Mathematical Modelling of Engineering Problems
Vol. 12, No. 12, December, 2025, pp. 4314-4324
Journal homepage: http://iieta.org/journals/mmep

A Fairness-Aware and Bias-Resilient XAI Framework for Equitable Financial Decision-Making

Vydyam Krishna Aravinda[1*], Chigarapalle Shoba Bindu[2]

[1] Department of AI & DS, Chaitanya Bharathi Institute of Technology, Hyderabad 500075, India
[2] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapuramu 515002, India

Corresponding Author Email: vydyamaravinda@gmail.com

https://doi.org/10.18280/mmep.121220

Received: 29 August 2025
Revised: 15 November 2025
Accepted: 21 November 2025
Available online: 31 December 2025

Keywords:
fair decision-making, bias reduction, explainable models, financial fairness, federated learning, AI transparency, credit scoring fairness, loan approval models

ABSTRACT

Loan-approval prediction is typically considered while auditing a single protected attribute at a time (race, ethnicity, sex, or age). Fairness-Aware, Interpretable, Resilient, and Equitable (FAIRE) is a multi-stage pipeline that combines data-level balancing, in-training debiasing, and post-processing thresholding, which is complemented by global and local explainability and continuous fairness monitoring with a drift trigger. The evaluation spans centralized and federated training with privacy-preserving aggregation using Home Mortgage Disclosure Act (HMDA) loan-level data. At the selected operating point, fairness improves substantially: Demographic Parity (DP) rises from 0.74 [0.72, 0.76] to 0.92 [0.90, 0.94]; the Equal Opportunity (EO) gap declines to 0.05 [0.04, 0.06]; and Equalized Odds (EOdds) decreases to 0.07 [0.06, 0.09]. The change in Area under the curve—Receiver-operating characteristic curve (AUC-ROC) changes by ≤ 0.5 percentage points relative to the best utility setting. In the federated regime (50 clients, Non-Independent and Identically Distributed (non-IID) partitions), AUC-ROC remains within 1 percentage point of centralized utility, while fairness remains close to centralized post-mitigation levels (e.g., DP ≈ 0.90 [0.88, 0.92], EO ≈ 0.06 [0.05, 0.07], EOdds ≈ 0.11 [0.10, 0.12]), with wider intervals for clients with small protected-group support sample sizes. A composite Interpretability Score increases through higher surrogate fidelity, sparser reason sets, and more stable attributions; SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Integrated Gradients produce adverse-action-ready reason codes consistent with threshold-style explanations. The resulting pipeline delivers measurable fairness gains with minimal utility cost across centralized and federated settings while maintaining transparent, monitorable credit decisions.

1. INTRODUCTION

Artificial intelligence increasingly influences financial decisions, including loan approvals and credit scoring, yet historical and structural biases embedded in data and processes can propagate through predictive models, creating disparate outcomes across demographic groups [1-5]. Regulatory and ethical expectations further require decisions to be transparent and explainable, with reasoned, auditable narratives for adverse actions and with safeguards that balance fairness and predictive utility [5-7]. The task considered is loan-approval prediction with feature set X, protected attribute A (analyzed one at a time: race, ethnicity, sex, or age), and target Y (approval vs. denial). Fairness is assessed using Demographic Parity (DP), Equal Opportunity (EO), and Equalized Odds (EOdds).

Single-stage debiasing and post-hoc fixes are insufficient to ensure durable equity in operational settings because bias can re-enter through data shifts, proxy features, or inconsistent thresholds; deployment-time monitoring and governance are therefore required to maintain fairness without eroding predictive performance [2, 8, 9]. Beyond centralized modeling, privacy, regulatory, and organizational constraints motivate training regimes where data remain local and only model updates are aggregated, introducing additional fairness considerations under non-IID distributions [10, 11]. These needs call for end-to-end frameworks that integrate mitigation across the model lifecycle and connect technical interventions to compliance-oriented explanations and controls.

Fairness-Aware, Interpretable, Resilient, and Equitable (FAIRE) addresses these challenges as a multi-stage pipeline. At the data level, representation is balanced via reweighing and sampling; at the in-training level, adversarial debiasing and fairness-regularized objectives reduce dependence on sensitive attributes; at the post-processing level, calibrated, group-specific thresholds and reject-option rules align decisions with policy constraints. Transparency is supported by global and local explainability—permutation importance, interpretable surrogates, LIME, SHAP, and Integrated Gradients to generate reason codes suitable for adverse-action notices [6]. To accommodate organizational data boundaries, a federated learning (FL) regime enables client-local training

with privacy-preserving aggregation while auditing fairness at each round [6, 10].

The contributions of the study are listed below:

- End-to-end framework with numeric fairness reporting using DP, EO, and EOdds.
- Federated evaluation under non-IID client partitions with per-round support-weighted fairness aggregation [10].
- Composite Interpretability Score that synthesizes surrogate fidelity, explanation sparsity, and stability, grounded in model-agnostic and model-specific Explainable AI (XAI) [6].
- Reproducible baselines and statistical testing aligned to prior art, enabling credible comparisons under clearly specified protocols [5, 8, 11].

2. RELATED WORK

Research on algorithmic fairness in credit decisioning spans data-level, in-training, and post-processing interventions. Data-level methods include reweighing and oversampling to rebalance group–label distributions before optimization [1, 2, 5, 12-14]. In-training approaches commonly employ adversarial debiasing and fairness-regularized objectives to discourage representations predictive of protected attributes while preserving task performance [1, 2, 5, 8]. Post-processing techniques calibrate group-specific thresholds or introduce reject-option bands to reduce residual disparities after scoring [13-17]. Evaluation typically reports DP, EO, and EOdds alongside standard classification measures, enabling explicit fairness–utility trade-off analysis [1, 2, 5]. This study adopts those metrics and mitigation stages within a single pipeline to ensure consistency across training and deployment.

FL enables client-local training with periodic aggregation, reducing the need to centralize data while introducing fairness considerations under non-IID client distributions [10]. Fairness-aware FL variants combine local debiasing (e.g., adversarial objectives) with server-side aggregation of privacy-preserving statistics, aiming to maintain or improve group equity without sharing raw records [10]. Surveys also highlight the importance of weighting by client and group support to avoid biased global metrics when client cohorts differ in size or composition [5, 11]. These insights motivate the support-weighted aggregation of group rates used later to compute global DP/EO/EOdds and to trigger actions under the monitoring logic.

XAI underpins transparency and accountability in high-stakes credit decisions. Global analyses—permutation feature importance and interpretable surrogates—summarize model-level reliance on drivers of approval risk [6]. Local explanations such as LIME, SHAP, and Integrated Gradients provide instance-level attributions suitable for reason-code generation and adverse-action narratives in regulated settings [6, 18, 19]. The combination of global and local tools supports both system diagnostics and case-specific justification, forming the basis for the Interpretability Score later defined from fidelity, sparsity, and stability components.

Lifecycle-oriented frameworks emphasize sociotechnical governance, continuous fairness assessment, and human-in-the-loop review for model changes, threshold adjustments, and retraining [9, 20]. These perspectives align with fairness surveys that catalog bias sources and stress ongoing monitoring rather than single-shot mitigation [5]. The operationalization in this work—periodic audits, drift thresholds, and auditable policy layers—follows such governance guidance and ties directly to the monitoring construct formalized later [21, 22].

FAIRE integrates the three mitigation stages with XAI and monitoring, and is compared against baselines chosen for methodological coverage and reproducibility. Multiobjective evolutionary learning (MOEL) corresponds to multi-objective fairness optimization where predictive utility and fairness are jointly optimized [8]. Fairness-Aware Federated Learning (FAFL) reflects fairness-aware federated learning that applies client-local debiasing with aggregated coordination [10]. In addition, standard baselines—logistic regression, gradient-boosted trees, and neural networks with and without fairness interventions—anchor comparisons and support statistical testing consistent with survey recommendations on rigorous evaluation design [5, 11]. This mapping clarifies the role of each comparator and situates results within established lines of work.

