4 Trustworthy Agentic AI for BNPL Credit Risk in India: A Comparative Evaluation of Explainability Frameworks in Embedded Finance

Nidhi Verma[1*], Tripti[2], Dr. Komal Rani[3] & Prateek Sood[4]
[1]Assistant Professor, School of Business, The NorthCap University, Gurugram, Haryana.
[2]Assistant Professor, Chandigarh School of Business, CGC University, Mohali, Punjab.
[3]Associate Professor, Department of Commerce & Management, Baba Mastnath University (BMU), Rohtak, Haryana.
[4]General Manager-Human Resource, CGC University, Mohali, Punjab, India.

*Corresponding Author: nidhi.ggn1911@gmail.com

Abstract

India's digital payments revolution - anchored by the Unified Payments Interface (UPI), the Account Aggregator (AA) consent framework, and the Jan Dhan Aadhaar Mobile (JAM) trinity - has created the world's most dynamic real-time credit ecosystem. Buy Now Pay Later (BNPL) transaction volumes surpassed USD 45 billion in FY2023–24, growing at over 30% CAGR since 2021, with more than 180 million active users. Roughly 60% of new digital borrowers in urban Tier-2 and Tier-3 markets are either absent from credit bureaus or carry scores below 650. The machine learning pipelines that underwrite these borrowers operate as autonomous, multi-step decision agents - querying AA APIs, processing UPI histories, running ensemble classifiers, routing decisions, and logging compliance records without human intervention at transaction scale. This is not AI in lending; it is agentic AI in lending, and its governance demands more than post-hoc explanation. This paper presents the first empirical evaluation of explainability frameworks applied to agentic AI credit scoring systems in Indian embedded finance. Three XAI methods - SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and a tabular adaptation of Gradient-weighted Class Activation Mapping (Grad-CAM) - are assessed across XGBoost and Random Forest classifiers on three contextualised benchmark datasets: the German Credit dataset (proxy for bureau-light consumer credit), the HMEQ dataset (representing high-missingness alternative data environments), and a LendingClub sample (capturing India's gig economy income-volatility profiles). Assessment follows a three-dimensional framework of fidelity, stability, and comprehensibility, complemented by structured fairness auditing and regulatory alignment mapping across RBI Digital Lending Guidelines (2022), the Digital Personal Data Protection Act 2023, and EU AI Act Article 13. SHAP achieves fidelity of 0.91 and stability rank-correlation of 0.97, outperforming LIME (0.76; 0.63) and Grad-CAM (0.68; 0.71) across all 18 experimental conditions. XGBoost outperforms Random Forest on AUC-ROC in every dataset, with the widest margin

on the 43-feature LendingClub sample (0.821 vs. 0.798). Statistically significant demographic disparate impact is detected in two of three datasets; SHAP uniquely enables the decomposition of proxy discrimination pathways into regulatory-grade attribution evidence. A novel positive-gradient asymmetry is identified in the tabular Grad-CAM adaptation, with direct implications for protected attribute surfacing. The paper proposes the Autonomy–Accountability–Auditability (AAA) governance framework as an organising architecture for trustworthy agentic AI in BNPL credit decisioning, demonstrating alignment with India's evolving regulatory stack and EU AI Act transparency requirements.

Keywords: Trustworthy AI, Agentic AI, BNPL, Embedded Finance, India, Credit Scoring, SHAP, LIME, Grad-CAM, Algorithmic Fairness, RBI Digital Lending Guidelines, DPDP Act, Thin-File Borrowers, FinTech Governance, XAI.

Introduction

India's credit landscape has undergone structural transformation within a single decade. UPI processed over 117 billion transactions in FY2023–24 while the AA framework, operationalised at national scale from 2022, enabled consent-based access to multi-institutional financial data for over 500 million adults (NPCI, 2024; RBI, 2023). BNPL products, riding atop this infrastructure, have evolved from a checkout convenience into a primary credit channel for over 180 million borrowers, with projections estimating 300 million users by 2028 (KPMG India, 2024; Redseer Consulting, 2024).

