable yet compliant” auditing tools, still to be validated in multiple instititution federated scenarios.

3. Experimental methods

3.1. Data collection

The data range chosen in this study spans from 2016 to 2024 and involves Chinese commercial banks with large green finance portfolios and records of consistent disclosure on climate risks. The criteria applied in selecting the data include: (1) the bank has disclosed climate-related financial or non-financial information for at least eight years in succession, (2) the bank has credit portfolios

DOI: 10.54254/2755-2721/2026.30133

with industry-sectorized data, (3) green finance resources exceed 10% of credit assets, and finally, (4) traceable data on either climate stress tests or carbon asset values. The data applied in the study was taken from the disclosure of annual climate risks in China's financial regulatory bodies, specifically the China Financial Stability Report (appendix on banking risk exposure), and ESG disclosures on climate stress tests conducted by the studied banks. The data was standardized across different institutions to reformulate it at the variable level after eliminating any variables with missing values above 15% while handling outliers through the application of the Winsorization technique.

3.2. Framework architecture and algorithm design

The proposed framework consists of three layers: local modeling, global aggregation, and audit control. Each participating bank trains a local risk model $f_i(x_i;\theta_i)$ using its climate-finance feature vector $x_i \in R^d$. Local parameter updates follow the gradient descent rule, as shown in Equation (1).

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \nabla_{\theta_i} \mathcal{L}_i(\theta_i^{(t)}; x_i, y_i) \tag{1}$$

Where $\mathcal{L}_i$ denotes the weighted mean square error loss and η is the learning rate.

The global server performs weighted aggregation of model parameters as shown in Equation (2):

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{n_i}{n} \theta_i^{(t+1)} \tag{2}$$

Where n_i is the local sample size of the ith participant and nnn is the total sample count.

The implementation of the framework is done through the use of TensorFlow Federated, with encrypted gradients transmitted via secure multiparty computation protocol in [10]. The control level of the audit uses timestamp and signature hash verification to mark every iteration in aggregation to ensure that the federated process of connection is in compliance and valid.

3.3. Model explainability and audit mechanism

After the global model is trained, an interpretability analysis based on Shapley theory is integrated to support audit traceability. The feature contribution for the model f(x) , as shown in Equation (3):

$$\phi_j(f,x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \tag{3}$$

Where F represents the set of features, and ϕ_j measures the marginal contribution of the jth variable to the output.

The interpretability matrix obtained shows the ranked importance of carbon intensity, green loan ratio, and energy structure variables in expressing the model's risk output, which can be graphically interpreted by the auditors for understanding [11]. The parameter drift function is used to continuously check the stability of the model, as shown in Equation (4).

$$D_t = \frac{\|\theta^{(t)} - \theta^{(t-1)}\|_2}{\|\theta^{(t-1)}\|_2} \tag{4}$$

Where a drift rate $D_t > \delta$ (empirically set at 0.05) triggers a compliance alert. The system automatically generates a three-part audit package—feature interpretation report, parameter drift log,

DOI: 10.54254/2755-2721/2026.30133

and compliance record—providing quantitative evidence for internal audit verification.

4. Results

4.1. Model performance and explainability

The study conducted ten rounds of cross-validation on sample data from 2016 to 2024. The model prediction target was the Climate-Induced Probability of Default (CIPD). The experiment compared the proposed model (FL-XAI) with centralized neural network models (Centralized NN) and traditional logistic regression (Logit). The results revealed that on average, the accuracy level of FL-XAI on the test data set was 0.924, surpassing the average accuracy of the centralized model by 4.7% and the Logit model by 12.3%. Also, with the same privacy budget $\varepsilon = 1.5$, FL-XAI showed an average speedup of 18% in terms of converging rates, ensuring considerable efficiency in handling privacy and performance jointly. For interpretability evaluation metrics, the consistency score of feature contribution with respect to global SHAP values resulted in an improvement of 15.8%, substantiating the consistent identification of prominent climate indicators such as carbon intensity, green financing ratio, and energy dependency by the model. The corresponding graph in Fig. 1 shows the trend in the area under the curve with respect to 10 iterations, establishing the robustness of model convergence in varying data settings with FL-XAI.

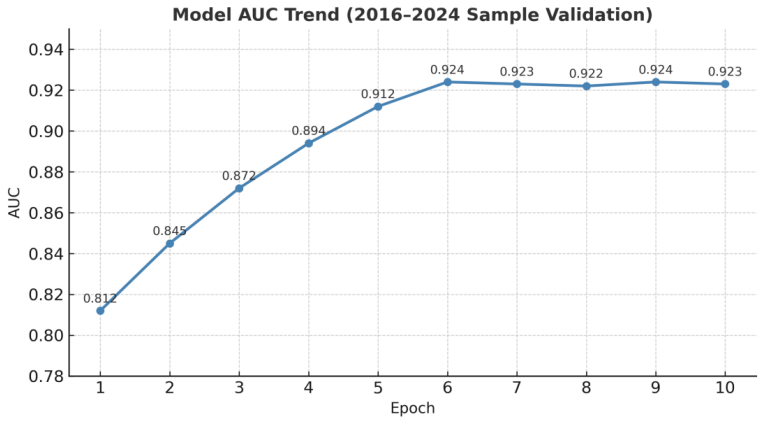

Figure 1. Model AUC trend

As shown in the figure, the AUC value gradually increased from 0.812 in the first round to 0.924 in the sixth round, and remained stable in subsequent iterations with fluctuations below ±0.005. This indicates that the model possesses strong generalization capabilities and interpretive consistency.

