FATE-Compliant ML Architecture with Blockchain-Verifiable Auditing: A Governance Framework for Ethical Compliance in FinTech

Samah S. Kareem
Isik University
Istanbul-Turkey
21comp9003@isik.edu.tr

Abstract

Concerns about fairness, accountability, transparency, and ethics (FATE) have intensified with the rapid adoption of artificial intelligence (AI) in financial technology (FinTech). Algorithmic lending systems risk entrenching socioeconomic and geographic disparities. To address these risks—and to align with emerging regulation such as the EU AI Act—we propose a multi-layer governance framework for credit-risk modelling that integrates (i) blockchain-verifiable auditing via smart contracts (Hyperledger Fabric); (ii) causal modelling with counterfactual, pathway-aware fairness quantified by the *Regional Inclusion Score*; (iii) a hybrid data pipeline that fuses local micro-data with controlled geospatial skew; and (iv) a fairness-aware gradient-boosted learner (FairXGBoost) embedded in a continuous monitoring loop. The smart-contract layer provides cryptographic auditability of data lineage, model updates, and alerts, operationalizing fairness as a regulatory constraint that triggers on-chain notifications when RIS declines. Across benchmark credit datasets with induced regional skew, the system maintained RIS at or above the regulatory threshold 0.85 while preserving competitive predictive performance, with automatic alerts when fairness degraded—demonstrating how causal analysis and ledger-backed governance can jointly yield reliable, regulator-verifiable AI for FinTech.

Introduction

FinTech AI's ethical imperative

Ethical flaws in FinTech's widespread use of AI are becoming more obvious, especially in algorithmic credit scoring. According to studies, applicants in economically disadvantaged regions continue to have approval rates that are 15 to 30% lower than those of their privileged counterparts (Olateju et al. 2024). Limited transparency in automated decision making increases these disparities and undermines public trust (Grimmelikhuijsen 2023). In line with the European Union AI Act, this work uses a compliance-first AI governance architecture to address these imperatives (European Commission, 2021). We ensure ethical compliance in high-stakes financial applications by going beyond statistical parity and capturing structural pathways of discrimination, building on causal fairness frameworks.

Limitations of Current Methods

Technical frameworks are still insufficient despite the importance of regulations. According to research, fairness pipelines have ongoing "technical debt" (Huang, Nourian, and Griest 2021), and current audit systems are criticized for lacking reliable and verifiable mechanisms (Singh 2023).

Auditability

According to (Singh 2023), post hoc audits alone are insufficient to guarantee regulatory compliance. We present blockchain-enabled auditability, which makes it possible to create immutable fairness logs that are securely encrypted (Aldboush and Ferdous 2023).

Regulatory Agility

It's still difficult to adjust dynamically to changing AI regulations. Agile compliance systems are essential, according to (Olatoye et al. 2024). We go one step further by incorporating constraint mapping APIs for automated regulatory synchronization (Mbiazi et al. 2023).

Methodology

Our methodology integrates causal explainability, a fairness-aware learning objective, and blockchain-verifiable auditing to enforce compliance in FinTech credit scoring. We first formalize a pathway-aware fairness metric, the *Regional Inclusion Score* (RIS), then describe data preparation, model architecture, training, and the auditing layer.

Causal Explainability Layer

We implement path-specific, counterfactual reasoning to isolate potentially unfair pathways (e.g., *Region* $\rightarrow$ *Income* $\rightarrow$ *Approval*) following prior work on counterfactual fairness (Chickering 2002). For an applicant with non-geographic covariates X_{-A} and region $A \in \{\text{Gaza}, \text{WestBank}\}$, we compare counterfactual predictions under interventions on A:

$$\widehat{Y}^{A\leftarrow\text{Gaza}}(X_{-A}) \quad \textbf{vs.} \quad \widehat{Y}^{A\leftarrow\text{WestBank}}(X_{-A}),$$

where $\widehat{Y}^{A\leftarrow a}$ denotes the model prediction when intervening on A via the do$(\cdot)$ operator while holding X_{-A} fixed. Pathway blocking is applied in the structural model over $\{$*income*, *education*, *mobile banking*, *region*$\}$;

Regional Inclusion Score (RIS)

Definition: Let A denote region (protected or quasi-protected), Y the true label, and $\hat{Y}$ the model decision. We define

$$\text{RIS} = \frac{\Pr(\hat{Y} = 1 \mid A = \text{disadvantaged})}{\Pr(\hat{Y} = 1 \mid A = \text{advantaged})},$$

and report RIS alongside accuracy and AUC throughout. We treat compliance as a hard constraint with a threshold τ=0.85, consistent with our governance policy and fairness targets.

Model Architecture

We use Gradient Boosted Decision Trees (XGBoost) due to their ability to capture non-linearities, gracefully handle missingness, and expose feature importance signals aligned with causal attribution.

Training Protocol We augment binary cross-entropy with a fairness regularizer that penalizes deviation of the training RIS from the compliance target:

$$\min_{\theta} \mathcal{L}_{\text{CE}}(\theta) + \lambda \left\| \text{RIS}_{\text{train}}(\theta) - \tau \right\|^2, \qquad \tau = 0.85,\ \lambda = 0.5,$$

where θ are model parameters and $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy. This balances predictive performance and fairness; geospatial features (aggregates and interactions) support pathway-aware adjustments.