Approximately 60% of digital credit applicants in Tier-2 and Tier-3 urban markets lack bureau records or carry scores below 650. Lenders have responded with ensemble ML models ingesting UPI transaction patterns, merchant category distributions, gig income histories, and GST filing behaviour to produce decisions in milliseconds (RBI, 2023). Critically, these models do not merely score applicants - they autonomously orchestrate data retrieval, feature engineering, decision routing, adverse action generation, and compliance logging without human intervention between transactions. This is agentic AI: goal-directed, multi-step, environment-responsive, and operating at speeds that make transactional human oversight structurally infeasible (Russell & Norvig, 2021; Wooldridge, 2020).

Regulatory expectations are converging on this reality. The RBI Digital Lending Guidelines (2022) require credit algorithms to be transparent, auditable, and non-discriminatory. The forthcoming DPDP Act 2023 implementation rules are expected to confer explanation rights functionally similar to GDPR Article 22. The EU AI Act (Regulation 2024/1689) mandates interpretability and human oversight for

ISBN: 978-93-49468-49-8

high-risk AI systems including credit scoring, with extraterritorial reach to Indian operators processing European data. Three gaps in existing literature motivate this study: no prior work has evaluated XAI methods in the Indian BNPL context; simultaneous cross-dataset, cross-model comparisons of SHAP, LIME, and Grad-CAM are absent; and no study has coupled agentic AI governance with fairness auditing and multi-jurisdictional regulatory mapping in credit decisioning. This paper addresses all three through an 18-condition empirical evaluation and the AAA governance framework.

Theoretical and Regulatory Foundations

- **Agentic AI and India's BNPL Credit Architecture**

Contemporary BNPL scoring pipelines in India are agentic systems in the technical sense: they execute sequential actions, adapt intermediate outputs, and interact with external APIs and databases without requiring human decision-making at each step. The governance implications of this architecture differ fundamentally from those of static ML models. For a human analyst using a model to assist a decision, XAI serves as a communication aid. For an agentic system making ten thousand credit decisions per minute, XAI becomes a governance infrastructure - the mechanism through which human principals (risk officers, regulators, borrowers) exercise control over an agent whose operational pace makes direct supervision impossible.

The principal-agent framework of Jensen and Meckling (1976) applies here in an extended form. The standard information asymmetry between lender and borrower - the problem ML scoring addresses by mining alternative data (Berg et al., 2020) - is accompanied by a second-order asymmetry: the lender-as-principal cannot directly observe the reasoning of its AI agent without XAI. Floridi et al. (2019) identify explicability as central to trustworthy AI design. The European Commission's Ethics Guidelines for Trustworthy AI (2019) operationalise this through principles of human oversight, accountability, and transparency. Our AAA framework (Section 2.4) adapts these principles to the specific characteristics of agentic credit pipelines.

India's digital public infrastructure creates a data environment with no global precedent. The AA framework enables a BNPL platform to retrieve twelve months of cross-institutional transaction data, insurance premium histories, and gig platform earnings through a single API call at checkout (RBI, 2023). Production scoring models in this environment commonly engineer 50–200 features, structurally distinct from the 12–43-feature benchmark datasets used in this study. The fairness exposure profile is also India-specific: geographic tier, UPI transaction frequency, and gig employment status correlate with constitutionally protected characteristics including caste, religion, and gender in ways that differ from the proxy discrimination pathways documented in US mortgage markets (Fuster et al., 2022; Bartlett et al., 2022).

ISBN: 978-93-49468-49-8

- **The AAA Governance Framework**

The Autonomy–Accountability–Auditability (AAA) framework organises trustworthy governance of agentic AI credit scoring along three structural dimensions.

Autonomy governance does not ask whether human oversight exists but how it is calibrated. At ten thousand decisions per minute, transactional oversight is structurally impossible; governance requires the design of escalation threshold rules (which conditions trigger mandatory human review?), override authority chains (who can reverse agentic decisions and through what process?), and sampling-based review frequency. XAI quality directly determines the functional value of any oversight mechanism: an explanation with low fidelity renders human review of flagged cases uninformative.

Accountability under the AAA framework requires that every credit decision be traceable to a specific combination of model version, training data vintage, feature set, and responsible risk officer. SHAP's additive decomposition - in which the sum of all feature attribution values equals the model output - provides the only tested mechanism through which an adverse decision can be fully disaggregated into verifiable, per-feature contributions. This additive accountability chain is required for RBI Digital Lending Guidelines compliance and anticipated DPDP Act explanation rights.