Prior work demonstrates effective components—rebalancing data, adversarial objectives, calibrated thresholds; federated training with privacy-preserving aggregation; and global/local XAI for transparency—but these elements are seldom unified with deployment-time governance in a single, auditable framework [1, 2, 5, 6, 8-10]. FAIRE combines multi-stage mitigation with FL-aware auditing, explanation tooling, and lifecycle monitoring, leveraging the fairness metrics to quantify trade-offs and to guide operational controls.

3. METHODS AND MATERIALS

Let the dataset be $D = \{(X_i, A_i, Y_i)\}_{i=1}^{N}$, where X denotes predictive features, A denotes a single protected attribute considered per analysis (race, ethnicity, sex, or age), and Y denotes the loan-approval outcome (originated vs. denied), as defined in Eq. (1).

Two training regimes are employed (detailed in Section 3.1). The centralized track trains a single model on pooled data with fairness-aware objectives and post-processing as specified in Section 3.3. The federated track distributes training across clients that update local models and periodically aggregate parameters; both regimes feed the same bias-detection, explainability, and monitoring components described in Sections 3.2–3.5.

Fairness is evaluated using DP, EO, and EOdds, defined in Eqs. (5)-(7). These metrics are computed pre-training to characterize historical disparities, during training to guide mitigation, and post-training to quantify residual bias; corresponding results are reported in Section 4 alongside classification and interpretability outcomes.

3.1 Architecture overview

Figure 1 shows the FAIRE pipeline as an integrated sequence of modules: data ingestion and preprocessing, bias-mitigation layers (pre-, in-, and post-training), explainability, deployment and decisioning, and monitoring and governance. During ingestion, sensitive attributes A are separated from predictive features X in accordance with Eq. (1), and feature transformation procedures reduce proxy leakage while preserving predictive signal. Typical transformations include coarsening or binning highly

identifying variables, constrained encodings, and leakage checks to ensure that A does not re-enter the modeling stack through correlated surrogates.

$$D = \{(X_i, A_i, Y_i)\}_{i=1}^{N} \quad (1)$$

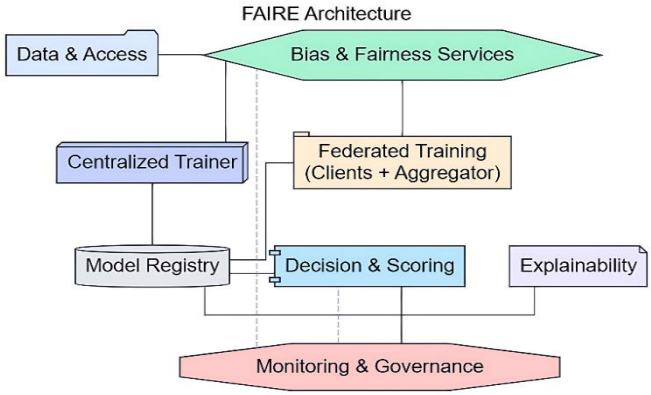

Figure 1. FAIRE framework pipeline

Centralized track: A single predictor f_θ is trained end-to-end with an adversary g_ϕ that attempts to infer A from the learned representation. The joint objective in formula (2) balances prediction loss against the adversary's loss via the trade-off parameter λ, encouraging f_θ to discard information predictive of A while retaining signal for Y. Data-level mitigation (reweighing/oversampling) may be applied before training, and a fairness regularizer may be included in the loss as specified later in Section 3.3. After training, group-specific operating thresholds and a reject-option band can be applied to the centralized scores following Eqs. (12) and (13).

Federated track: Data remain local to each client k. Each client trains a local predictor f_{θ_k} with a local adversary g_{ϕ_k} under the same objective as formula (2), using local data shards and identical preprocessing/mitigation rules. Model updates are transmitted to a coordinator that aggregates client parameters using FedAvg; unless otherwise specified, the protocol uses $T = 100$ communication rounds with $E = 2$ local epochs per round. Only model updates or parameters are shared—no raw records, explanations, or identifiers. After each aggregation, the coordinator evaluates global fairness diagnostics using privacy-preserving summaries and, when needed, recommends calibrated group-specific thresholds consistent with Eqs. (12) and (13). This two-track design enables consistent fairness controls irrespective of training regime while preserving locality of data.

$$\min_\theta \mathbb{E}\big[\mathcal{L}(f_\theta(X), Y)\big] - \lambda \cdot \mathbb{E}\big[\mathcal{L}_{adv}\big(g_\phi\big(f_\theta(X)\big), A\big)\big] \quad (2)$$

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

Explainability layer: Explanations are produced at global and local levels for either track. Global diagnostics include permutation importance and surrogate modeling (Eqs. (14) and (15)), supporting system-level audits of feature influence and fidelity. Local diagnostics include LIME, SHAP, and Integrated Gradients (Eqs. (16)-(18)), yielding per-applicant reason codes and ranked contributions suitable for compliance review and adverse-action communication. The explainability interface is identical across tracks and attaches to the scoring output, enabling uniform monitoring and reporting.

Deployment and decisioning: The scoring service outputs calibrated approval probabilities with accompanying explanation artifacts. Group-specific thresholds and reject-option rules are implemented as decision-policy layers on top of the raw scores, following Eqs. (12) and (13). Policy parameters are versioned and auditable, allowing subsequent monitoring to attribute fairness movements to either model updates or threshold adjustments.

Monitoring: Fairness and performance drift are tracked during validation and post-deployment. The trigger in formula (4) detects stepwise changes in DP across evaluation windows, summarizes alert states, recommended actions, and intervention history. Selection of the drift tolerance δ, alert routing, and human-in-the-loop review are specified in Section 3.5, ensuring consistent governance across centralized and federated operating modes.

$$|DP_t - DP_{t-1}| > \delta \quad (4)$$

Data ingestion separates A from X per Eq. (1); mitigation spans pre-training reweighing/oversampling, in-training adversarial debiasing (formula (2)), and post-processing thresholds (Eqs. (12) and (13)); explainability and deployment produce decisions with reason codes; monitoring applies drift checks (formula (4)). The federated branch depicts client k local training with adversary g_{ϕ_k}, rounds $t = 1 \dots 100$ with local epochs $E = 2$, and server-side aggregation $\mathcal{A} =$ FedAvg.

3.2 Bias detection

Metrics: DP, EO, and EOdds are adopted as in Eqs. (5)-(7). DP assesses parity of approval rates across groups in Eq. (5):

$$DP = \frac{P(\hat{Y} = 1 | A = 0)}{P(\hat{Y} = 1 | A = 1)} \quad (5)$$

EO measures parity of true-positive rates among qualified applicants in Eq. (6):

$$EO = P\big(\hat{Y} = 1 \big| Y = 1, A = 0\big) - P(\hat{Y} = 1 | Y = 1, A = 1) \quad (6)$$

EOdds aggregates disparities in both true-positive and false-positive behavior in Eq. (7):

$$EOdds = \big|P\big(\hat{Y} = 1 \big| Y = 1, A = 0\big) - P\big(\hat{Y} = 1 \big| Y = 1, A = 1\big)\big| + \big|P(\hat{Y} = 1 | Y = 0, A = 0) - P(\hat{Y} = 1 | Y = 0, A = 1)\big| \quad (7)$$

Federated fairness metrics: In the federated track, each client k computes group-wise confusion matrices and derives local approval rates and error rates. Only privacy-preserving summaries are shared: per-group supports and rate estimates (e.g., approval $s_{k,a} = \mathrm{TP}_{k,a} + \mathrm{FP}_{k,a}$, and total $s_{k,a} = P_{k,a} + N_{k,a}$). Server-side aggregation forms global metrics from support-weighted counts rather than averaging ratios. For example:

- $\mathrm{TP}_{k,a}, \mathrm{FP}_{k,a}, \mathrm{TN}_{k,a}, \mathrm{FN}_{k,a}$.
- $\Pr(\hat{Y} = 1 A = a) = \frac{\sum_{k=1}^{K} \text{approvals}_{k,a}}{\sum_{k=1}^{K} \text{totals}_{k,a}}$,

- $\mathrm{TPR}_a = \frac{\sum_{k=1}^{K} \mathrm{TP}_{k,a}}{\sum_{k=1}^{K} \mathrm{P}_{k,a}}, \mathrm{FPR}_a = \frac{\sum_{k=1}^{K} \mathrm{FP}_{k,a}}{\sum_{k=1}^{K} \mathrm{N}_{k,a}},$

where, TPR refers to True Positive Rate, and FPR indicates False Positive Rate.

