Data Intelligence and Informatics Practice

Review
Volume 1(1), 2026, 42−55
https://doi.org/10.53623/diip.v1i1.1273

Comparative Analysis of Machine Learning Models for Credit Risk Assessment in FinTech Applications

Ayesha Tariq, Muhammad Hassan Raza*

COMSATS University Islamabad, Islamabad 45550, Pakistan

*Corresponding author: m.hassan.raza44@gmail.com

SUBMITTED: 23 June 2026; REVISED: 31 July 2026; ACCEPTED: 2 August 2026

ABSTRACT: Credit risk assessment is a core function of financial technology (FinTech) platforms, underpinning lending decisions, pricing, and portfolio risk management. Over the past decade, machine learning (ML) has progressively displaced traditional statistical scorecards as the dominant analytical paradigm, driven by the availability of large, high-dimensional, and often alternative data sources together with the need for higher predictive accuracy. This review provides a comparative analysis of ML models used for credit risk assessment in FinTech applications, spanning traditional statistical classifiers, tree-based ensemble methods, and deep learning architectures. Following a structured literature search and screening process, we synthesize empirical evidence on predictive performance across benchmark and proprietary FinTech datasets, critically contrast model families under varying data conditions, and examine the trade-off between predictive power and interpretability that shapes real-world adoption. We further discuss the role of alternative data and explainable artificial intelligence (XAI) in expanding financial inclusion while preserving regulatory compliance, distinguishing global from local interpretability, and we outline open challenges, including data imbalance, concept drift, algorithmic fairness, privacy, and model governance. The review concludes with directions for future research, emphasizing hybrid and federated learning architectures, standardized benchmarking protocols, and explainability-by-design frameworks for credit risk applications in emerging FinTech markets.

KEYWORDS: Credit risk assessment; machine learning; FinTech; credit scoring; ensemble learning; deep learning

1. Introduction

Credit risk, defined as the probability that a borrower will fail to meet contractual repayment obligations, has historically been managed through statistical scorecards built on logistic regression and discriminant analysis [1, 2]. These models have long been favoured by lenders and regulators because their linear, additive structure produces transparent and auditable decisions [1]. However, the emergence of financial technology (FinTech) has fundamentally altered both the scale and the nature of the credit assessment problem. FinTech lenders, including peer-to-peer (P2P) platforms, digital banks, and mobile-money providers, originate loans rapidly, often to thin-file or unbanked borrowers, and increasingly rely on non-traditional,

high-dimensional data such as digital footprints, transaction histories, and behavioural signals rather than conventional credit bureau records [3, 4].

This shift toward alternative data and automated underwriting has, in turn, catalysed the adoption of machine learning (ML) techniques capable of capturing non-linear relationships and complex feature interactions that traditional models cannot represent [5, 6]. Tree-based ensembles such as random forest and gradient boosting, together with deep neural architectures including multilayer perceptrons, convolutional networks, and recurrent networks, have repeatedly demonstrated superior discriminatory power over logistic regression benchmarks across diverse credit datasets [6−8]. At the same time, the FinTech credit environment introduces challenges that are less pronounced in traditional banking, including severe class imbalance between defaulting and non-defaulting borrowers [9, 10], rapid concept drift in borrower behaviour, and heightened scrutiny from regulators regarding fairness, transparency, and explainability of automated decisions [2, 11].

A substantial and growing body of literature has compared the performance of individual ML algorithms on specific datasets, yet comparative evidence remains fragmented across disparate model families, evaluation protocols, and FinTech use cases such as P2P lending, digital microcredit, and SME financing [7, 12, 13]. Existing systematic reviews have tended to focus narrowly on either algorithmic benchmarking [6] or on financial inclusion and alternative data [3, 14], leaving a gap for a review that jointly synthesizes model comparison, performance evaluation methodology, and the operational and regulatory challenges specific to FinTech deployment.

This review distinguished itself from prior comparative and systematic reviews in three ways. Lessmann et al. [6] provided an influential algorithmic benchmarking study but predated the widespread adoption of alternative data and deep learning architectures in FinTech, and therefore did not address explainability or fairness. Shi et al. [12] and Tian et al. [13] presented broader systematic reviews of machine learning in credit risk and internet financial risk management, respectively, but primarily synthesized methodological trends rather than critically comparing model families under different data conditions. Roy and Vasa [5] reviewed machine learning and artificial intelligence methods for FinTech credit scoring and risk management, whereas Berg et al. [3] and Dushimimana et al. [14] focused specifically on alternative data and financial inclusion rather than cross-model performance. In contrast, this review (i) compared the performance of major machine learning model families across three data conditions, namely traditional bureau-scale datasets, alternative-data-rich FinTech datasets, and highly imbalanced thin-file populations; (ii) synthesized dataset characteristics, validation methods, explainability techniques, fairness considerations, and predictive performance for representative comparative studies; and (iii) integrated explainability, fairness, privacy, and governance issues by distinguishing global and local interpretability and evaluating their implications for regulatory compliance within a unified framework for FinTech applications.