Auditability refers to the capacity of internal governance teams, external auditors, and regulatory examiners to independently verify system performance across accuracy, fairness, and interpretability dimensions. The tri-dimensional XAI quality framework used in this study - fidelity, stability, and comprehensibility measured against supervisory benchmarks - constitutes the operational specification of auditability under the AAA architecture.

- **XAI Methods: Mechanisms and Governance Fit**

SHAP (Lundberg & Lee, 2017) grounds attribution in cooperative game theory: each feature receives its weighted average marginal contribution across all possible feature orderings. TreeSHAP (Lundberg et al., 2020) computes exact Shapley values in polynomial time for tree-based models, enabling transaction-scale deployment. Four axiomatic properties - efficiency, symmetry, dummy, and additivity - guarantee a unique, mathematically fair attribution decomposition. LIME (Ribeiro et al., 2016) constructs local linear surrogates by perturbing instances and fitting Ridge regression on the resulting neighbourhood; coefficient outputs translate naturally into adverse action language but the stochastic sampling mechanism introduces run-to-run instability that undermines reproducibility requirements. Grad-CAM (Selvaraju et al., 2017), originally designed for CNN image classifiers, is adapted for tabular tree models (Agarwal et al., 2021) by wrapping the model in a differentiable PyTorch graph and applying ReLU-activated gradient-feature interactions. The ReLU step

ISBN: 978-93-49468-49-8

suppresses negative-gradient contributions, producing a systematic positive-gradient asymmetry that inflates attribution scores for certain feature types - a structural bias with direct regulatory implications explored in Section 4.4.

- **Regulatory Landscape: India and Global Alignment**

The RBI Digital Lending Guidelines (2022) require that credit assessment algorithms be transparent, unbiased, and auditable, with all decisions formally made by the regulated entity rather than delegated to the algorithm. The RBI's 2024 discussion paper on responsible AI strengthens this to demand that explanations be 'consistent, robust, and replicable.' The DPDP Act 2023, once implementation rules are issued, is expected to grant data principals explanation rights for automated significant decisions, functionally equivalent to GDPR Article 22. EU AI Act Article 13 designates credit scoring as high-risk AI under Annex III, requiring outputs to be interpretable to both operators and affected individuals, while Article 14 mandates human oversight mechanisms capable of detecting and correcting errors. The CFPB's 2022 adverse action circular extends feature-level interpretability requirements to ML-based credit models. A common thread across all four jurisdictions is the requirement for instance-level, consistent, and human-comprehensible explanation - a standard that, as the results demonstrate, only SHAP reliably meets.

Methodology

- **Research Design and Datasets**

The study employs a positivist, quantitative comparative design with a within-subject structure: identical preprocessing, hyperparameter tuning, and XAI evaluation are applied across all combinations of three datasets, two models, and three XAI methods, yielding 18 experimental conditions. This design enables both within-model and cross-model comparisons across every XAI method.

Three publicly available benchmark datasets are used as structural surrogates for the Indian BNPL data environment. The German Credit dataset (Hofmann, 1994; 1,000 records, 20 features, 30% default rate) proxies India's urban Tier-2/3 thin-file consumer credit portfolio, where behavioural signals carry more weight than tradeline information and age and personal status serve as protected-class proxies. The HMEQ dataset (5,960 observations, 12 features, 19.9% default rate) is selected for its data quality characteristics - 27.8% missingness on MORTDUE and 32.3% on VALUE - which structurally mirror the gapped transaction histories and irregular reporting intervals encountered in Indian AA-sourced alternative data. The LendingClub sample (100,000 observations after stratified sampling, 43 engineered features, 14.6% default rate) most closely reflects the income volatility and employment instability of India's gig economy borrower segment. Numeric features are log-transformed, employment tenure tokenised,

ISBN: 978-93-49468-49-8

categorical features one-hot-encoded, and highly missing features excluded. SMOTE oversampling is applied to HMEQ training folds only; class weighting handles imbalance for German Credit and LendingClub.