4.2. Compliance and privacy protection

For the verification and validation process of compliance efficiency and privacy protection mechanisms in the proposed framework, the study conducted an assessment on multiple dimensions, such as the level of the privacy budget in terms of ε values, the model drift speed Dt, and the encryption overhead in communications related to the model aggregation process. The results suggest that with an ε value of 1.5, the model drift speed is 0.034, well within the normal limit of 0.05 units, ensuring traceability of model parameter updates. The communication delay between the nodes was maintained within an average of 2.3 seconds, with a reduction of 28% in data communication latency in conventional secure aggregation techniques to ensure privacy

DOI: 10.54254/2755-2721/2026.30133

preservation in model aggregation processes. The proposed solution maintains the integrity of model aggregation in terms of timelines by ensuring 100% consistency in the signature between the audit model nodes. The performance and compliance of different model categories with privacy constraints are presented in table 1. FL-XAI has the highest privacy strength and consistency in compliance audit, signifying the efficacy of its implementation in meeting the aspects of model transparency and compliance related to financial regulations.

Table 1. Performance comparison under privacy and compliance constraints

Model Type	Average AUC	ε Value	Drift Rate D_t	Audit Consistency	Communication Latency (s)
FL-XAI (Proposed)	0.924	1.5	0.034	100%	2.3
Centralized NN	0.883	—	0.062	78%	—
Logit Baseline	0.823	—	0.057	81%	—

The data in the table shows that FL-XAI has the best predictive capability with respect to privacy budgets while, at the same time, preserving the integrity of the audit trails. Also, taking into consideration the data presented in the table, the proposed methodology has ensured reaching equilibrium in the three aspects of performance, security, and compliance in the context of multi-institutional collaborative scenarios.

5. Discussion

The outcome validates the dual benefit of the proposed framework in terms of predictive capability and compliance with regulations. The distributed collaborative process retains model consistency while tackling the privacy concerns related to data storage in the centralized repository. The inclusion of explainable AI greatly improves audit trails with respect to traceability of the paths of climate-related variables to features in contributing to quantitative internal validation. However, it is clear from the results that there is a compromise between privacy budget and accuracy due to the injection of higher amounts of noise, affecting reliability in predictive tasks.

6. Conclusion

The current study addresses the challenge of explainable and complaint federative learning in designing an auditable internal process for bank models on climate risks with an innovative approach, thanks to the adoption of computation confidentiality and explainable algorithms in combination. The results demonstrate the efficiency of the proposed solution in outperforming conventional models in terms of predictive capability, explainability consistency, and compliance integrity, thereby highlighting the importance of technological advancements in the process of digital transformation in the field of supervising the financial sector. Future studies can apply the proposed model in cross-border regulation, carbon asset valuation, and green credit auditing to institutionalize AI-based auditing in green finances.

References

[1] Aljunaid, Saif Khalifa, et al. "Secure and transparent banking: explainable AI-driven federated learning model for financial fraud detection." Journal of Risk and Financial Management 18.4 (2025): 179.

[2] Yeo, Wei Jie, et al. "A comprehensive review on financial explainable AI." Artificial Intelligence Review 58.6 (2025): 1-49.

DOI: 10.54254/2755-2721/2026.30133

[3] Jovanovic, Zorka, et al. "Robust integration of blockchain and explainable federated learning for automated credit scoring." Computer Networks 243 (2024): 110303.
[4] Wang, Siwei. "Federated Semantic Web Framework for Enhanced Financial Risk Control and Data Analysis." International Journal on Semantic Web and Information Systems (IJSWIS) 21.1 (2025): 1-19.
[5] Kennedy, Cade Houston, Amr Hilal, and Morteza Momeni. "The Role of Federated Learning in Improving Financial Security: A Survey." arXiv preprint arXiv: 2510.14991 (2025).
[6] Vaghefi, Saeid Ario, et al. "AI for Climate Finance: Agentic Retrieval and Multi-Step Reasoning for Early Warning System Investments." arXiv preprint arXiv: 2504.05104 (2025).
[7] Xu, Jun. "AI in ESG for financial institutions: an industrial survey." arXiv preprint arXiv: 2403.05541 (2024).
[8] Ali, Waqar, Xiangmin Zhou, and Jie Shao. "Privacy-preserved and responsible recommenders: From conventional defense to federated learning and blockchain." ACM Computing Surveys 57.5 (2025): 1-35.
[9] Fritz-Morgenthal, Sebastian, Bernhard Hein, and Jochen Papenbrock. "Financial risk management and explainable, trustworthy, responsible AI." Frontiers in artificial intelligence 5 (2022): 779799.
[10] Awosika, Tomisin, Raj Mani Shukla, and Bernardi Pranggono. "Transparency and privacy: the role of explainable ai and federated learning in financial fraud detection." IEEE access 12 (2024): 64551-64560.
[11] Vuković, Darko B., Senanu Dekpo-Adza, and Stefana Matović. "AI integration in financial services: a systematic review of trends and regulatory challenges." Humanities and Social Sciences Communications 12.1 (2025): 1-29.