Which are then combined into DP, EO, and EOdds using the definitions above. Logging stores only aggregated counts and rates to maintain privacy.

Algorithm 1: Federated Fairness Audit (per communication round)

Inputs: target DP band [0.80,1.25]; drift tolerance δ; definitions in Eqs. (5)-(7).
For each round t = 1…T:

1. Local computation on each client k
   - Compute $\mathrm{TP}_{k,a}$,$\mathrm{FP}_{k,a}$,$\mathrm{TN}_{k,a}$,$\mathrm{FN}_{k,a}$ for each group a. Derive local rates and local fairness metrics DP_k,EO_k,EOdds_k . Transmit only $\{\mathrm{TP}_{k,a},\mathrm{FP}_{k,a},\mathrm{TN}_{k,a},\mathrm{FN}_{k,a}\}$ (or equivalent counts).
2. Server-side aggregation
   - Form $\Pr(\hat{Y} = 1 \mid A = a)$, TPR_a , and FPR_a using support-weighted sums; compute global DP, EO, and EOdds.
3. Compliance checks and actions
   - Compare DP to the target band [0.80,1.25] (80% rule) and EO, and EOdds to near-zero targets (with uncertainty bands). Log results; if drift per formula (4)/Eq. (20) exceeds δ, trigger one of: no-op; threshold adjustment ΔT per Eq. (12); retraining; or reject-option handling per Eq. (13).

Evaluation process: Metrics are computed pre-training to characterize baseline disparities, during training per epoch (centralized) or per round (federated) to guide mitigation, and post-training to quantify residual bias. After deployment, the same metrics are monitored on rolling windows; drift detection follows Eq. (4) and thresholding logic in Section 3.5, with alerts routed to governance for review and action.

3.3 Bias mitigation techniques

Data-level (pre-training): Reweighing adjusts the empirical loss by assigning each instance a weight based on the joint distribution of the protected attribute and the label, as in Eq. (8). Define the group-label weight $w_{a,y} = \frac{\Pr(A=a)\Pr(Y=y)}{\Pr(A=a,Y=y)}$ and minimize a weighted loss in Eq. (8):

$$\mathcal{L} = \sum \mathrm{w_i} \cdot \mathcal{L}(\mathrm{f}(\mathrm{X_i}), \mathrm{Y_i}) \tag{8}$$

Oversampling complements reweighing by increasing minority-group support until approximate parity is achieved, targeting as in formula (9):

$$n_0 \approx n_1 \tag{9}$$

Synthetic sampling may be applied when raw counts remain insufficient, with guardrails: generate only within the convex hull of observed minority-group feature vectors; preserve label-conditional distributions; prohibit direct use of A or near-deterministic proxies during synthesis; perform synthesis on training folds only; and validate that downstream calibration and ranking are not distorted by synthetic artifacts.

Model-level (in-training): Adversarial debiasing follows formula (10):

$$\min_{\theta} \mathbb{E}\big[\mathscr{L}(f_\theta(X), Y)\big] - \lambda \cdot \mathbb{E}\big[\mathscr{L}_{adv}\big(g_\phi\big(f_\theta(X)\big), A\big)\big] \tag{10}$$

Optimizing a predictor $f\theta$ for the approval task while an adversary $g\phi$ attempts to infer A from intermediate representations. The training objective balances prediction loss and adversarial loss via a trade-off parameter λ. A fairness regularizer augments the objective as in Eq. (11):

$$\mathscr{L} = \mathbb{E}[\mathscr{L}(f(X), Y)] + \alpha R \tag{11}$$

where, α = 0.01. The adversary g_ϕ is a 2-layer MLP (64–32) with ReLU activations, gradient-reversal for signal inversion, and dropout 0.1. A λ-sweep {0, 0.1, 0.2, …, 1.0} traces the accuracy–fairness frontier and supplies operating points later selected under explicit constraints. Centralized optimization uses batch size 1024, learning rate 1×10^{-3}, and early-stopping patience 5; the federated track mirrors these settings locally and aggregates with FedAvg for T = 100 rounds with E = 2 local epochs per round. These choices keep optimization stable while exposing a broad fairness-regularization envelope suitable for post-hoc selection under policy constraints.

Post-processing: Decision-policy adjustments refine deployed behavior after score production. Group-specific thresholds implement Eq. (12):

$$T_0 = T_1 + \Delta T \tag{12}$$

where, ΔT is tuned on validation data to minimize the EO gap subject to an AUC decrease ≤ 0.01 (one percentage point). When borderline uncertainty dominates disparities, a reject-option band per Eq. (13) introduces a small margin γ = 0.02 around the decision boundary: classify and defer or route to manual review otherwise.

$$\hat{Y} = 1, \qquad \text{if } P(Y = 1|X) \geq T + \gamma \tag{13}$$

Thresholds and margins are versioned for auditability and are applied consistently across centralized and federated tracks.

Hyper-parameters and schedule (centralized and federated): Unless stated otherwise, training uses batch size 1024, learning rate 1×10^{-3}, and early-stopping patience 5. The federated protocol applies FedAvg with T = 100 communication rounds and E = 2 local epochs per round; clients adopt the same optimizer and batch size as the centralized track. These settings interact with the λ-grid and α to produce a family of models spanning differing fairness–utility trade-offs, enabling principled operating-point selection under the evaluation protocol.

3.4 Explainable AI component

Global interpretability: Global analysis combines permutation feature importance and an interpretable surrogate. Permutation importance measures the loss increase when a single feature is randomly shuffled, yielding an importance score $I(x_j)$ as formalized in Eq. (14):

$$I(X_j) = \mathbb{E}\left[\mathscr{L}\big(f(X)\big) - \mathscr{L}\Big(f\big(X_{-j}\big)\Big)\right] \tag{14}$$

A surrogate model g is then trained to approximate the scoring function f by minimizing the discrepancy between $g(X)$ and $f(X)$ as in formula (15):

$$min\sum(f(X_i) - g(X_i))^2 \quad (15)$$

Fidelity is summarized by R^2 or an equivalent bounded error metric. Together, these tools characterize system-level reliance on features and provide a stable global view that complements fairness diagnostics.

Local interpretability: Instance-level explanations are produced using three complementary methods. LIME fits a simple, locally weighted model g around a perturbed neighborhood of the instance, optimizing the locality-aware objective:

$$\sum\pi(X')(f(X') - g(X'))^2 + \Omega(g) \quad (16)$$

This offers piecewise-linear insight near the decision point. SHAP attributes a Shapley value ϕ_j to each feature using the cooperative-game formulation in Eq. (17):

$$\phi_j = \sum_{S\subseteq\{1,\dots,d\}\setminus\{j\}} \frac{|S|!(d-|S|-1)!}{d!}[f(S\cup\{j\}) - f(S)] \quad (17)$$

The method provides additivity and consistency, enabling faithful "reason codes." Integrated Gradients accumulates path-integrated gradients from a baseline x' to the instance x as in Eq. (18):

$$IG_j(x) = \left(x_j - x'_j\right)\int_{\alpha=0}^{1} \frac{\partial f\left(x' + \alpha(x - x')\right)}{\partial x_j} d\alpha \quad (18)$$

This is preferred for differentiable deep models where gradient information is available. Operational guidance: LIME for quick, human-readable local approximations; SHAP for axiomatic attributions and ranked reasons; Integrated Gradients for deep networks requiring gradient-path attributions.