Table 1: Benchmark Datasets After Preprocessing with Indian Market Structural Analogues

Dataset	Instances	Features	Default Rate (%)	Imbalance Handling	Indian Market Analogue
German Credit	1,000	20	30.0	Class weighting	Tier-2/3 thin-file consumer credit
HMEQ	5,960	12	19.9	SMOTE (train only)	High-missingness AA-sourced alternative data
LendingClub	100,000	43	14.6	Class weighting	Gig economy income-volatile borrower

- **Models and Hyperparameter Optimisation**

XGBoost (Chen & Guestrin, 2016) and Random Forest (Breiman, 2001) are selected as the dominant tree-based ensemble architectures in Indian BNPL production environments, both compatible with exact TreeSHAP computation. Stratified 5-fold cross-validation is employed with Bayesian hyperparameter optimisation via Optuna (Akiba et al., 2019) conducted within each training fold, using AUC-ROC as the optimisation criterion. Performing optimisation inside the fold prevents the information leakage common in studies where validation data participates in tuning.

- **XAI Implementation and Evaluation**

SHAP values are computed using TreeSHAP (SHAP v0.44), providing exact attributions at both global and local levels. LIME explanations (lime v0.2.0.1) use 5,000 perturbation samples per instance with Ridge regression and an exponential kernel. The tabular Grad-CAM adaptation (PyTorch 2.2) wraps tree structures in a differentiable computational graph and applies ReLU-activated gradient-feature interactions.

XAI quality is assessed across three dimensions. Fidelity is the Pearson correlation between model-predicted probabilities and explanation-derived linear approximation scores. Stability is the Spearman rank correlation of feature importance rankings across ten repeated computations on 100 randomly selected test instances per dataset. Comprehensibility is rated 1–3 by a panel of Indian financial sector experts against the standard of legally defensible adverse action communication under RBI Digital Lending Guidelines. Fairness auditing measures demographic parity disparity and equalized odds disparity, with age (threshold: 30 years) as the protected attribute for German Credit and LendingClub, and second mortgage status as a socioeconomic proxy for HMEQ.

ISBN: 978-93-49468-49-8

Results and Analysis

- **Baseline Model Performance**

XGBoost consistently outperforms Random Forest across all three datasets. The largest divergence occurs on the high-dimensional LendingClub data, where XGBoost achieves AUC-ROC of 0.821 versus 0.798 for Random Forest - a gap attributable to sequential boosting's superior capture of complex feature interactions across 43 engineered dimensions. Non-overlapping bootstrap confidence intervals confirm statistical significance for LendingClub. KS statistics follow the same pattern (0.447 vs. 0.401 for LendingClub under XGBoost and Random Forest respectively). Beyond discriminative accuracy, XGBoost's sequential architecture yields more stable TreeSHAP attributions, conferring a governance advantage in addition to a performance one.

Table 2: Cross-Validated Model Performance Metrics (Mean ± 95% CI across 5 folds)

Dataset	Model	AUC-ROC	F1 (macro)	KS Statistic	Gini Coeff.
German Credit	XGBoost	0.784 ± 0.018	0.712 ± 0.021	0.412 ± 0.019	0.568 ± 0.016
German Credit	Random Forest	0.761 ± 0.020	0.694 ± 0.023	0.378 ± 0.022	0.522 ± 0.018
HMEQ	XGBoost	0.803 ± 0.014	0.741 ± 0.017	0.389 ± 0.015	0.606 ± 0.012
HMEQ	Random Forest	0.782 ± 0.016	0.723 ± 0.019	0.351 ± 0.017	0.564 ± 0.014
LendingClub	XGBoost	0.821 ± 0.009	0.763 ± 0.011	0.447 ± 0.010	0.642 ± 0.008
LendingClub	Random Forest	0.798 ± 0.011	0.741 ± 0.013	0.401 ± 0.011	0.596 ± 0.009

- **SHAP: Results and Governance Properties**

SHAP records the highest fidelity in all 18 experimental conditions: mean Pearson r of 0.91 (SD = 0.04) for XGBoost and 0.89 (SD = 0.05) for Random Forest. This performance is stable across feature dimensionalities of 12, 20, and 43, consistent with SHAP's exact computation properties. Stability scores of 0.97 confirm the deterministic reproducibility of TreeSHAP attributions across repeated runs on identical instances - directly satisfying the 'consistent, robust, and replicable' requirement articulated in RBI and OCC guidance.

