

Systematic Review

Credit Risk Scoring in the Age of AI: A Systematic Comparison of Traditional, ML, and DL Models Based on Accuracy, Stability, and Interpretability

Radouane Aboulmaouda ^{1,*} , Khadija Slimani ^{1,2}  and Nour El Houda Chaoui ¹ 

¹ LaRI Laboratory, Faculty of Science, Ibn Tofail University (UIT), BP 242, Kenitra 14000, Morocco; khadija.slimani@esiea.fr (K.S.); nourelhouda.chaoui@uit.ac.ma (N.E.H.C.)

² EsieaLab LDR, Higher School of Computer Science, Electronics and Automation (ESIEA), 75005 Paris, France

* Correspondence: rad.aboulmaouda@gmail.com or aboulmaouda.radouane@uit.ac.ma

Abstract

With the broad application of machine learning (ML) and deep learning (DL) models in financial markets, it has become increasingly important to evaluate their performance compared with traditional methods, especially for credit risk scoring in financial services. This mechanism determines the probability of default (PD) for borrowers, in other words, how likely a client is to fail to repay their debt. With the rapid development and growing availability of ML/DL technologies, it is essential for banks, financial institutions, auditors and regulatory bodies to assess whether these approaches truly outperform traditional models in terms of accuracy, stability and interpretability or whether their complexity comes at a cost. A systematic review following PRISMA examined credit risk scoring models. From 520 initial articles, 117 were analyzed to compare ML/DL approaches with traditional methods. This review evaluates accuracy, stability and interpretability, offering guidance for model selection in real-world credit scoring. Logistic regression remains essential in regulated contexts requiring transparency, supporting informed decisions on balancing performance and explainability.

Keywords: credit scoring; accuracy; stability; interpretability; traditional models; machine learning; deep learning; PRISMA; review

1. Introduction

Credit risk evaluation is fundamental in the decision-making process of banks and other financial institutions. Credit scoring is a statistical or machine learning technique, typically represented using a scorecard, which helps predict the probability of borrowers turning delinquent and hence defaulting (Altman, 1968; Siddiqi, 2012). This is a forward-looking tool that allows lenders to make quick decisions by using historical data to get a hint about future trends. Reliable models that can score credits are critical not only for assessing the risk of default probability of borrowers but also for pricing interest rates, controlling the credit portfolio's risk and managing regulatory requirements (Xia et al., 2018). Beyond these fundamental uses, credit scoring is adopted for automating loan approval processing, improving customer segmentation and optimizing collection procedures to increase operational efficiency and profitability. Statistical models, such as logistic regression and decision trees, have been commonly used in the domain of credit scoring. Such models provide simplicity and transparency, which are essential for compliance, financial regulation and building trust among stakeholders (Lessmann et al., 2015; Thomas et al., 2002). Nonetheless,



Academic Editors: Herbert Kimura
and Leonardo Fernando Cruz Basso

Received: 15 June 2026

Revised: 10 July 2026

Accepted: 11 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

with the rise in big data and machine learning, more advanced algorithms are being used to improve the accuracy of prediction while dealing with a variety of imbalanced datasets.

The development of credit scoring is a great step forward in financial tools and analysis. Prior to the 20th century, credit scoring was based primarily on subjective information (i.e., lenders' intuition) and qualitative data (Altman, 1968). Statistical techniques started the formalization process in the 1950s; discriminant analysis models were developed at that time for assessing risk using financial ratios. The Z-Score model of Edward Altman (Altman, 1968) was the first step through the door to using a systematic approach based on historical financial data to measure the chances of bankruptcy (Altman, 1968; Wiginton, 1980).

The 1980s and 1990s saw the digitization of financial systems that significantly changed credit scoring with automated credit-scoring models like the FICO score, which is now regarded as a standard in consumer credit rating in the United States (Crook et al., 2007). These systems normalized and sped the credit determining process to enable a wider availability of credit. In the same period, more sophisticated statistical techniques were incorporated into research ensembles to improve their prediction (Lessmann et al., 2015).