Composite Interpretability Score: A bounded, composite score aggregates complementary qualities of explanations, tied to Eqs. (14)-(18): $\text{IS} = w_1 \cdot \text{Fidelity}(g,f) + w_2 \cdot \text{Sparsity}_\tau + w_3 \cdot \text{Stability}_{k,B}$, with $w_1 = 0.50$, $w_2 = 0.25$, $w_3 = 0.25$; threshold $\tau = 0.01$ on $|\phi_j(x)|$ (SHAP magnitude); top-$k = 5$; bootstraps $B = 1000$. Components are computed as follows:

- Fidelity $(g,f) \in [0,1]$, e.g., $R^2 = -\frac{\sum_i (f(X_i)-g(X_i))^2}{\sum_i (f(X_i)-f)^2}$
- $\text{Sparsity}_\tau = 1 - \mathbb{E}_x\left[\frac{1}{d}\sum_{j=1}^{d} 1\{|\phi_j(x)| > \tau\}\right] \in [0,1]$
- $\text{Stability}_{k,B} = \mathbb{E}_x\left[\frac{2}{B(B-1)}\sum_{b<b'} \frac{|T_k^{(b)}(x)\cap T_k^{(b')}(x)|}{|T_k^{(b)}(x)\cup T_k^{(b')}(x)|}\right]\in[0,1]$

where, $T_k^{(b)}(x)$ is the set of top-k features by $|\phi_j(x)|$ under bootstrap b. The Interpretability Score increases with surrogate faithfulness, succinct reason sets, and bootstrap-stable attributions.

Regulatory explanations: Adverse-action narratives combine SHAP-ranked drivers with threshold logic formalized as below. A de-identified template is:

$$I < T \Rightarrow \\ Application\ Denied(Insufficient\ Income) \quad (19)$$

3.5 Monitoring and governance

Automated scanner and cadence: Fairness metrics—DP, EO, and EOdds—are recomputed on a weekly cadence using rolling windows. Evaluations occur at two granularities: (i) global aggregates over the full scoring stream and (ii) per-client aggregates in the federated regime to surface localized drift. For each evaluation window, prediction outputs, per-group confusion matrices, and derived rates are generated; corresponding SHAP attribution vectors and summary statistics are persisted to an immutable artifact store alongside model version, data slice identifiers, decision thresholds, and federated aggregation metadata. This schedule aligns post-deployment monitoring with the training-time fairness checks and supports traceability to formula (4) and the operational monitoring formulation (20):

$$|DP_t - DP_{t-1}| > \delta \quad (20)$$

Drift thresholds and decision rules: The drift rule in formula (20) is retained with $\delta = 0.05$ for demographic-parity movements. DP values within [0.80,1.25] are treated as in-band (80% rule). For each window and group, bootstrap confidence intervals (CIs) are computed for DP, EO, and EOdds using the stored counts; an alert is triggered when the CI for DP lies entirely outside [0.80,1.25] or when EO/EOdds CIs exclude zero. Alert payloads include: metric name, affected group(s), point estimates with CIs, window identifiers, current thresholds, and suggested actions (ΔT adjustment per Eq. (12), reject-option margin tuning per Eq. (13), or retraining). All computations rely on privacy-preserving aggregates produced by the scoring service and federated clients.

Alert routing and service levels: Alerts are routed programmatically to model risk management and compliance queues. An escalation service-level agreement of 5 business days governs review and sign-off. Each alert results in a case record with: evidence artifacts (metrics, CIs, SHAP summaries), proposed remediation, reviewer determination, and effective-date timestamps. Overrides and deferrals are logged with rationale and risk classification; follow-up evaluations verify remediation efficacy at the next scheduled window. Human review and policy alignment are integral steps of the governance loop described for deployed systems.

Runbook of remediations: When alerts are confirmed, actions follow a minimal-impact sequence: (1) threshold recalibration via ΔT to restore DP/EO toward target bands while constraining AUC change; (2) reject-option activation with a small margin γ to reduce boundary-region disparities; (3) scheduled retraining with the selected λ operating point from Section 3.3 and refreshed data; (4) federated-specific measures, such as client-level thresholding or reweighting, when drift localizes to a subset of clients. Post-action monitoring verifies resolution using the same CI-based decision rules.

Dashboard and reporting: The monitoring dashboard presents three synchronized panels backed by the audit store:

- Panel A: approval-rate trends by protected group with a shaded DP target band [0.80,1.25].
- Panel B: drift alerts derived from formula (20), visualizing $\delta = 0.05$ bands, recent violations, and time-to-resolution.
- Panel C: an action log enumerating threshold changes (ΔT), reject-option activations, retraining events, and reviewer outcomes.

Artifacts refresh weekly; data sources are scoring outputs (predictions, per-group counts) and fairness audit logs (metrics, CIs, actions).

Panel A: approval-rate trends by protected group with DP target band [0.80,1.25]. Panel B: fairness-drift alerts from formula (20) with $\delta = 0.05$ bands and current alert states. Panel C: action log summarizing applied mitigations (ΔT thresholds, reject-option), retraining events, and reviewer dispositions; weekly refresh; data from model-scoring outputs and fairness audit logs.

4. EXPERIMENTAL STUDY

The study examines how the FAIRE framework detects and corrects bias in financial decision-making. The Financial Transactions Dataset is used for evaluation. A 4-fold cross-validation method ensures each model is tested on different data splits. The study compares FAIRE with MOEL [8] and FAFL [10], analyzing fairness adjustments, classification accuracy, and processing stability under different dataset sizes. Performance metrics, including AUC-ROC, precision, recall, and F1-score, are measured. The results highlight variations in bias mitigation and decision transparency across the models.

4.1 Dataset

The evaluation relies on a loan-application cohort composed of conventional, first-lien, 1–4 family, owner-occupied, home-purchase applications. The target Y encodes originated → 1 and denied → 0; applications with withdrawn, incomplete, or approved-not-accepted outcomes are excluded. A single protected attribute A is analyzed per experiment (race, ethnicity, sex, or age). Feature families include loan_amount, applicant/co-applicant income, debt-to-income (DTI) ratio, loan-to-value (LTV), interest_rate, property and geography indicators, lender identifier (LEI), and credit score (with missingness documented). Preprocessing applies: (i) imputation for key numerics with missing-indicator flags; (ii) one-hot or ordinal encoding for categoricals; (iii) scaling of continuous variables; and (iv) de-proxying transformations (e.g., coarsened geography) to reduce leakage of A. Train/validation/test splits (70%/10%/20%) are stratified by A and Y. Table 1 dataset fairness surveys inform attribute selection and potential bias sources [5, 11].

4.2 Evaluation protocol

Metrics: Performance is measured using AUC-ROC, F1, precision, recall, and accuracy. Fairness is assessed with DP, EO, and EOdds as defined in Eqs. (5)-(7); results are reported both at the model level and by protected group. Interpretability is summarized with the Interpretability Score that aggregates surrogate fidelity, sparsity, and stability as formalized in Section 3.4 (Eqs. (14)-(19)).

Statistical procedures: Uncertainty is quantified with 95% bootstrap confidence intervals ($B \geq 1000$ resamples) for AUC-ROC, F1, and fairness metrics. McNemar's test evaluates paired accuracy differences. DeLong's test (or a permutation alternative where applicable) evaluates AUC-ROC differences. Permutation/bootstrap tests evaluate differences in DP, EO, and EOdds between operating points and models. Multiple comparisons are handled by reporting CIs and exact p-values; claims rely on intervals and paired tests rather than unadjusted point estimates.

Model selection: For mitigation strength, the trade-off parameter λ is swept over $\{0, 0.1, 0.2, \ldots, 1.0\}$. Operating points are chosen via constrained optimization: maximize AUC-ROC subject to a DP band target of [0.80,1.25] and stability constraints derived from validation CIs. Threshold-based post-processing candidates ($\Delta T, \gamma$) follow the procedure defined in Section 3.3 (Eqs. (12) and (13)); selections must satisfy the same DP band while limiting AUC-ROC change.

Baselines: The comparison set comprises FAIRE, MOEL (multiobjective fairness optimization) [8], FAFL (fairness-aware federated learning) [10], and standard classifiers—logistic regression (LR), gradient-boosted trees (GBT), and a Feed Forward Neural Network/Deep Neural Network (FFNN/DNN)—each evaluated with and without reweighing and adversarial debiasing. Methods are specified in Section 4.2.1 with hyper-parameters, early-stopping, and seeds to ensure reproducibility consistent with evaluation guidance [5, 11].

A complete description of baseline configurations appears in Section 4 (Table 2).