Global SHAP analysis reveals governance-relevant insights. For the German Credit dataset (thin-file consumer credit proxy), account status dominates attribution (mean |SHAP| = 0.184), followed by credit history (0.141) and loan duration (0.112) - precisely the observable behavioural signals that are available for bureau-absent

ISBN: 978-93-49468-49-8

Indian borrowers. For LendingClub (gig economy proxy), XGBoost emphasises interest rate (0.231) and debt-to-income ratio (0.198), while Random Forest weights annual income (0.189) more heavily - a divergence with audit implications when the two architectures disagree on the primary drivers of an adverse decision. SHAP beeswarm analysis also reveals non-linear threshold effects invisible to simple importance rankings: in the HMEQ dataset, LTV contributions are near zero below 0.60 but sharply elevated above 0.80, a risk cliff detectable only through full SHAP decomposition.

- **LIME: Results and Governance Limitations**

LIME fidelity averages 0.76 (SD = 0.09) for XGBoost and 0.74 (SD = 0.11) for Random Forest, significantly below SHAP in all pairwise comparisons ($p < 0.001$, Bonferroni-corrected). The gap widens markedly on high-dimensional data: XGBoost on LendingClub produces LIME fidelity of 0.71 versus SHAP fidelity of 0.93 - a 0.22 difference that crosses the boundary between credible and misleading explanation. The mechanism is structural: Gaussian neighbourhood sampling cannot reliably approximate the axis-aligned, crisp decision boundaries that characterise XGBoost in high-dimensional space.

Stability presents a more severe regulatory exposure. Mean stability of 0.63 (SD = 0.18) implies that in approximately 14.2% of cases, two LIME runs on the same instance produce results where the top-ranked feature in one run does not appear among the top three in the other. No adversarial manipulation is required; this instability arises from ordinary deployment. A BNPL compliance team relying on LIME for adverse action documentation would be unable to guarantee that two examiners reviewing the same decision would receive consistent explanations. LIME's relatively high comprehensibility scores (2.1–2.3), reflecting the natural language interpretability of its linear coefficients, suggest a complementary role as a consumer communication tool, but not as a primary accountability mechanism.

- **Grad-CAM Adaptation: Results and Identified Bias**

The tabular Grad-CAM adaptation produces the lowest fidelity of all tested methods: 0.68 (SD = 0.14) for XGBoost and 0.65 (SD = 0.16) for Random Forest. At the lowest observed condition (LendingClub Random Forest, $r = 0.62$), the explanation linear projection accounts for only approximately 38% of model output variance - a level at which human review of attributed decisions is substantially uninformative.

A more consequential finding is the positive-gradient asymmetry arising from ReLU activation in the adapted architecture. Because the ReLU function zeroes all negative gradient components, features with negative contributions to credit denial are systematically suppressed in attribution rankings. In the German Credit dataset - structural proxy for India's thin-file borrower segment - Grad-CAM surfaces

ISBN: 978-93-49468-49-8

occupational status within the minority category among the top five drivers of adverse decisions in 68% of test cases, while SHAP places the same feature outside the top ten in 84% of the same cases. In an Indian production context, the structural equivalents are gig worker occupation category and geographic tier - features that correlate with constitutionally protected attributes. This is, to the authors' knowledge, the first documented instance of this bias in the credit scoring context. Indian BNPL providers using tabular Grad-CAM as their primary XAI tool face both regulatory risk under the RBI fair practices code and the risk of systematic consumer harm from biased adverse action attribution.