This review addressed these gaps through four objectives. First, it categorized and compared the principal families of machine learning models used for credit risk assessment, including statistical classifiers, ensemble and boosting methods, and deep learning architectures, highlighting their respective strengths and limitations. Second, it synthesized comparative evidence on predictive performance using widely adopted evaluation metrics, including the area under the receiver operating characteristic curve (AUC), the Kolmogorov–

Smirnov (KS) statistic, and F1-score, across FinTech and benchmark credit datasets under different data conditions. Third, it examined FinTech-specific applications and challenges, including the use of alternative data for financial inclusion, the regulatory and ethical implications of automated credit decisions, and the role of explainable artificial intelligence (XAI) in balancing predictive accuracy with interpretability. Finally, it identified emerging developments, including blockchain-enabled federated credit scoring and large language model-based risk assessment, and proposed directions for future research.

2. Review Methodology

2.1. Search Strategy.

Literature was identified through structured searches of Scopus, Web of Science, IEEE Xplore, and ScienceDirect. Search strings combined the terms “credit risk”, “credit scoring”, “machine learning”, “FinTech”, “peer-to-peer lending”, “deep learning”, “alternative data”, and “explainable artificial intelligence”, combined using Boolean AND/OR operators and restricted to titles, abstracts, and keywords. The primary search window covered publications from 2012 to 2025, capturing the period of rapid FinTech and ML-based credit scoring growth, with a small number of earlier foundational studies on statistical credit scoring retained where directly relevant to establishing the traditional baseline against which ML methods are compared.

2.2. Screening and Eligibility Criteria.

The overall study identification, screening, eligibility assessment, and inclusion process is illustrated in Figure 1. The initial database search identified 410 records. After removing duplicates, 312 unique articles remained and were screened based on their titles and abstracts.

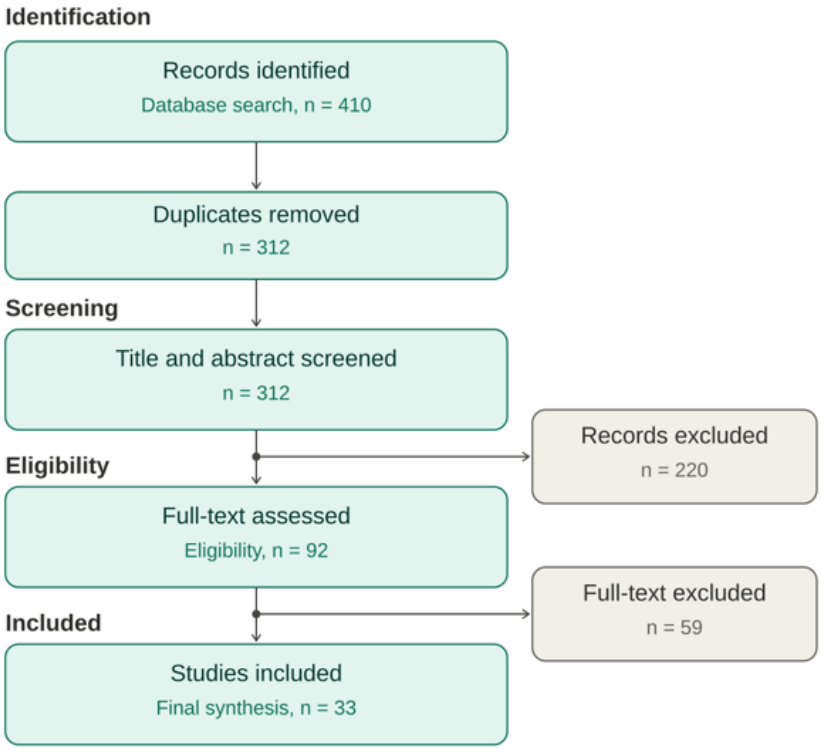

Figure 1. PRISMA flow diagram of the literature search and study selection process.

Studies were included if they (i) applied or compared one or more machine learning models for credit risk assessment, credit scoring, or default prediction within FinTech, peer-to-peer (P2P) lending, digital lending, or comparable retail and SME credit contexts; (ii) were published in peer-reviewed journals; (iii) reported empirical performance results, methodological developments, or structured review findings; and (iv) were available in full-text English. Studies were excluded if they focused exclusively on conventional banking without FinTech or alternative-data relevance, were conference-only publications, lacked sufficient methodological detail, or duplicated findings reported in other eligible studies. Following title and abstract screening, 92 articles underwent full-text assessment against the eligibility criteria. After excluding studies that did not meet the inclusion requirements, 33 articles were retained for qualitative synthesis. This systematic selection process ensured that the review incorporated high-quality and relevant evidence while minimizing duplication and methodological bias.

3. Machine Learning Model Families for Credit Risk Assessment.

Table 1 summarizes the principal machine learning model families used in credit risk assessment, their representative algorithms, key strengths, and limitations. These models differ substantially in terms of predictive capability, interpretability, computational requirements, and suitability for different credit data environments.