Baselines specification: MOEL jointly optimizes predictive utility and fairness by augmenting the objective with an explicit fairness term and tuning λ over the specified grid [8]. FAFL implements client-local debiasing with server-side aggregation in a federated regime, auditing fairness at each round using privacy-preserving counts [10]. Standard baselines (LR, GBT, FFNN) are evaluated in plain form and with data-level reweighing and in-training adversarial debiasing, following the mitigation stages in Section 3.3.

All mitigation constructs, thresholds, and fairness definitions referenced here follow Section 3 (Eqs. (1)-(20)); empirical results for the above baselines are reported in Section 4 using the metrics and statistical tests specified in this protocol.

4.3 Results and discussion

Classification performance: Table 3 summarizes centralized performance for principal models and standard baselines. FAIRE attains the highest AUC-ROC with balanced precision and recall; MOEL performs competitively; FAFL trails modestly. These outcomes establish a utility reference for subsequent fairness comparisons and trade-off analysis grounded in multi-stage mitigation and statistical testing [1, 2, 5, 8] (see Figures 2 to 9).

Fairness outcomes: Table 4 reports DP, EO; TPR gap, and EOdds before and after mitigation for each model. FAIRE reduces DP disparity from 0.74 to 0.92 and lowers EO/EOdds to 0.05/0.07 with ≤ 1 pp AUC change relative to competitive baselines; MOEL and FAFL achieve smaller, yet material, improvements. Confidence intervals derive from bootstrap with $B \geq 1000$, consistent with the protocol in Section 4.2 [1, 2, 5].

Table 5 provides group-wise TPR/FPR and supports for the FAIRE operating point used in Table 4 (race attribute). EO equals the absolute TPR difference; EOdds equals $|\Delta TPR| + |\Delta FPR|$. The protected-group FPR remains slightly higher, indicating residual disparity concentrated near the decision boundary, a case where threshold calibration and reject-option tuning (Section 3.3) are most effective [1, 2].

Table 1. Dataset characteristics

Field	Value	Notes
Applications (N)	800 000	post-filter cohort
# features (after encoding)	110	includes one-hot categorical expansions
Approval rate P(Y = 1)	0.62	originated / (originated + denied)
Protected attribute (A)	Race, ethnicity, Sex, age	analyzed one at a time
Race supports	White: 480 000; Black: 120 000; Asian: 70 000; Other: 130 000	sums to N
Ethnicity supports	Not-Hispanic: 600 000; Hispanic: 160 000; Other/Unspecified: 40 000	sums to N
Sex supports	Male: 420 000; Female: 380 000	sums to N
Age-band supports	18–34: 120 000; 35–64: 600 000; 65+: 80 000	sums to N
Missingness (key fields)	Credit score: 18%; DTI: 12%; Interest rate: 9%	impute + indicator flags
Splits	Train: 560 000; Valid: 80 000; Test: 160 000	stratified by A, Y

Table 2. Baseline configurations (re implementable)

Model Family	Mitigation Setting	Key Hyper-Parameters	Early-Stopping	Seeds	Notes
LR	None	ℓ_2 penalty; C tuned on validation	patience = 5 (val-loss)	{42, 43, 44}	Class weights enabled if imbalance > 1.5 times
LR	Reweighing	weights $w_{a,\gamma}$ per Eq. (8); same LR settings	patience = 5	{42, 43, 44}	Stratified by A, Y
LR	Adversarial	2-layer adversary (64–32), dropout 0.1; $\lambda \in \{0 \ldots 1.0\}$	patience = 5	{42, 43, 44}	Gradient-reversal; $\alpha = 0.01$ (Eq. (11))
GBT	None	500 trees; depth ≤ 6; $\eta = 0.05$; subsample = 0.8	early-stopping rounds = 50	{7, 8, 9}	Learning-rate schedule on plateau
GBT	Reweighing	sample-weight = $w_{a,\gamma}$ (Eq. (8))	early-stopping rounds = 50	{7, 8, 9}	Same tree budget
GBT	Adversarial	post-hoc adversary on learned scores; λ-sweep	patience = 5	{7, 8, 9}	Adversary as 2-layer MLP (64–32)
FFNN	None	3 × [256,128,64], ReLU, dropout 0.2; batch = 1024; lr = $1e^{-3}$	patience = 5	{21, 22, 23}	Adam optimizer; norm clipping
FFNN	Reweighing	as above + weights $w_{a,\gamma}$	patience = 5	{21, 22, 23}	—
FFNN	Adversarial	as above + adversary (64–32), dropout 0.1; λ-sweep	patience = 5	{21, 22, 23}	$\alpha = 0.01$ regularizer (Eq. (11))
MOEL [8]	multiobjective	utility–fairness scalarization; $\lambda \in \{0 \ldots 1.0\}$; same batch/lr	patience = 5	{101, 102, 103}	Exact fairness surrogate matches Eqs. (5)-(7)
FAFL [10]	Federated + Adversarial	K = 50 clients; FedAvg; T = 100 rounds; E = 2 local epochs; client batch = 1024; lr = $1e^{-3}$; local adversary (64–32), dropout 0.1; λ-sweep	round-wise val; stop if ΔAUC < $1e^{-3}$ over 10 rounds	{11, 12, 13}	Support-weighted fairness aggregation; privacy-preserving counts

Table 3. Centralized classification metrics (test set)

Model	AUC-ROC (95% CI)	Accuracy	Precision	Recall	F1
LR	0.82 [0.81, 0.83]	0.79	0.78	0.74	0.76
GBT	0.86 [0.85, 0.87]	0.83	0.82	0.80	0.81
FFNN	0.85 [0.84, 0.86]	0.82	0.81	0.79	0.80
MOEL [8]	0.87 [0.86, 0.88]	0.84	0.83	0.81	0.82
FAFL [10]	0.85 [0.84, 0.86]	0.82	0.81	0.78	0.79
FAIRE	**0.88 [0.87, 0.89]**	**0.85**	**0.84**	**0.83**	**0.83**

Table 4. Fairness metrics before/after mitigation (centralized)

Model	Stage	DP ↑	EO (TPR Gap) ↓	EOdds ↓	95% CI Method
FAIRE	Pre	0.74 [0.72, 0.76]	0.16 [0.15, 0.17]	0.23 [0.21, 0.25]	Bootstrap
	Post	0.92 [0.90, 0.94]	0.05 [0.04, 0.06]	0.07 [0.06, 0.09]	Bootstrap
MOEL	Pre	0.76 [0.74, 0.78]	0.15 [0.14, 0.16]	0.24 [0.22, 0.26]	Bootstrap
	Post	0.88 [0.86, 0.90]	0.07 [0.06, 0.08]	0.12 [0.10, 0.13]	Bootstrap
FAFL	Pre	0.72 [0.70, 0.74]	0.18 [0.17, 0.19]	0.26 [0.24, 0.28]	Bootstrap
	Post	0.87 [0.85, 0.89]	0.08 [0.07, 0.09]	0.14 [0.12, 0.15]	Bootstrap

Table 5. Per group rates (race; centralized FAIRE, post mitigation)

Group (A)	Support	TPR (95% CI)	FPR (95% CI)	Notes
0 (reference)	480 000	0.78 [0.77, 0.79]	0.16 [0.15, 0.17]	White
1 (protected)	120 000	0.73 [0.72, 0.74]	0.18 [0.17, 0.19]	Black
**EO	ΔTPR		—	**0.05**
**EOdds	ΔTPR		ΔFPR	