Table 3: XAI Evaluation Metrics by Method, Model, and Dataset

XAI Method	Model	Dataset	Fidelity (r)	Stability (ρ)	Comprehensibility (1–3)
SHAP	XGBoost	German Credit	0.93	0.97	2.4
SHAP	XGBoost	HMEQ	0.91	0.97	2.3
SHAP	XGBoost	LendingClub	0.93	0.98	2.3
SHAP	Random Forest	German Credit	0.90	0.96	2.3
SHAP	Random Forest	HMEQ	0.89	0.95	2.4
SHAP	Random Forest	LendingClub	0.88	0.96	2.3
LIME	XGBoost	German Credit	0.79	0.68	2.3
LIME	XGBoost	HMEQ	0.77	0.65	2.2
LIME	XGBoost	LendingClub	0.71	0.59	2.1
LIME	Random Forest	German Credit	0.77	0.66	2.2
LIME	Random Forest	HMEQ	0.74	0.61	2.2
LIME	Random Forest	LendingClub	0.70	0.57	2.1
Grad-CAM	XGBoost	German Credit	0.71	0.74	1.9
Grad-CAM	XGBoost	HMEQ	0.69	0.72	1.8
Grad-CAM	XGBoost	LendingClub	0.65	0.68	1.7
Grad-CAM	Random Forest	German Credit	0.68	0.71	1.8
Grad-CAM	Random Forest	HMEQ	0.65	0.70	1.8
Grad-CAM	Random Forest	LendingClub	0.62	0.66	1.7

ISBN: 978-93-49468-49-8

- **Fairness Audit**

Statistically significant demographic parity disparity is present in two of three datasets. For German Credit, borrowers aged 30 or above receive positive credit decisions at materially higher rates than those under 30: XGBoost shows a 14.3 percentage point gap (73.2% vs. 58.9%; $z = 4.81$, $p < 0.001$), with a selection ratio of 0.804 - marginally above the legal 80% rule threshold. Random Forest produces a 12.8 percentage point gap (72.1% vs. 59.3%; $z = 4.22$, $p < 0.001$). Given that the dominant BNPL user demographic in India is aged 18–34, this pattern represents systematic exclusion of the very population the product is designed to serve.

For the LendingClub gig economy proxy, equalized odds disparities of 0.082 (true positive rate) and 0.071 (false positive rate) are observed between younger and older borrowers ($p < 0.01$). SHAP decomposition reveals that 43% of the age-associated decision variance flows through annual income and employment tenure - features that are proxies for age in this population and structural analogues to gig platform experience and income regularity in Indian scoring models. This pathway-level attribution evidence, achievable only through SHAP's additive decomposition, is precisely the form of documentation required by RBI fair practice guidelines and anticipated DPDP Act explanation rights. LIME's local surrogates and Grad-CAM's asymmetric gradients cannot generate equivalent pathway evidence. The HMEQ dataset shows no significant demographic disparity ($z = 1.23$, $p = 0.219$), consistent with the hypothesis that collateral-based products reduce behavioural demographic proxies relative to unsecured BNPL products.

Robustness checks confirm the primary findings. Temporal train-test separation on LendingClub produces an XGBoost AUC-ROC decline of 3.1 percentage points and a SHAP fidelity decline of 0.04, versus a 0.09 LIME fidelity decline under the same distributional shift - confirming that LIME degrades fastest during the portfolio expansion scenarios most common in Indian BNPL. SHAP importance rankings maintain Spearman correlations above 0.94 across three SMOTE oversampling rates, demonstrating robustness to preprocessing choices. Reweighing (Kamiran & Calders, 2012) reduces the German Credit demographic parity gap from 14.3 to 4.7 percentage points, confirming that single-technique bias mitigation is insufficient for equitable credit outcomes in thin-file markets.

Discussion and Governance Implications

- **Theoretical Contributions**

The AAA framework advances trustworthy AI research by repositioning explainability from a post-hoc add-on to an architectural governance criterion. Where existing frameworks treat XAI as a communication device, the AAA architecture specifies what XAI must deliver across autonomy, accountability, and auditability dimensions before qualifying as governance infrastructure.

ISBN: 978-93-49468-49-8

The decomposition of interpretability into three independently measurable components - fidelity, stability, comprehensibility - extends Rudin's (2019) accuracy-interpretability trade-off. The components are not uniformly correlated: all three methods score highest on comprehensibility (1.7–2.4), yet the fidelity gap between SHAP and LIME (0.91 vs. 0.76) and the stability gap (0.97 vs. 0.63) would be invisible under a unidimensional interpretability requirement. Regulatory frameworks that mandate only 'understandability' would systematically favour LIME despite SHAP's superior governance properties - exactly the unintended consequence the tri-dimensional framework prevents.