Table 1. Comparative overview of major machine learning model families used in credit risk assessment.

Model family	Representative algorithms	Key strengths	Key limitations
Statistical / linear	Logistic regression, linear discriminant analysis, ANFIS	High interpretability; regulatory acceptance; low computational cost	Assumes linear separability; weaker discrimination on non-linear, high-dimensional data
Tree-based ensemble	Random forest, XGBoost, LightGBM, gradient boosting	Captures non-linear interactions; robust to outliers; strong tabular performance	Less transparent; risk of overfitting without tuning; sensitive to class imbalance
Deep learning	MLP, CNN, LSTM, autoencoders	Learns hierarchical representations; effective with large/alternative data and sequential history	Requires large training data; computationally intensive; limited native interpretability
Hybrid / statistical-ensemble	U-MIDAS logit with group lasso, fuzzy-neural hybrids	Balances interpretability with variable selection and non-linearity	Added model complexity; fewer standardized implementations

3.1. Traditional Statistical and Linear Classifiers.

Logistic regression remains the most widely deployed credit scoring technique in both traditional banking and FinTech because its coefficients map directly onto interpretable odds ratios, supporting regulatory scorecards and reason-code generation for adverse action notices [1, 2, 15]. Linear discriminant analysis (LDA) and related parametric classifiers share this transparency advantage but assume linear separability and normally distributed predictors, assumptions frequently violated by skewed financial and behavioural variables [6, 16]. Hybrid extensions, such as the Adaptive Neuro-Fuzzy Inference System (ANFIS), combine fuzzy rule bases with neural learning to relax the linearity assumption while retaining a degree of rule-based interpretability, and have shown improved classification accuracy over pure logistic baselines on credit card datasets [16]. Support vector machine variants, including fuzzy

entropy-weighted formulations applied to peer-to-peer lending investment decisions, have also been used to introduce non-linear decision boundaries while retaining a degree of margin-based interpretability [17]. Despite consistently weaker discriminatory performance relative to non-linear alternatives in large-scale benchmarking studies [6], statistical classifiers continue to serve as the baseline against which newer ML methods are evaluated, and frequently remain the model of choice where regulatory transparency outweighs marginal accuracy gains.

3.2. Tree-Based Ensemble and Boosting Methods.

Ensemble methods aggregate multiple weak learners, typically decision trees, to produce classifiers with substantially improved generalization relative to single models. Random forest constructs an ensemble of de-correlated trees via bootstrap aggregation and random feature subsampling, and has been shown to match or exceed logistic regression accuracy across multiple credit and loan-approval datasets [12]. Gradient boosting variants, including XGBoost and LightGBM, build trees sequentially to minimize residual error and have become the dominant ensemble approach in recent FinTech credit-scoring studies owing to their regularization mechanisms and computational efficiency on large tabular datasets [7, 18]. Comparative studies applying logistic regression, random forest, and XGBoost to FinTech personal-loan data consistently report gradient boosting achieving the highest accuracy and recall, with random forest a close second, while logistic regression lags on both discrimination and recall under class imbalance [18−20]. Hybrid statistical-boosting models, such as the U-MIDAS logit framework with group lasso penalty, have further been proposed to combine the interpretability of regression with the variable-selection capacity of regularized ensembles for corporate default prediction [21]. Similarly, hybrid architectures such as the time series analysis-based multilayer perceptron (TSA-MLP) have been proposed to identify credit default problems by combining engineered temporal features with neural classification layers [22].

3.3. Deep Learning Architectures.

Deep learning models extended the non-linear modelling capability of ensemble methods by learning hierarchical feature representations directly from raw or minimally engineered data, building on advances in artificial neural network theory and pattern recognition [23]. Multilayer perceptrons (MLPs) and behavioural deep credit-scoring networks demonstrated strong predictive performance on transactional and behavioural banking datasets, frequently outperforming both logistic regression and classical ensemble methods when sufficient training data was available [4]. Convolutional neural networks (CNNs) were adapted for credit scoring by transforming tabular data into pixel-matrix representations, enabling the extraction of local feature interactions while maintaining post-hoc explainability through saliency-based techniques [8]. Recurrent architectures, particularly long short-term memory (LSTM) networks, proved well suited to sequential repayment-history data and were applied to forecast platform-level default rates in peer-to-peer lending with greater temporal sensitivity than static classifiers [24]. Although deep learning models often achieved the highest predictive accuracy on large, alternative-data-rich FinTech datasets, their limited interpretability and greater computational requirements remained important challenges for practical deployment.