Blockchain-Verifiable Auditing

We deploy a permissioned ledger (Hyperledger Fabric) to provide immutable, regulator-verifiable monitoring. Chaincode records: (i) model version and training metadata (hashes of code/data, time, configuration); (ii) data lineage events; (iii) periodic fairness summaries (RIS, CIs) and drift indicators; and (iv) alerts when RIS $< \tau$. Each record includes a timestamp, signer identity, and cryptographic digests, enabling auditors to replay state and verify that remediation (e.g., retraining or throttling) was triggered upon violation. This operationalizes fairness as a governance constraint rather than a post-hoc metric.

Evaluation and Results

Experimental Setup

We evaluate on the augmented German Credit dataset with regional annotations distinguishing *West Bank* (reference) and *Gaza* (protected). To create a measurable fairness gap, we inject a synthetic approval-rate penalty for the protected group, then measure fairness using RIS (Eq. 2) alongside standard predictive metrics.

Models are trained and assessed with stratified group k-fold cross-validation (we use k=5), ensuring balanced representation across regions. Unless stated, we report fold-averaged means with 95% bootstrap CIs (1,000 replicates). Significance for pairwise model comparisons (AUC and RIS) is assessed via paired bootstrap on fold predictions.

Baselines. (1) *Unconstrained XGBoost*: trained only for accuracy. (2) *Post-hoc Regional Penalty*: a simple evaluation-time baseline applying a fixed penalty at prediction. **Proposed**: regularized training objective in Eq. (3) with continuous monitoring.

Core Results

We observe the expected trade-off between accuracy and RIS. The baseline achieves the highest raw accuracy (82%) but violates the fairness requirement with RIS=0.63<0.85. Both fairness-aware variants raise RIS substantially; the proposed framework attains RIS=0.90 with competitive accuracy (79%), satisfying the compliance target (τ=0.85 from Eq. 2).

Model	AUC	Accuracy	RIS
Baseline XGBoost	0.86	0.82	0.63
FairXGBoost (eval variant)	0.85	0.81	0.86
Proposed framework	0.84	0.79	0.90

Table 1: Test performance (means; 95% CIs omitted for brevity).

The baseline exhibits larger geospatial false-positive-rate gaps (ΔFPR = 0.12) and higher counterfactual inconsistency (0.28); our framework reduces these to 0.04 and 0.12, respectively.

require(ris ¿= 85, "Unfair model detected"); emit FairnessAlert(ris, block.timestamp);

Conclusion and Future Work

We introduced a governance framework for ethical compliance in FinTech credit scoring that combines causal diagnostics, a fairness-aware learning objective, and blockchain-verifiable auditing. On an augmented credit dataset with induced regional skew, unconstrained models achieved strong accuracy but violated fairness requirements, whereas our framework met the compliance target for the Regional Inclusion Score (RIS; Eq. 2) with competitive predictive performance. The ledger-backed monitor provided real-time, auditable oversight by emitting alerts whenever RIS ¡ 0.85, thereby operationalizing fairness as a governance constraint rather than a post-hoc metric.

Limitations and External Validity. Our study relies on a *single* benchmark (German Credit) with *synthetically injected* regional skew. While this controlled setting is useful for mechanism testing, it constrains generalizability.

Future Work. (i) Real-world validations with the Palestine Monetary Authority and additional markets; (ii) adaptive, jurisdiction-aware thresholds rather than a fixed τ; and (iii) scaling the monitoring layer with privacy-preserving mechanisms suitable for production SLAs.

References

Aldboush, H. H. H.; and Ferdous, M. 2023. Building Trust in Fintech: An Analysis of Ethical and Privacy Considerations in the Intersection of Big Data, AI, and Customer Trust. *International Journal of Financial Studies*, 11(3): 90.

Chickering, D. M. 2002. Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3: 507–554.

Grimmelikhuijsen, S. 2023. Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision Making. *Public Administration Review*, 83: 241–262.

Huang, C.; Nourian, A.; and Griest, K. 2021. Hidden technical debts for fair machine learning in financial services. In *NeurIPS Workshop on Machine Learning in Finance*.

Mbiazi, D.; Bhange, M.; Babaei, M.; Sheth, I.; and Kenfack, P. J. 2023. Survey on AI Ethics: A Socio-Technical Perspective. arXiv preprint.

Olateju, O. O.; Okon, S. U.; Olaniyi, O. O.; Samuel-Okon, A. D.; and Asonze, C. U. 2024. Exploring the Concept of Explainable AI and Developing Information Governance Standards for Enhancing Trust and Transparency in Handling Customer Data. *Journal of Engineering Research and Reports*, 26(7): 244–268.

Olatoye, F. O.; Awonuga, K. F.; Mhlongo, N. Z.; Ibeh, C. V.; Elufioye, O. A.; and Ndubuisi, N. L. 2024. AI and Ethics in Business: A Comprehensive Review of Responsible AI Practices and Corporate Responsibility. *International Journal of Science and Research Archive*, 11(1): 1433–1443.

Singh, C. 2023. Artificial Intelligence and Deep Learning: Considerations for Financial Institutions for Compliance with the Regulatory Burden in the United Kingdom. *Journal of Financial Crime*, 30.