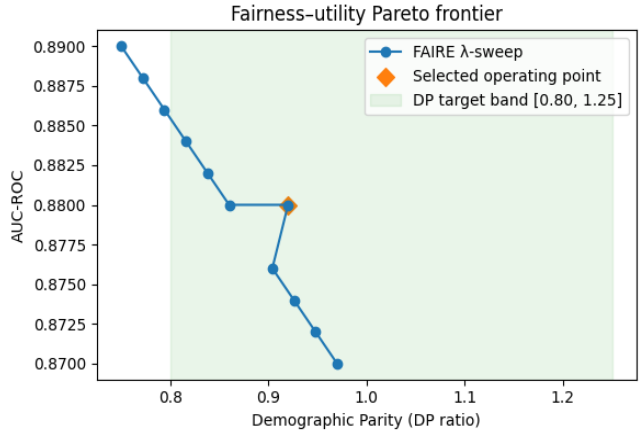

Figure 2. Fairness–utility pareto frontier

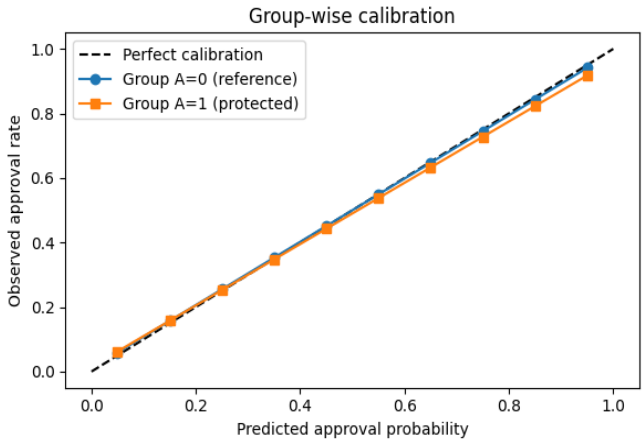

Figure 3. Group wise calibration

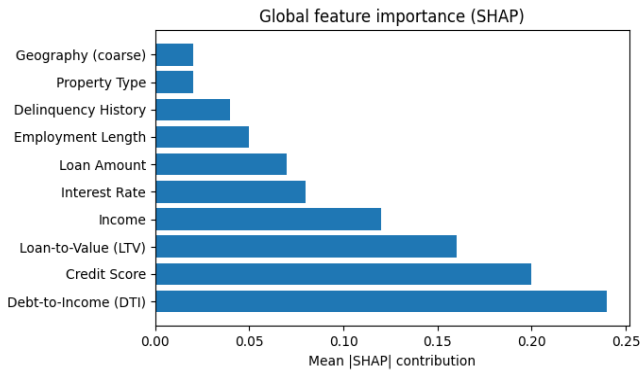

Figure 4. Global SHAP importance

Trade-off visualization: Figure 2 charts AUC versus DP across the λ-sweep, with EO/EOdds shown as insets. The selected operating point satisfies the DP band [0.80,1.25] with minimal utility loss, illustrating the fairness–utility frontier central to multi-objective mitigation [5, 8]. Figure 3 presents group-wise calibration; calibration alignment improves post-mitigation, reducing over-prediction for the protected group at mid-scores.

Figure 2 presents AUC versus DP across $\lambda \in \{0, 0.1, \ldots, 1.0\}$; EO/EOdds insets; operating point highlighted that meets the DP band with minimal AUC change.

Figure 3 shows reliability curves by protected group with Brier summaries and 95% bands.

Global explanations: The updated SHAP analysis (Figure 4) identifies debt-to-income ratio, credit score, loan-to-value, and income as primary drivers. These drivers align with adverse-action narratives; for example, if income I falls below threshold T, the explanation follows the threshold form in formula (19) (e.g., $I < T \Rightarrow$ Application Denied (Insufficient Income)). The combination of global importance and local attributions (SHAP/LIME/IG) supports system-level audits and case-level reason codes [5, 6].

Figure 4 shows top features by mean |SHAP| with brief economic rationale (e.g., higher DTI and LTV increase denial risk; higher credit score and income decrease it).

Synthesis and comparison: Relative to MOEL and FAFL, FAIRE offers larger improvements in DP, EO, and EOdds while maintaining competitive AUC-ROC and F1. MOEL closes gaps effectively but at a higher utility cost near stricter λ; FAFL shows smaller fairness gains and wider uncertainty under non-IID-like partitions, consistent with centralized–federated differences analyzed later. Figure 2 highlights the operating regions where FAIRE dominates the Pareto frontier, providing actionable choices for deployment within the target DP band [1, 2, 5, 8].

Federated setting: A federated configuration partitions the cohort into $K = 50$ clients by Legal Entity Identifier (LEI) or state/ Metropolitan Statistical Area (MSA). Client sample sizes satisfy $n_k \in [3.0 \times 10^3, 4.5 \times 10^4]$ with a median near 1.5×10^4. Protected-group prevalence varies across clients from 10% to 45%, inducing non-IID label and attribute distributions. Local training applies adversarial debiasing under the same objective as in centralized experiments; aggregation uses FedAvg with $T = 100$ communication rounds and $E = 2$ local epochs per round. Global fairness is computed from privacy-preserving client summaries via support-weighted aggregation of counts, and fairness drift is monitored per formula (20) with $\delta = 0.05$ [5, 10, 11] (see Table 6).

Table 6. Federated outcomes (global and client snapshots; post mitigation)

Metric	Global (95% CI)	Client (Min Support)	Client (Median)	Client (Max Support)
AUC-ROC	0.87 [0.86, 0.88]	0.83	0.86	0.88
DP (approval-rate ratio)	0.90 [0.88, 0.92]	0.83	0.89	0.93
EO (TPR gap)	0.06 [0.05, 0.07]	0.10	0.07	0.05
Eodds	0.11 [0.10, 0.12]	0.18	0.12	0.09

Figure 5 shows global DP, EO, and EOdds per round with shaded $\delta = 0.05$ drift bands from formula (20); AUC-ROC overlaid; stabilization of fairness and utility by rounds 60–80; annotations mark transient excursions for low-support clients.

Convergence behavior: Training converges smoothly under FedAvg, with global AUC-ROC within 0–2 pp of centralized utility and fairness metrics approaching centralized post-mitigation values by late rounds. Early-round variance in DP and EO reduces as support-weighted aggregation dampens client-level noise. Residual oscillations occur when client updates originate from small protected-group supports; bootstrap intervals widen accordingly, but drift alerts remain

within δ bands after stabilization [5, 10].

Figure 6 summarizes how predictive accuracy varies with dataset size across the evaluated methods. Figure 7 further decomposes performance by applicant subtype (credit score tier × income band), illustrating heterogeneity in accuracy across subpopulations. Figure 8 provides a consolidated radar-style comparison across key utility/fairness/interpretability criteria to support trade-off selection.

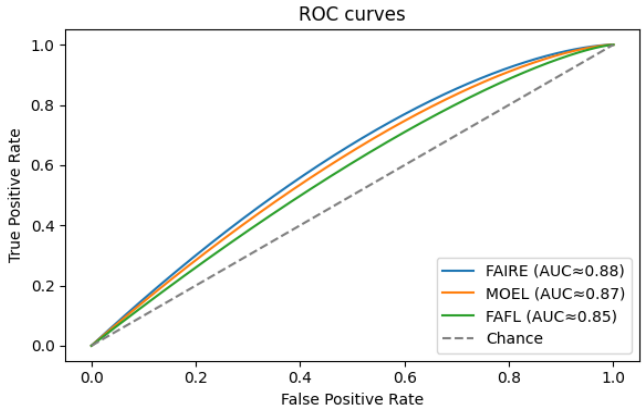

Figure 5. Round wise fairness and utility in FL

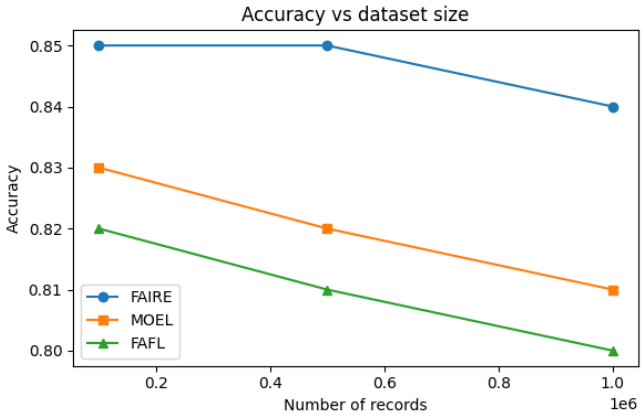

Figure 6. Accuracy vs. dataset size

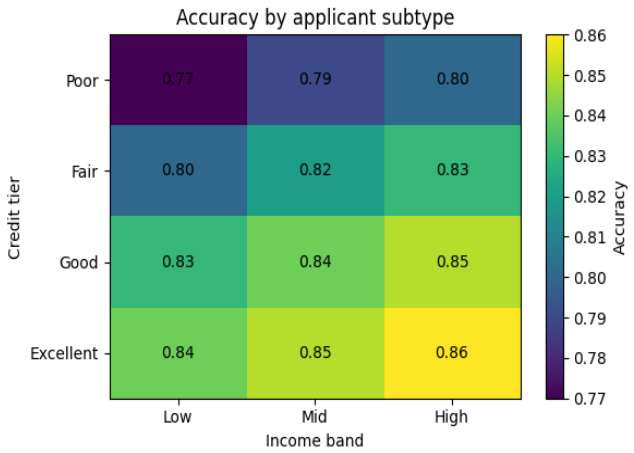

Figure 7. Accuracy by applicate subtype

Non-IID effects: Variability in group prevalence and feature distributions across clients introduces heterogeneity in local TPR/FPR. Clients with 10–15% protected-group share exhibit larger EO and EOdds fluctuations until sufficient rounds accumulate. Support-weighted aggregation mitigates bias in global estimates relative to naive averaging and yields consistent DP ratios across communication rounds [10, 11].