4. Comparative Performance Analysis

4.1. Evaluation Metrics and Benchmark Datasets.

Comparative evaluation of credit risk models relies on a consistent set of metrics that capture both overall classification accuracy and discriminatory power under class imbalance. Accuracy, precision, recall, and F1-score quantify classification correctness at a chosen decision threshold, whereas the area under the receiver operating characteristic curve (AUC) and the Kolmogorov–Smirnov (KS) statistic assess threshold-independent discriminatory ability, the latter being a long-standing industry standard for scorecard validation [1]. Large-scale benchmarking exercises spanning multiple real-world credit datasets have established that AUC and the Brier score together provide a more reliable basis for model comparison than accuracy alone, since accuracy can be misleadingly inflated by the majority non-default class [6, 25]. FinTech-specific studies commonly draw on Lending Club and analogous P2P platform data, German and Australian credit benchmark datasets, and proprietary digital-lender portfolios, with sample sizes and default rates varying substantially across studies, which complicates direct cross-study comparison [18, 19].

4.2. Synthesis of Comparative Performance Evidence.

Across the reviewed comparative studies, a consistent ranking pattern emerged in which gradient boosting methods (XGBoost and LightGBM) and random forest generally achieved the highest AUC and classification accuracy, while deep learning architectures performed competitively and occasionally outperformed ensemble methods on large alternative-data-rich datasets. In contrast, logistic regression consistently exhibited lower discriminatory performance but remained competitive in terms of calibration, transparency, and regulatory interpretability [6, 7]. As summarized in Table 2, large-scale benchmarking studies demonstrated that heterogeneous ensembles and random forest consistently ranked among the best-performing classifiers across multiple retail credit datasets [6]. Studies using Chinese and European P2P lending datasets further reported prediction accuracies exceeding 90%, with random forest, gradient boosting, and deep neural networks outperforming conventional statistical models [18, 19]. For sequential repayment data, long short-term memory (LSTM) networks produced lower forecasting errors than traditional time-series approaches, highlighting the advantage of deep learning for modelling temporal credit-risk patterns [24]. Furthermore, convolutional neural network (CNN)-based approaches enhanced predictive performance while incorporating saliency-based explainable AI methods to improve model transparency [8]. Despite these advances, Table 2 also indicates that most studies relied on single-platform datasets, reported limited external validation, and rarely evaluated fairness or algorithmic bias, highlighting important gaps in the current literature. Complementary analyses of P2P lending data identified loan purpose, credit grade, and borrower indebtedness as consistently influential predictors of default across different modelling approaches [26, 27]. The representative studies presented in Table 2 illustrate the diversity of datasets, validation strategies, explainability techniques, and methodological limitations discussed throughout this review, rather than providing an exhaustive summary of all 33 included studies.

Table 2. Representative comparative findings from reviewed studies on machine learning models for credit risk and default prediction.

Dataset characteristics	Validation method	Models compared / best performer	Reported metrics	Explainability approach	Fairness / key limitations	Reference
8 real-world retail credit-scoring datasets; varying size and default rates	10-fold cross-validation; H-measure ranking	41 classifiers incl. logistic regression, RF, boosting, SVM; heterogeneous ensembles / RF top-ranked	AUC, H-measure, Brier score	Not addressed (pre-dates XAI benchmarking)	No fairness auditing; findings pre-date alternative-data FinTech contexts	[6]
Renrendai Chinese P2P platform; large-scale loan-level data	Train/test split with stratified sampling	RF, XGBoost, GBM, neural network; random forest best	Accuracy > 90%, recall	Not addressed	Single-platform data; limited external validity	[18]
P2P lending data (UK/EU)	Train/test split; AUC comparison	Logistic regression, SVM, deep neural networks; deep neural network best	AUC	Not addressed	Aggregate platform-level features; limited borrower-level granularity	[19]
P2P platform default-rate time series	Rolling-window time-series validation	LSTM vs. statistical time-series models; LSTM best	Forecast error (RMSE)	Not addressed	Platform-level, not loan-level; limited transferability across platforms	[24]
Public credit-scoring benchmark datasets (image-transformed tabular features)	Train/test split with cross-validation	CNN (image-transformed tabular data) vs. classical ML; CNN with XAI layer best	AUC, accuracy	Saliency-based XAI layer integrated with CNN	Explainability method specific to image-transformed representation; not directly comparable to SHAP/LIME	[8]

4.3. Critical Comparison Across Data Conditions.

The reviewed comparative evidence indicates that the relative ranking of model families is not fixed but shifts systematically with the underlying data conditions. On small-to-moderate, low-dimensional bureau-style datasets such as the German and Australian credit benchmarks, logistic regression and simple ensembles remain highly competitive with more complex alternatives, and the marginal AUC gains from gradient boosting or deep learning are often modest relative to the added complexity and reduced transparency [6]. On large, high-dimensional, alternative-data-rich FinTech datasets, such as P2P lending platforms and digital-footprint-based lenders, the advantage of tree-based ensembles and deep learning widens substantially, with gradient boosting and random forest achieving accuracy above 90 percent and deep neural networks matching or exceeding ensembles when sufficient training volume is available [18, 19]. Under severe class imbalance and thin-file conditions typical of emerging-market FinTech lending, model family robustness diverges further: ensemble and boosting methods tolerate moderate imbalance without adjustment, whereas logistic regression and, to a lesser extent, deep learning architectures require explicit resampling or cost-sensitive adjustment to maintain acceptable recall on the minority default class [9, 10]. These data-condition-dependent patterns suggest that model selection in FinTech credit risk assessment should be guided by dataset scale, dimensionality, and class balance rather than by a universal

ranking of model families, reinforcing the case for standardized, multi-condition benchmarking protocols.