The advent of big data analytics in the early 2010s also ushered in a new era with ongoing work on machine learning (ML) techniques for credit scoring. ML algorithms, such as support vector machines (SVM), neural networks and ensemble methods, have better prediction than traditional models by using large and complex datasets including alternative data sources such as transactional behavior or digital footprints (Huang et al., 2007; Baesens et al., 2003). In 2018, developments such as the method of heterogeneous ensemble combining bagging and stacking improved the accuracy of model predictions when handling problems associated with similar data variety and imbalance (Xia et al., 2018).

With the advent of big data, machine learning and AI (including generative AI) in recent years since 2020 also present major challenges for practitioners to determine which is the most suitable scoring model. Although the traditional final-chain model, such as logistic regression, is applied for simplicity, interpretability and regulatory acceptance reasons, it often shows lower predicting power than modern ML methods (Lessmann et al., 2015). Practitioners are more interested in making use of sophisticated algorithms, which purport to be more accurate and robust. However, these more sophisticated models inherently suffer from interpretability and stability issues, which are often barriers to be accepted in industrial sectors dominated by transparency and explainability requirements (Rudin, 2019; Montevechi et al., 2024). Although ML and DL models offer tremendous possibilities for exploitation of unstructured data, they bring concerns about "black-box" behavior and compliance (Tjoa & Guan, 2021). Reconciling these trade-offs between accuracy, interpretability and stability has been identified as a core concern because practitioners need to negotiate the benefits of innovating while establishing trust and complying with regulatory institutions (Klein & Walther, 2024).

This review intends to offer a holistic comparison of standards and machine learning-based credit-scoring models by systemically reviewing the scientific literature published from 2018 to 2026. The aim is to compare these models in terms of three fundamental factors: accuracy, stability and interpretability.

Accuracy refers to a model's ability to correctly classify the borrowers as good or bad credit risks and is often associated with metrics like area under the curve (AUC) and precision–recall performance (Lessmann et al., 2015; James, 2021). Stability refers to the robustness of a model with respect to different datasets and time periods, which is fundamental in a non-stable economic environment where economic shocks or borrowers' behavior can alter the reliability of models (Kwon et al., 2025; Montevechi et al., 2024). Interpretability refers to whether stakeholders, like bank managers and regulators, can understand and

trust the model's decision-making, which is essential in meeting regulatory requirements as well as the confidence of the stakeholder (James, 2021; Klein & Walther, 2024).

Adopting the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology for systematic reviews (Page et al., 2021a), relevant studies were identified that provide insights into how different credit-scoring models perform across these three dimensions. By focusing on recent advancements and emerging trends, this review seeks to guide researchers and practitioners in choosing the most appropriate models for their specific needs. It also intends to highlight the limitations in existing research and suggest avenues for future work.

This article is structured as follows: the first section describes the method employed to conduct the systematic review, including the selection criteria and the main items of analysis. The following section reports the results, organized around the three dimensions: accuracy, stability, and interpretability. Finally, this article discusses the implications of these findings for practitioners and researchers and outlines potential avenues for future research on credit risk assessment.

2. Methodology

For the review methodology, PRISMA guidelines address transparency and reporting of systematic reviews to promote replicable practices in the selection and synthesis of literature (Page et al., 2021a). The aim of this review is to contrast traditional and ML models under three different dimensions, namely accuracy, stability and interpretability, for credit risk scoring. Publications between 2018 and 2026 were screened. The process contains five essential stages: literature search, study selection, data extraction, evaluation and findings.