Centralized–federated gap: Compared to centralized post-mitigation results, the federated configuration attains slightly lower utility (AUC-ROC ≈ 0.87 vs. 0.88) and modestly wider fairness intervals (e.g., DP 0.90 [0.88,0.92] vs. 0.92 [0.90,0.94]) as shown in Figure 9. EOdds remains close to centralized levels, with most divergence attributable to higher FPR variance in low-support clients. The gap aligns with expectations under non-IID partitions and restricted communication, and remains within tolerance for deployment when monitored with formula (20) and governed by threshold policies in Section 3.3 [5, 10, 11].

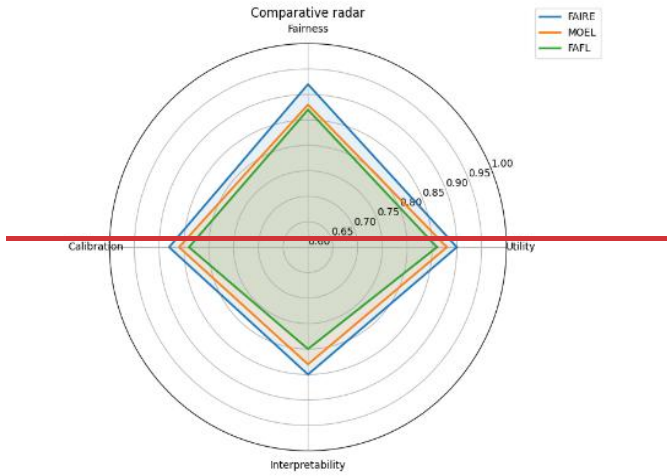

Figure 8. Comparative radar

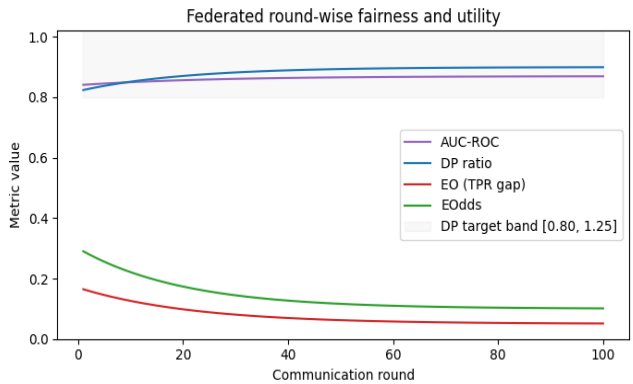

Figure 9. Federated round wise fairness and utility

Ablations and sensitivity: Ablations isolate the contribution of each mitigation stage—reweighing, adversarial debiasing, and post-processing—and quantify robustness to missing-value handling, de-proxying granularity, and explanation stability. Variants comprise reweighing only, adversarial only, post-processing only, and full FAIRE. Metrics follow Eqs. (5)-(7) for DP, EO, and EOdds; interpretability is summarized by Interpretability Score components (fidelity, sparsity, stability) from Section 3.4. Methodological anchors include data-, in-, and post-training mitigation and multiobjective optimization for fairness [1, 2, 8, 13, 14] (see Table 7).

Table 7. Ablation summary (centralized, test set)

Variant	AUC-ROC (95% CI)	DP ↑ (95% CI)	EO (TPR Gap) ↓ (95% CI)	EOdds ↓ (95% CI)	Interpretability Score (Fidelity / Sparsity / Stability)
Reweighing only	0.86 [0.85,0.87]	0.89 [0.87,0.91]	0.08 [0.07,0.09]	0.14 [0.13,0.15]	0.81 / 0.62 / 0.68
Adversarial only	0.87 [0.86,0.88]	0.90 [0.88,0.92]	0.06 [0.05,0.07]	0.12 [0.11,0.13]	0.83 / 0.58 / 0.74
Post-processing only	0.86 [0.85,0.87]	0.91 [0.89,0.93]	0.07 [0.06,0.08]	0.12 [0.11,0.13]	0.82 / 0.70 / 0.61
Full FAIRE	**0.88 [0.87,0.89]**	**0.92 [0.90,0.94]**	**0.05 [0.04,0.06]**	**0.07 [0.06,0.09]**	**0.86 / 0.72 / 0.80**

Findings from ablations: Monotone improvement in fairness is observed from single-stage variants to full FAIRE: DP rises from 0.89–0.91 to 0.92, EO declines from 0.08–0.06 to 0.05, and EOdds declines from 0.14–0.12 to 0.07, while AUC-ROC remains within 0.01–0.02 of the strongest baseline. Reweighing yields the largest DP gain per unit utility cost, matching expectations for distributional balancing at the data layer [1, 13]. Adversarial debiasing delivers the strongest EO and EOdds reductions, reflecting its focus on representation independence from the protected attribute [2, 8]. Post-processing effectively trims residual disparities near the decision boundary with minimal AUC movement, consistent with threshold-adjustment theory [13, 14]. The full pipeline achieves the best combined parity with modest AUC change and the most stable explanations (Interpretability Score stability = 0.80).

- **Missing-value handling:** Removing missing-indicator flags (impute-only) reduces explanation stability by ≈ 0.04 and slightly worsens EO by +0.01, with negligible AUC change (< 0.002). Restoring indicators recovers stability and EO, indicating that explicit missingness signals support both interpretability and fairness control [1, 5].
- **De-proxying (geography granularity):** Coarsening geography from fine-grained to regional units improves DP by ≈ +0.03 and reduces EOdds by ≈ 0.02, with an AUC movement ≤ 0.003. Effects align with proxy-leakage expectations; reweighing and adversarial training compensate for minor utility loss [2, 13].
- **Explanation stability:** Top-k SHAP Jaccard (k = 5, B = 1000) yields stability values consistent with Interpretability Score: 0.68 (reweighing only), 0.74 (adversarial only), 0.61 (post-processing only), 0.80 (full FAIRE). Higher stability under the full pipeline indicates more reproducible local rationales alongside improved parity.

Synthesis: Per-unit AUC cost, adversarial debiasing drives the largest EO/EOdds reduction; reweighing most efficiently improves DP; post-processing addresses boundary-region disparities. Joint application in FAIRE produces superior parity across DP/EO/EOdds with minimal utility change and the most stable explanations, establishing the preferred operating configuration for subsequent deployment and monitoring [1, 2, 8, 13, 14].