4.4. Class Imbalance and Robustness Considerations.

Default events are rare relative to non-default observations in most credit portfolios, and this imbalance materially affects model evaluation and selection, a recurring theme in broader reviews of financial distress and corporate failure prediction [28]. Resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been shown to improve recall for minority default classes in P2P lending without substantially degrading precision [9], while cost-sensitive learning approaches that assign asymmetric misclassification costs to false negatives directly align model optimization with the asymmetric financial consequences of approving a defaulting borrower [10]. Comparative evidence indicates that ensemble and boosting methods are generally more robust to moderate imbalance than logistic regression, but that severe imbalance still requires explicit resampling or cost-sensitive adjustment regardless of model family, and that reported performance gains from any single technique are highly dataset-dependent, reinforcing the need for standardized benchmarking protocols when comparing studies across the literature [6].

5. FinTech Applications, Challenges, and Emerging Trends

5.1. Alternative Data and Financial Inclusion.

A defining feature of FinTech credit risk assessment is the use of alternative data sources, including digital footprints, mobile-phone usage patterns, and social and transactional metadata, to score borrowers who lack sufficient traditional credit-bureau history. Evidence from a major European digital lender shows that simple digital-footprint variables, such as device type, e-mail provider, and time of application, carry information content comparable to traditional credit bureau scores for predicting consumer default [3]. Similarly, machine learning models trained on airtime mobile-loan repayment data have been used to construct credit scores for unbanked populations in emerging markets, demonstrating that alternative behavioural data can substitute for missing bureau records without a substantial loss of predictive accuracy [14]. FinTech lenders relying on such automated, ML-driven underwriting pipelines have also been shown to originate loans with measurable efficiency gains relative to traditional commercial banks [29]. These findings support the broader proposition that ML-enabled alternative credit scoring can expand financial inclusion, particularly in markets, such as Pakistan and other South Asian economies, where formal credit history coverage remains limited and informal or thin-file borrowers represent a large share of potential FinTech customers [30].

5.2. Regulatory, Ethical, and Interpretability Challenges.

The predictive advantages of ensemble and deep learning models are tempered by reduced transparency relative to traditional scorecards, creating tension with regulatory requirements for explainable, non-discriminatory lending decisions. Explainable artificial intelligence (XAI) techniques, including SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), have been integrated into credit-scoring pipelines to provide

post-hoc, instance-level justifications for otherwise opaque model outputs, enabling lenders to generate adverse-action reason codes from black-box predictions [8,11].

Explainability approaches in credit risk applications can be usefully distinguished along two dimensions. Global interpretability methods, such as permutation feature importance, partial dependence plots, and global surrogate models, characterize overall model behaviour across the full borrower population and are primarily used for model validation, ongoing performance monitoring, and the documentation required by model risk management frameworks. Local interpretability methods, including SHAP and LIME, instead generate instance-level explanations for individual credit decisions and are the primary technical mechanism for producing borrower-facing adverse-action reason codes required under regulations such as the U.S. Equal Credit Opportunity Act (Regulation B) and analogous provisions in other jurisdictions, as well as for supporting the general right to explanation associated with automated decision-making under data protection regimes such as the EU General Data Protection Regulation. Effective regulatory compliance in FinTech credit risk assessment therefore typically requires both interpretability layers: global explanations to satisfy model governance and audit requirements, and local explanations to satisfy individual borrower-facing transparency obligations [8,11,31].

Within FinTech lending specifically, explainability frameworks have been used to reconcile lender risk-based pricing decisions with borrower-facing transparency requirements [31]. Beyond interpretability, FinTech credit models raise concerns regarding algorithmic fairness, since alternative data variables can act as proxies for protected characteristics and inadvertently encode disparate impact across demographic groups, an issue that has motivated explicit fairness-aware model auditing using group fairness metrics such as demographic parity and equalized odds, alongside conventional performance evaluation [13]. Data privacy and cross-border data governance present additional constraints, particularly where alternative data is sourced from third-party digital platforms, requiring FinTech lenders to reconcile predictive performance with compliance obligations under evolving financial-sector regulation, including data minimization principles and, increasingly, privacy-preserving architectures such as the federated learning approaches discussed in Section 5.3 [32].