Scope of the review window. The systematic search is deliberately bounded to 2018–2026 to capture the period in which ML/DL credit-scoring methods and their explainability and stability tooling matured. Foundational and landmark contributions that predate this window, discriminant analysis (Altman, 1968), the FICO era of automated scoring, and the benchmark studies of Baesens et al. (2003) and Lessmann et al. (2015) are not part of the screened corpus but are integrated narratively in the Introduction to preserve the evolutionary context of scoring models; extending the systematic corpus backwards is noted as a complementary avenue (Section 5).

In accordance with PRISMA recommendations to ensure methodological transparency and prevent unnecessary duplication, this systematic review has been registered in PROSPERO, the international registry for systematic review protocols (registration number CRD420261426881; <https://www.crd.york.ac.uk/PROSPERO/view/CRD420261426881>, accessed on 13 July 2026).

2.1. Literature Search

A thorough search was conducted in several scientific databases, such as ScienceDirect and Google scholar, to obtain a coverage of relevant studies. We included studies published between 2018 and 2026 that focus on credit scoring, credit risk assessment, or credit evaluation. Eligible papers must use at least one standard or machine learning model commonly applied in credit scoring, such as logistic regression, linear discriminant analysis, generalized linear models, Naive Bayes, decision trees (including CART), Bayesian networks, gradient boosting methods (XGBoost, LightGBM, CatBoost, AdaBoost), random forests, rotation forests, k-nearest neighbors, support vector machines, or neural and deep learning architectures (CNN, MLP, ResNet, Inception, etc.). Studies were retained only if they reported model performance metrics, such as accuracy, stability, interpretability,

explainability, AUC, or overall predictive performance. The full list of sources, the exact search string and the screening criteria are detailed below to ensure reproducibility.

Search sources and query. To ensure reproducibility, the search drew on four complementary sources: ScienceDirect, Google Scholar, Semantic Scholar and OpenAlex aggregated through the Elicit research assistant. The following Boolean query was executed over the 2018–2026 window: (“credit scoring” OR “credit risk” OR “credit assessment” OR “credit evaluation” OR “scorecard”) AND year: [2018 TO 2026] AND (“logistic regression” OR “linear discriminant analysis” OR “generalized linear model” OR “naive bayes” OR “decision tree” OR CART OR “bayesian network” OR “gradient boosting” OR XGBoost OR “light gradient boosting” OR LGBM OR “categorical boosting” OR CatBoost OR AdaBoost OR “random forest” OR “rotation forest” OR “k-nearest neighbors” OR “support vector machine” OR SVM OR “neural network” OR “deep learning” OR “convolutional neural network” OR CNN OR “multi-layer perceptron” OR MLP OR ResNet OR “inception network”) AND (accuracy OR stability OR interpretability OR explainability OR AUC OR “model performance”). The query returned 520 records, which were exported for screening.

Screening criteria. Records were screened on title and abstract against eight explicit, jointly applied criteria:

1. Credit-domain relevance (credit scoring, credit risk assessment or loan-default prediction);
2. Implementation of at least one of 28 predefined model types (LR, LDA, GLM, NB, DT, CART, BNet, GBM, SGB, XGB, LGBM, CatB, AdaB, RaF, RoF, ARF, LeBa, K-NN, SVM, NNs, DNNs, CNNs, ANN, MLP, FCN, ResNet, Inc);
3. Reporting of at least one performance metric (AUC, Gini, KS, F1, accuracy, recall or precision);
4. Empirical validation on real-world or simulated credit data;
5. Sufficient methodological detail for data extraction;
6. A genuine credit-scoring application. Each record was assessed holistically against all eight criteria.

2.2. Study Selection

The method of selection was clearly designed not to overlook pertinent and good studies. Out of 520 initially identified articles, 76 were removed as duplicates, and 444 entries remained. These articles were subjected to an initial screening, resulting in the exclusion of 145 studies for being unrelated to credit scoring or falling outside the financial domain. This reduced the total number of articles to 299 for review. The titles and abstracts of these remaining articles were then reviewed, leading to the exclusion of an additional 102 studies that did not meet the inclusion criteria. The full-text articles of the 197 that survived after application of the exclusion criteria were then reviewed according to the following criteria: no articles that only discussed technical mathematical algorithms of machine learning without applying them to credit scoring or those lacking clear results or were not published in English. Following this process, 117 articles that met all inclusion criteria were selected for the final analysis. Based on the template proposed by Page et al. (2021b), the results are presented in Figure 1.