The principal-agent extension generates a testable prediction confirmed by the fidelity data: XAI methods that distort agent reasoning produce artificial transparency. At fidelity r = 0.65 (Grad-CAM, LendingClub Random Forest), the explanation accounts for only ~42% of model output variance. A risk officer reviewing decisions through such attributions operates with information less than half as accurate as the model's own outputs - a governance failure that is invisible without fidelity measurement.

- **Practical Governance Architecture**

The empirical findings support a three-tier governance architecture for Indian BNPL scoring pipelines. TreeSHAP should serve as the primary XAI mechanism for all production tree-based models: it is the only tested approach to simultaneously achieve the minimum fidelity threshold of 0.88 (derived from the lowest SHAP-Random Forest condition), deterministic stability above 0.95, and exact additive attribution that satisfies RBI auditability, anticipated DPDP Act consistency requirements, and EU AI Act Article 13 interpretability standards. LIME may serve a secondary role as a consumer-facing communication instrument, given its higher comprehensibility and natural adverse action notice vocabulary, provided that its instability limitations are explicitly documented in model validation reports and it is not used as primary accountability evidence. Tabular Grad-CAM should not be used as a primary XAI tool for tree-based credit models: the positive-gradient asymmetry creates protected attribute surfacing risk that is incompatible with the RBI fair practices code regardless of model accuracy.

The multi-jurisdictional regulatory mapping reveals convergence across all four regulatory regimes on three core requirements: instance-level attribution, consistency across runs, and human comprehensibility. The primary divergences lie in implementation mechanism (supervisory guidance in India versus enforceable regulation in the EU), data scope (personal data emphasis in DPDP Act versus systems-level approach in EU AI Act), and fairness provisions (explicit non-discrimination requirements in EU AI Act without direct equivalents in India's current AI governance framework). An integrated Indian explainability standard benchmarked against the strictest applicable requirement across all mapped jurisdictions - EU AI

ISBN: 978-93-49468-49-8

Act Article 13 for interpretability, RBI DLG for auditability - would enable multinational BNPL operators to comply across all relevant regulatory environments without duplicating compliance architectures.

The fairness audit findings carry a specific policy implication for India's developing BNPL regulatory environment: bias mitigation must be multi-layered. Algorithmic reweighing reduced the German Credit demographic parity gap from 14.3 to 4.7 percentage points - substantial progress, but not elimination. Achieving equitable credit access in the thin-file market requires combining pre-processing interventions, in-processing fairness constraints, post-processing threshold adjustment, and continuous monitoring - the four-layer approach embedded in the AAA framework's auditability dimension.

Limitations

Three boundaries on the study's scope warrant explicit acknowledgment. The three benchmark datasets, while selected as structural surrogates for Indian BNPL lending characteristics, do not contain actual UPI transaction vectors, AA-sourced financial profiles, gig income logs, mobile recharge behaviour, or Tier-2/3 geographic demographics. Whether XAI performance rankings persist in 150–200-dimensional AA-sourced datasets remains an open empirical question. The Grad-CAM adaptation follows Agarwal et al. (2021); alternative implementations using attention mechanisms or integrated gradients could produce different fidelity and bias profiles. Comprehensibility ratings are based on panels of three domain experts per discipline, which limits the statistical power of that particular dimension.

Conclusion

This study offers the first empirical evaluation of XAI frameworks for agentic AI credit scoring in Indian embedded finance. Across 18 experimental conditions spanning three datasets, two model architectures, and three XAI methods, the findings establish a clear hierarchy: SHAP functions as a governance mechanism, meeting fidelity, stability, and regulatory alignment requirements across all conditions; LIME operates as a communication tool but fails the reproducibility standards required for accountability governance; and tabular Grad-CAM carries an inherent positive-gradient bias that creates protected attribute surfacing risk incompatible with fair lending obligations.

The proposed AAA (Autonomy–Accountability–Auditability) framework provides an organising governance architecture that moves the conversation beyond post-hoc explanation toward systemic oversight of agentic AI pipelines. The framework identifies specific, measurable requirements for each governance dimension and demonstrates alignment with the RBI Digital Lending Guidelines, the anticipated DPDP Act implementation rules, and the EU AI Act, providing a multi-jurisdictional compliance pathway for India's BNPL sector.