5.3. Emerging Trends: Federated Learning, Blockchain, and Hybrid Architectures.

Recent research has begun to address the data-sharing and trust constraints inherent in cross-institution credit scoring through federated learning architectures, which allow multiple FinTech lenders or credit bureaus to jointly train shared models without centralizing sensitive borrower data, and blockchain-based frameworks that combine federated learning with explainability layers to provide auditable, decentralized credit-scoring infrastructure [32]. Internet financial risk management research more broadly highlights a shift toward hybrid pipelines that integrate traditional risk indicators with machine-learning-derived behavioural and network features, alongside calls for more standardized, reproducible evaluation frameworks across the fragmented FinTech risk-management literature [13]. Text-based soft information extracted from loan descriptions and borrower narratives has similarly been shown to improve default prediction beyond structured financial variables alone [33]. Looking forward, the integration of large-scale alternative data, federated and privacy-preserving learning, and explainability-by-design principles is likely to define the next generation of credit risk assessment systems in FinTech, particularly in emerging markets seeking to balance rapid

digital-lending growth with financial-inclusion and regulatory objectives. Common evaluation metrics used in machine learning-based credit risk studies is shown in Table 3.

Table 3. Common evaluation metrics used in machine learning-based credit risk studies.

Metric	Description	Interpretation in credit risk context
Accuracy	Proportion of correctly classified borrowers (defaulters and non-defaulters) out of all classified cases	Can be misleading under severe class imbalance typical of default data
Precision / Recall	Precision = correctly predicted defaulters / all predicted defaulters; recall = correctly predicted defaulters / all actual defaulters	Recall is critical for minimizing missed defaulters; precision controls false alarms
F1-score	Harmonic mean of precision and recall	Balances false positives and false negatives for the minority default class
AUC-ROC	Area under the receiver operating characteristic curve, plotting true positive rate against false positive rate	Threshold-independent measure of overall discriminatory power; widely used for model ranking
Kolmogorov–Smirnov (KS) statistic	Maximum separation between cumulative distributions of predicted scores for defaulters and non-defaulters	Traditional scorecard validation metric; high KS indicates strong class separation

Figure 2 presents the end-to-end machine learning workflow synthesized from the reviewed literature, illustrating that effective credit risk assessment extends beyond model development to encompass data acquisition, preprocessing, model training, evaluation, explainability, deployment, governance, and continuous monitoring. Rather than representing a linear process, the workflow incorporates an iterative feedback loop in which model performance is continuously monitored for data drift and concept drift, enabling periodic retraining and recalibration as new credit data become available. This lifecycle perspective reflects the growing emphasis on responsible AI deployment in FinTech, where predictive performance, regulatory compliance, and operational reliability must be maintained throughout model implementation.

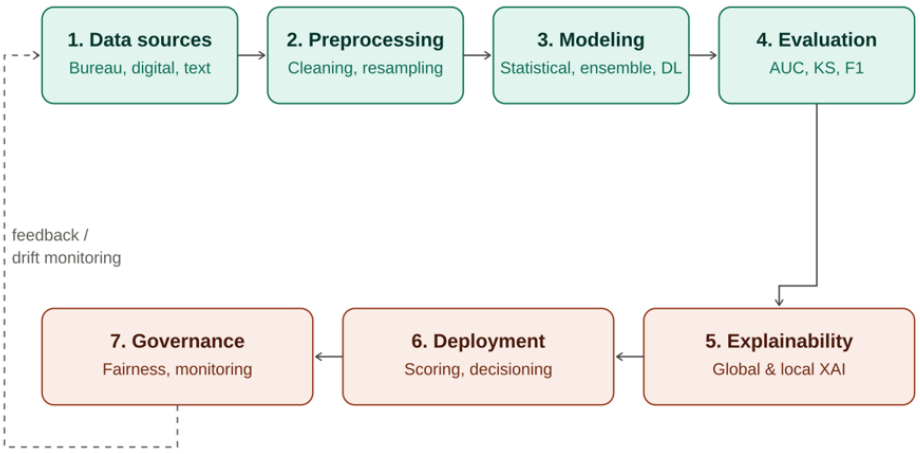

Figure 2. End-to-end machine learning workflow for credit risk assessment.

The principal technical and governance challenges encountered across this workflow are summarized in Table 4. Class imbalance remains one of the most frequently reported technical obstacles, with resampling techniques and cost-sensitive learning commonly adopted to improve minority-class prediction [9, 10]. As shown in Table 4, limited interpretability of

ensemble and deep learning models has been addressed through explainable AI approaches, including global methods for governance and model auditing, and local methods such as SHAP, LIME, and saliency techniques for explaining individual credit decisions [8, 11]. The literature has also highlighted growing concerns regarding algorithmic fairness, data privacy, and benchmarking inconsistency. Fairness-aware auditing, privacy-preserving learning architectures, and standardized benchmarking protocols have therefore emerged as complementary strategies to enhance transparency, regulatory compliance, and the robustness of machine learning-based credit risk assessment in FinTech environments [2, 6, 13, 32]. Collectively, Figure 2 and Table 4 demonstrate that successful deployment of machine learning models depends not only on predictive accuracy but also on the integration of explainability, fairness, governance, and continuous model management throughout the model lifecycle.

Table 4. Challenges and mitigation strategies for ML-based credit risk assessment.