2.3. Study Type Distribution

Out of the 117 studies retained, 82 (70.1%) are comparative analyses testing multiple models simultaneously, enabling direct cross-model conclusions; 25 (21.4%) evaluate a single model against a benchmark; and 10 (8.5%) are systematic reviews or survey type contributions. The dominance of comparative studies (82) is the principal strength of this corpus for drawing cross-model evidence.

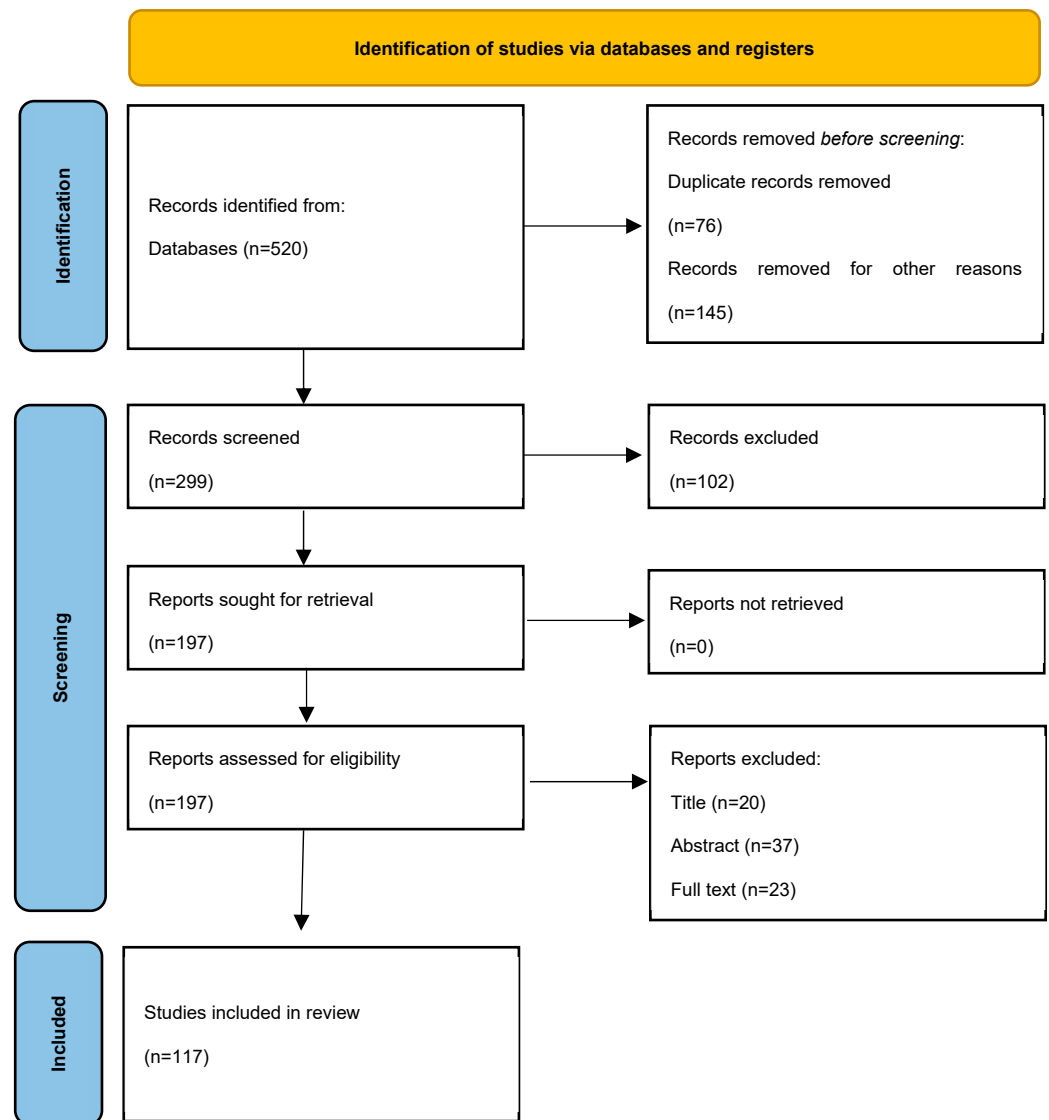


Figure 1. Enhanced PRISMA workflow diagram.

Quality and risk-of-bias considerations. A first level of quality control was embedded in the eligibility criteria: only peer-reviewed studies and those providing sufficient methodological detail (dataset characteristics, model specification and validation procedure) were retained. We nonetheless acknowledge that a formal item-level risk-of-bias appraisal, for instance using PROBAST, the tool designed for prediction-model studies, was not applied to each included study, so the studies are weighted equally in the synthesis. This is stated as a limitation, and a structured PROBAST-based appraisal is identified as a priority for the extended version of this work (Section 5).

2.4. Models Considered in the Review

Table 1 outlines the models covered in this review, categorized into standard models, machine learning models and deep learning models.

Table 1. Overview of models considered for credit-scoring analysis.

Category	Model Family	Specific Models
Standard models	Logistic Regression	Logistic Regression (LR) Penalized Logistic Regression (PLR) Penalized Logistic Tree Regression (PLTR)
	Linear Models	Linear Discriminant Analysis (LDA)