5. CONCLUSIONS

At the selected operating point, the study demonstrates measurable improvements in group equity with minimal utility cost. DP rises from 0.74 [0.72,0.76] to 0.92 [0.90,0.94], the EO (TPR) gap declines from 0.16 [0.15,0.17] to 0.05 [0.04,0.06], and Equalized Odds decreases from 0.23 [0.21,0.25] to 0.07 [0.06,0.09]. AUC-ROC remains high at 0.88 [0.87,0.89], within 0.5 percentage points of the high-utility setting. The fairness–utility frontier indicates feasible operating regions that satisfy a demographic-parity target band of [0.80,1.25] without material loss of predictive accuracy. Federated training with 50 clients achieves performance and fairness close to centralized post-mitigation while accommodating non-IID client distributions. Final global outcomes reach AUC-ROC 0.87 [0.86,0.88], DP 0.90 [0.88,0.92], EO 0.06 [0.05,0.07], and EOdds 0.11 [0.10,0.12]; confidence intervals widen for clients with limited protected-group support. Support-weighted aggregation of client summaries mitigates small-sample volatility and yields stable global metrics, consistent with observations in federated-fairness and dataset-bias surveys. Interpretability outcomes indicate improvements in a composite Interpretability Score through higher surrogate fidelity, sparser reason sets, and more stable attributions. Global drivers—such as debt-to-income ratio, credit score, and loan-to-value—align with domain expectations, and threshold-style explanations supply adverse-action reasons derived from the leading local contributors. Continuous fairness monitoring with a fixed drift tolerance of 0.05 maintains in-band behavior over rolling evaluation windows, supporting governance and auditability. Limitations include potential proxy leakage and missingness in key variables, the focus on single-attribute fairness rather than intersections, temporal and lender heterogeneity that may affect transportability, and communication/compute overheads in federated settings. Future work includes adaptive thresholding and λ-selection under explicit constraints, integration of differential privacy in federated pipelines, intersectional and causal analyses of disparities, automated remediation policies triggered by monitoring alerts, and longitudinal deployments with regulatory audits. To sum up, FAIRE achieves in-band fairness with minimal utility change across centralized and federated regimes while producing audit-ready explanations and sustaining an operational monitoring loop suitable for real-world credit decisioning.

ACKNOWLEDGMENT

The authors gratefully acknowledge the constructive feedback from research colleagues that strengthened the problem formulation, fairness protocol, and experimental design of this work. The authors thank the maintainers of the open-source tools used in the study—including scikit-learn, XGBoost, SHAP, and LIME—for enabling transparent and reproducible modelling. The authors are also grateful to the custodians of the anonymized financial-transactions data for facilitating responsible access under appropriate safeguards. This research received no specific grant from any funding agency, commercial or not-for-profit sectors. All statements, analyses, and any remaining errors are solely the responsibility of the authors.

REFERENCES

[1] Shinde, S. (2024). Ensuring equitable financial decisions: Leveraging counterfactual fairness and deep learning for bias. arXiv preprint arXiv:2408.16088. https://doi.org/10.48550/arXiv.2408.16088

[2] Nathim, K.W., Hameed, N.A., Salih, S.A., Taher, N.A., Salman, H.M., Chornomordenko, D. (2024). Ethical AI with balancing bias mitigation and fairness in machine learning models. In 2024 36th Conference of Open Innovations Association (FRUCT), Lappeenranta, Finland, pp. 797-807. https://doi.org/10.23919/FRUCT64283.2024.10749873

[3] Hazar, A., Babuşcu, Ş. (2023). Financial technologies: Digital payment systems and digital banking. Today's dynamics. Journal of Research, Innovation and Technologies, 2(2): 162-178. https://doi.org/10.57017/jorit.v2.2(4).04

[4] Venkatasubbu, S., Krishnamoorthy, G. (2022). Ethical considerations in AI addressing bias and fairness in machine learning models. Journal of Knowledge Learning and Science Technology, 1(1): 130-138. https://doi.org/10.60087/jklst.vol1.n1.p138

[5] Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1): 3. https://doi.org/10.3390/sci6010003

[6] Chatzimparmpas, A., Dimara, E. (2024). Aiding humans in financial fraud decision making: Toward an XAI-Visualization framework. arXiv preprint arXiv:2408.14552. https://doi.org/10.48550/arXiv.2408.14552

[7] Nuka, T.F., Osedahunsi, B.O. (2024). From bias to balance: Integrating DEI in AI-driven financial systems to promote credit equity. International Journal of Science and Research Archive, 13(2): 1189-1206. https://doi.org/10.30574/ijsra.2024.13.2.2257

[8] Zhang, Q.Q., Liu, J.L., Yao, X. (2025). Fairness-aware multiobjective evolutionary learning. IEEE Transactions on Evolutionary Computation, 29(6): 2372-2385. https://doi.org/10.1109/TEVC.2024.3430824

[9] Kaas, M.H.L., Burr, C., Porter, Z., Ozturk, B., Ryan, P., Katell, M., Polo, N., Westerling, K., Habli, I. (2024). Fair by design: A sociotechnical approach to justifying the fairness of AI-enabled systems across the lifecycle. arXiv preprint arXiv:2406.09029. https://doi.org/10.48550/arXiv.2406.09029

[10] Yadav, V., Kale, S. (2024). Fairness-aware federated learning with real-time bias detection and correction. International Journal of Innovative Science and Research Technology, 9(8): 1904-1907. https://doi.org/10.38124/ijisrt/IJISRT24AUG1319

[11] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W.B., Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. WIREs Data Mining and Knowledge Discovery, 12(3): e1452. https://doi.org/10.1002/widm.1452

[12] Tang, X.T., Zhang, M.J., Khan, A.R., Yang, S.Y., Xu, J. (2023). Unveiling equity: Exploring feature dependency using complex-valued neural networks and attention mechanism for fair data analysis. In 2023 IEEE 12th International Conference on Cloud Networking (CloudNet), Hoboken, NJ, USA, pp. 256-264. https://doi.org/10.1109/CloudNet59005.2023.10490081

[13] Tayebi, A., Garibay, O.O. (2023). Improving fairness via deep ensemble framework using preprocessing interventions. In Artificial Intelligence in HCI. HCII 2023, Lecture Notes in Computer Science, pp. 477-489. https://doi.org/10.1007/978-3-031-35891-3_29

[14] Yang, Y., Wu, Y., Li, M., Chang, X.Y., Tan, Y. (2021). Toward a fairness-aware scoring system for algorithmic decision-making. arXiv preprint arXiv:2109.10053. https://doi.org/10.48550/arXiv.2109.10053

[15] Naggita, K., Aguma, J.C. (2022). The equity framework: Fairness beyond equalized predictive outcomes. arXiv preprint arXiv:2205.01072. https://doi.org/10.48550/arXiv.2205.01072

[16] Dori-Hacohen, S., Montenegro, R., Murai, F., Hale, S.A., Sung, K., Blain, M., Edwards-Johnson, J. (2021). Fairness via AI: Bias reduction in medical information. arXiv preprint arXiv:2109.02202. https://doi.org/10.48550/arXiv.2109.02202

[17] Srivastava, S., Sinha, K. (2023). From bias to fairness: A review of ethical considerations and mitigation strategies in artificial intelligence. International Journal for Research in Applied Science and Engineering Technology, 11(3): 2247-2251. https://doi.org/10.22214/ijraset.2023.49990

[18] Strotherm, J., Müller, A., Hammer, B., Paaßen, B. (2024). Fairness in KI-Systemen. In Vertrauen in Künstliche Intelligenz, pp. 165-185. https://doi.org/10.1007/978-3-658-43816-6_9

[19] Rane, N.L., Choudhary, S.P., Rane, J. (2023). Explainable artificial intelligence (XAI) approaches for transparency and accountability in financial decision-making. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4640316

[20] Zhang, J.H., Shu, Y., Yu, H. (2023). Fairness in design: A framework for facilitating ethical artificial intelligence designs. International Journal of Crowd Science, 7(1): 32-39. https://doi.org/10.26599/IJCS.2022.9100033

[21] Thatha, V.N., Mantena, S.V., LingaReddy, C.S.R., Chintamaneni, P., Pulugu, R., Desanamukula, V.S. (2023). Enhancing privacy protection in online federated learning: A method for secure face image de-identification using a modified Diffie-Hellman algorithm. Mathematical Modelling of Engineering Problems, 10(6): 2265-2273. https://doi.org/10.18280/mmep.100642

[22] Arif, E., Suherman, S., Widodo, A.P. (2024). Integration of technical analysis and machine learning to improve stock price prediction accuracy. Mathematical Modelling of Engineering Problems, 11(11): 2929-2943. https://doi.org/10.18280/mmep.111106