Challenge	Description	Mitigation Strategy	Ref.
Class imbalance	Default cases are much fewer than non-default cases, leading to biased model training.	Resampling techniques (SMOTE, over-/under-sampling) and cost-sensitive learning.	[9, 10]
Limited interpretability	Ensemble and deep learning models provide limited explanation of individual predictions.	Global interpretability (feature importance, surrogate models) and local XAI methods (SHAP, LIME, saliency maps).	[8, 11]
Algorithmic fairness	Alternative data may act as proxies for protected attributes, resulting in biased decisions.	Fairness auditing using group fairness metrics and bias-mitigation techniques.	[13]
Data privacy	Alternative and cross-platform data raise privacy and regulatory compliance concerns.	Federated learning, privacy-preserving learning, data minimization, and regulatory-compliant governance.	[2, 32]
Benchmarking inconsistency	Differences in datasets, metrics, and validation protocols limit cross-study comparability.	Standardized benchmarking using multiple real-world datasets and common evaluation protocols.	[6]

6. Conclusion and Future Directions

This review has compared the principal families of machine learning models applied to credit risk assessment in FinTech applications, namely statistical classifiers, tree-based ensembles, and deep learning architectures, and has synthesized comparative evidence on their predictive performance across benchmark and FinTech-specific datasets identified through a structured literature search and screening process. The reviewed literature consistently indicates that ensemble methods, particularly gradient boosting and random forest, offer the strongest balance of accuracy, robustness, and computational efficiency for tabular credit data, while deep learning architectures provide additional gains on large, high-dimensional, or sequential alternative-data sources at the cost of reduced native interpretability. Traditional statistical classifiers, although generally outperformed on discriminatory metrics, remain relevant where regulatory transparency is paramount, and the relative ranking of model families shifts systematically with dataset scale, dimensionality, and class balance. Beyond model comparison, this review highlights that FinTech credit risk assessment is shaped as much by data and governance considerations as by algorithmic choice. The use of alternative data sources has demonstrated genuine potential to expand financial inclusion in markets with limited traditional credit-bureau coverage, including Pakistan and other South Asian economies, but this potential is contingent on addressing class imbalance, ensuring algorithmic fairness, and embedding explainability, distinguishing global model-governance explanations

from local borrower-facing explanations, directly into model design rather than treating it as an afterthought. Future research should prioritize the development of standardized, multi-dataset benchmarking protocols to enable reliable cross-study comparison; the design of hybrid architectures that combine the interpretability of statistical models with the predictive capacity of ensembles and deep networks; and the further exploration of federated and privacy-preserving learning to enable collaborative credit risk modelling across institutions without compromising borrower data privacy. As FinTech lending continues to expand into emerging markets, explainability-by-design and fairness-aware modelling will be essential to ensuring that gains in predictive accuracy translate into responsible and inclusive credit risk management.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this review article.

Author Contributions

Both authors contributed equally to the conceptualization, literature search, analysis, writing, and revision of this manuscript. Both authors have read and approved the final version of the manuscript.

Data Availability Statement

No new data were created or analyzed in this study. Data sharing is not applicable to this article, as this is a review article that synthesizes previously published literature. All sources discussed are cited in the reference list.

References

[1] Crook, J.N.; Edelman, D.B.; Thomas, L.C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*, 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100.

[2] Onay, C.; Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*, *26*, 382–405. https://doi.org/10.1108/JFRC-06-2017-0102.

[3] Berg, T.; Burg, V.; Gombović, A.; Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies*, *33*, 2845–2897. https://doi.org/10.1093/rfs/hhz099.

[4] Ala'raj, M.; Abbod, M.F.; Majdalawieh, M.; Jum'a, L. (2022). A deep learning model for behavioural credit scoring in banks. *Neural Computing and Applications*, *34*, 5839–5866. https://doi.org/10.1007/s00521-021-06695-z.

[5] Roy, J.K.; Vasa, L. (2024). Machine learning and artificial intelligence method for FinTech credit scoring and risk management: A systematic literature review. *International Journal of Business Analytics*, *11*, 1–23. https://doi.org/10.4018/IJBAN.347504.

[6] Lessmann, S.; Baesens, B.; Seow, H.-V.; Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*, 124–136. https://doi.org/10.1016/j.ejor.2015.05.030.

[7] Bastos, J.A. (2022). Predicting credit scores with boosted decision trees. *Forecasting*, *4*, 925–935. https://doi.org/10.3390/forecast4040050.

[8] Dastile, X.; Celik, T. (2021). Making deep learning-based predictions for credit scoring explainable. *IEEE Access*, *9*, 50426–50440. https://doi.org/10.1109/ACCESS.2021.3068854.

[9] Costello, F.J.; Lee, K.C. (2019). Exploring the performance of synthetic minority over-sampling technique (SMOTE) to predict good borrowers in P2P lending. *Journal of Digital Convergence*, *17*, 71–78. https://doi.org/10.14400/JDC.2019.17.9.071.

[10] Bahnsen, A.C.; Aouada, D.; Ottersten, B. (2014). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *41*, 6609–6619. https://doi.org/10.1016/j.eswa.2014.04.001.