	Decision Trees	Decision Tree (DT) Classification and Regression Trees (CART) C4.5 Decision Tree
	Naive Bayes	Naive Bayes (NB) Gaussian Naive Bayes (GNB)
	K-Nearest Neighbors	K-Nearest Neighbors (K-NN)
	Bayesian Networks	Bayesian Network (BNet)
Machine Learning models	Gradient Boosting	Extreme Gradient Boosting (XGBoost) Light Gradient Boosting Machine (LightGBM) Categorical Boosting (CatBoost) Gradient Boosting Machine (GBM) Adaptive Boosting (AdaBoost)
	Random Forest	Random Forest (RF) Rotation Forest (RoF) Adaptive Random Forest (ARF)
	Support Vector Machine	Support Vector Machine (SVM)
	Hybrid/Ensemble	Stacking Ensemble Models (SEM) Two-Stage Hybrid Models (TSH) TabNet Factorization Machine (FM)
Deep Learning models	Neural Networks	Multi-layer Perceptron (MLP) Deep Neural Network (DNN) Convolutional Neural Network (CNN) Long Short-Term Memory Network (LSTM) Artificial Neural Network (ANN) Bidirectional LSTM (BiLSTM) Graph Conv. Network (GCN)

3. Review Findings

3.1. Annual Distribution of the Publications

The 117 retained studies span 2019 to 2026, with a pronounced acceleration from 2022. The two years 2024 and 2025 alone account for 48 studies (41% of the corpus), reflecting the convergence of three trends: the widespread availability of Kaggle-hosted credit datasets, the maturation of explainability frameworks (SHAP, LIME), and growing regulatory pressure for auditable AI in lending. The 2026 cohort (eight studies already indexed as of the search date) confirms sustained growth. Importantly, the shift toward comparative multi-model studies (82 of 117) accelerates from 2021 onward, enabling the cross-tier evidence synthesis central to this review. An overview of these trends is shown in Figure 2.