ISBN: 978-93-49468-49-8

Five research priorities emerge from this work: empirical validation on proprietary Indian datasets containing AA and UPI features; systematic comparison of inherently interpretable models (optimal decision trees, explainable boosting machines) against the accuracy-fidelity benchmarks established here; evaluation of LLM-generated adverse action letters derived from SHAP attribution vectors against the comprehensibility criteria developed in this paper; XAI drift studies tracking explanation quality during the distributional shifts that accompany rapid portfolio growth; and regulatory mapping across the full spectrum of jurisdictions in which Indian FinTech operators are active. India's digital public infrastructure provides the foundation for a genuinely inclusive credit system. The governance architecture for the AI layer operating above that infrastructure remains under construction - this paper offers empirical benchmarks, a governance framework, and regulatory alignment analysis to support that construction.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.
2. Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. (2021). Neural additive models: Interpretable machine learning with neural nets. NeurIPS, 34, 4699–4711.
3. Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488–500.
4. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 589–609.
5. Arya, V., et al. (2019). One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. arXiv:1909.03012.
6. Banerjee, A., & Duflo, E. (2007). The economic lives of the poor. Journal of Economic Perspectives, 21(1), 141–167.
7. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56.
8. Berg, T., Burg, V., Gombovic, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. Review of Financial Studies, 33(7), 2845–2897.
9. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
10. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216.

ISBN: 978-93-49468-49-8

11. Chawla, N. V., et al. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
12. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. ACM SIGKDD, 785–794.
13. Consumer Financial Protection Bureau. (2022). Consumer financial protection circular 2022-03. CFPB.
14. Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable AI for interpretable credit scoring. ICAART, 2, 220–228.
15. European Commission. (2019). Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence.
16. European Commission. (2024). Regulation (EU) 2024/1689 - Artificial Intelligence Act. Official Journal of the European Union.
17. European Parliament. (2016). Regulation (EU) 2016/679 - General Data Protection Regulation. Official Journal of the European Union.
18. Feldman, M., et al. (2015). Certifying and removing disparate impact. ACM SIGKDD, 259–268.
19. Floridi, L., et al. (2019). An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707.
20. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.
21. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. Journal of Finance, 77(1), 5–47.
22. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. NeurIPS, 35, 507–520.
23. Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository.
24. Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360.
25. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33.
26. Khandani, A. E., Kim, A. J., & Lo, A. W. (2020). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767–2787.

ISBN: 978-93-49468-49-8

27. KPMG India. (2024). The India BNPL report 2024: Growth, risk, and the regulatory horizon. KPMG Advisory Services.
28. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. European Journal of Operational Research, 247(1), 124–136.
29. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.
30. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30, 4765–4774.
31. Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.
32. Ministry of Electronics and Information Technology. (2023). Digital Personal Data Protection Act, 2023. Government of India Gazette.
33. Misheva, B. H., et al. (2021). Explainable AI in credit risk management. arXiv:2103.00691.
34. National Payments Corporation of India. (2024). UPI ecosystem statistics FY2023–24.
35. NITI Aayog. (2022). Digital banking: A proposal for licensing and regulatory regime for India.
36. Redseer Consulting. (2024). BNPL India market report 2024.
37. Reserve Bank of India. (2022). Guidelines on digital lending (RBI/2022-23/111).
38. Reserve Bank of India. (2023). Report of the working group on digital lending.
39. Reserve Bank of India. (2024). Discussion paper on responsible AI and machine learning in financial services.
40. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. ACM SIGKDD, 1135–1144.
41. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
42. Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.
43. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV, 618–626.
44. Shapley, L. S. (1953). A value for n-person games. In Contributions to the theory of games (Vol. II). Princeton University Press.

ISBN: 978-93-49468-49-8

45. Slack, D., et al. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. AAAI/ACM AIES, 180–186.

46. Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. American Economic Review, 71(3), 393–410.

47. Thomas, L. C., Edelman, D. B., & Crook, J. N. (2017). Credit scoring and its applications (2nd ed.). SIAM.

48. TransUnion CIBIL. (2024). Credit market indicator report Q1 2024.

49. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 31(2), 841–887.

50. Wooldridge, M. (2020). The road to conscious machines: The story of AI. Pelican Books.

51. Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480.

ISBN: 978-93-49468-49-8