[11] Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*, 203–216. https://doi.org/10.1007/s10614-020-10042-0.

[12] Shi, S.; Tse, R.; Luo, W.; D'Addona, S.; Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications*, *34*, 14327–14339. https://doi.org/10.1007/s00521-022-07472-2.

[13] Tian, X.; Tian, Z.; Khatib, S.F.A.; Wang, Y. (2024). Machine learning in internet financial risk management: A systematic literature review. *PLOS ONE*, *19*, e0300195. https://doi.org/10.1371/journal.pone.0300195.

[14] Dushimimana, B.; Wambui, Y.; Lubega, T.; McSharry, P. (2020). Use of machine learning techniques to create a credit score model for airtime loans. *Journal of Risk and Financial Management*, *13*, 180. https://doi.org/10.3390/jrfm13080180.

[15] Hand, D.J.; Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A*, *160*, 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x.

[16] Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three-stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, *222*, 168–178. https://doi.org/10.1016/j.ejor.2012.04.009.

[17] Cho, P.; Chang, W.; Song, J.W. (2019). Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision. *IEEE Access*, *7*, 16925–16939. https://doi.org/10.1109/ACCESS.2019.2896474.

[18] Xu, J.; Lu, Z.; Xie, Y. (2021). Loan default prediction of Chinese P2P market: A machine learning methodology. *Scientific Reports*, *11*, 18759. https://doi.org/10.1038/s41598-021-98361-6.

[19] Turiel, J.D.; Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, *7*, 191649. https://doi.org/10.1098/rsos.191649.

[20] Setiawan, N.; Suharjito, D. (2019). A comparison of prediction methods for credit default on P2P lending using machine learning. *Procedia Computer Science*, *157*, 38–45. https://doi.org/10.1016/j.procs.2019.08.139.

[21] Jiang, C.; Xiong, W.; Xu, Q.; Liu, Y. (2021). Predicting default of listed companies in mainland China via U-MIDAS Logit model with group lasso penalty. *Finance Research Letters*, *38*, 101487. https://doi.org/10.1016/j.frl.2020.101487.

[22] Jiang, J.; Meng, X.; Liu, Y.; Wang, H. (2022). An enhanced TSA-MLP model for identifying credit default problems. *SAGE Open*, *12*. https://doi.org/10.1177/21582440221094586.

[23] Abiodun, O.I.; Kiru, M.U.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Umar, A.M.; Linus, O.U.; Arshad, H.; Kazaure, A.L.; Gana, U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access*, *7*, 158820–158846. https://doi.org/10.1109/ACCESS.2019.2945545.

[24] Liang, L.; Cai, X. (2020). Forecasting peer-to-peer platform default rate with LSTM neural network. *Electronic Commerce Research and Applications*, *43*, 100997. https://doi.org/10.1016/j.elerap.2020.100997.

[25] Noriega, J.P.; Rivera, L.A.; Herrera, J.A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Data*, *8*, 169. https://doi.org/10.3390/data8110169.

[26] Serrano-Cinca, C.; Gutiérrez-Nieto, B.; López-Palacios, L. (2016). Determinants of default in P2P lending. *PLOS ONE*, *11*, e0164967. https://doi.org/10.1371/journal.pone.0164967.

[27] Ariza-Garzón, M.J.; Iturriaga, F.J.L.; Arroyo, J.; Caparrini, A. (2021). Risk-return modelling in the p2p lending market: Trends, gaps, recommendations and future directions. *Electronic Commerce Research and Applications*, *49*, 101079. https://doi.org/10.1016/j.elerap.2021.101079.

[28] Sun, J.; Li, H.; Huang, Q.-H.; He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, *57*, 41–56. https://doi.org/10.1016/j.knosys.2013.12.006.

[29] Hughes, J.P.; Jagtiani, J.; Moon, C.G. (2022). Consumer lending efficiency: Commercial banks versus a FinTech lender. *Financial Innovation*, *8*, 1–39. https://doi.org/10.1186/s40854-021-00326-1.

[30] Khalid, S.; Khan, M.A.; Mazliham, M.S.; Alam, M.M.; Aman, N.; Taj, M.T.; Zaka, R.; Jehangir, M. (2022). Predicting risk through artificial intelligence based on machine learning algorithms: A case of Pakistani nonfinancial firms. *Complexity*, *2022*, 6858916. https://doi.org/10.1155/2022/6858916.

[31] Babaei, G.; Giudici, P.; Raffinetti, E. (2023). Explainable FinTech lending. *Journal of Economics and Business*, *125–126*, 106126. https://doi.org/10.1016/j.jeconbus.2023.106126.

[32] Jovanovic, Z.; Hou, Z.; Biswas, K.; Muthukkumarasamy, V. (2024). Robust integration of blockchain and explainable federated learning for automated credit scoring. *Computer Networks*, *243*, 110303. https://doi.org/10.1016/j.comnet.2024.110303.

[33] Jiang, C.; Wang, Z.; Wang, R.; Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, *266*, 511–529. https://doi.org/10.1007/s10479-017-2668-z.