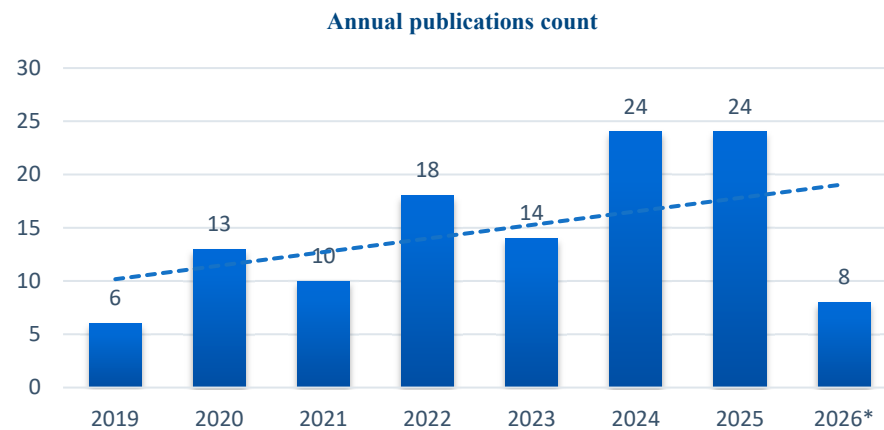


Figure 2. Annual distribution of publications retained in the review. * Data as of April 2026. The count is provisional and may increase as the year progresses.

3.2. Distribution Across Model Families

Figure 3 shows the distribution of studies across model families. Logistic regression leads with 60 studies, reflecting its enduring relevance in regulated environments where interpretability is paramount. ML ensemble methods follow closely: Gradient boosting (58), random forest (50) and support vector machines (48), driven by their strong predictive performance on financial data. Neural networks (45) represent the deep learning strand, though their adoption remains constrained by interpretability and regulatory concerns.

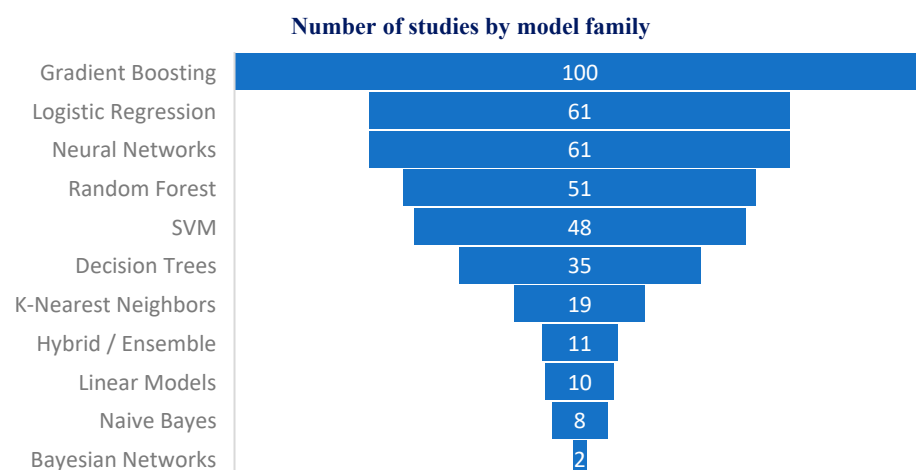


Figure 3. Frequency of evaluation across scoring model families.

Overall, the distribution confirms a clear tension between accuracy-driven ML/DL approaches and the transparency demands that keep traditional models firmly in the race.

3.3. Dataset Characteristics

Dataset transparency is a significant methodological weakness: 75 of 117 studies (64.1%) do not clearly document the dataset name or source. Among named datasets, the German Credit UCI (22 studies) and Lending Club (14 studies) dominate. The geographic concentration on German and US datasets limits generalizability to other credit markets. The Lending Club dataset emerged as the most frequently used data source. Dataset sizes ranged considerably from 1000 entries to over 2.8 million records, as detailed in Table 2.

Table 2. Key characteristics of main datasets employed across the reviewed studies.

Dataset Source	Dataset Name	Size	Region	Type	Studies
UCI machine learning repository (22 studies)	German credit dataset	1000 entries	Germany	Public	(Emmanuel et al., 2024; Zong, 2025; Seo, 2020; Arora & Kaur, 2020; H. Li et al., 2024; Quan & Sun, 2024; Mathibela & Maposa, 2026; Bahloul et al., 2026; Paz et al., 2025; Atif & Salmi, 2022; Manik et al., 2022; X. Zhang et al., 2024b; Bahadori & Hassanzadeh, 2026; Aruleba & Sun, 2025; Gerunov, 2020; Kuyoro et al., 2022; Chi et al., 2021; Zhou et al., 2020)
	Australian credit datasets	690 entries	Australia	Public	(Emmanuel et al., 2024; Yu et al., 2025; Pertiwi et al., 2024a; Quan & Sun, 2024; Paz et al., 2025; Aruleba & Sun, 2025; Chi et al., 2021)
	Taiwan credit datasets	30,000 entries	Taiwan	Public	(Emmanuel et al., 2024; Sugiyarto et al., 2019; Almajid, 2022; Paz et al., 2025; Ugarte Chacon et al., 2023; Gerunov, 2020)
Lending Club (14 studies)	Lending Club	25,535 entries (time period 2007–2018)	US	Public	(Nwafor et al., 2024; Liu et al., 2022b; Yu et al., 2025; Pertiwi et al., 2024b; Kim & Cho, 2019; Munsarif et al., 2022; Berhane et al., 2024; Paz et al., 2025; Bahadori & Hassanzadeh, 2026)
	P2P Lending Club (LC)	2,876,629 loans and 150 features	USA	Public, real-world	(Trinh, 2024; Mendrofa et al., 2023; Arora & Kaur, 2020; Berhane et al., 2024)
	HELOC Dataset, Lending Club (LC) Dataset	HELOC—10,000 instances, LC—2.3 million loan applications	Likely USA	Real-world	(Demajo et al., 2020)
Kaggle Platform (7 studies)	Home Credit dataset	307,511 rows and 122 columns	-	Real-world	(Hernes et al., 2023)
	Loan Default Prediction/vehicle-loan-default-prediction	255,347 entries	-	Real-world, public	(Ukpabi et al., 2025; Rao et al., 2023; B. Han, 2024)
	Credit risk dataset	111,000 entries	-	Public, real-world	(Kanaparthi, 2024; Gür et al., 2025; X. Han et al., 2025)
	Credit Score Classification Clean Data Give Me Some Credit project	31,695 entries 150,000 entries	-	Public, real-world	
CSMAR database (2 studies)	GEM, STAR, SMB datasets	27,032,572 samples	China	Public, real-world	(Du & Zhang, 2023; C. Shen & Wu, 2025)
	CSMAR database	32,246 data points from 3762 enterprises			

This concentration on small, mostly European and US public benchmarks, with 64.1% of studies not disclosing their data source, is a material threat to external validity: conclu-

sions about model superiority may not transfer to emerging-market portfolios or to real domestic-bank credit samples, which remain markedly under-represented in the corpus. Findings should therefore be read as conditional on this dataset profile.

The empirical evidence of the reviewed corpus rests on a small number of recurring public benchmarks, complemented by larger or proprietary sources that better reflect operational credit-scoring conditions. The UCI Machine Learning Repository alone supports 22 studies through three reference datasets: the German credit dataset (1000 entries, Germany) is the most widely used, with 17 studies relying on it (Emmanuel et al., 2024; Zong, 2025; Seo, 2020; Arora & Kaur, 2020; H. Li et al., 2024; Quan & Sun, 2024; Mathibela & Maposa, 2026; Bahloul et al., 2026; Paz et al., 2025; Atif & Salmi, 2022; Manik et al., 2022; X. Zhang et al., 2024b; Bahadori & Hassanzadeh, 2026; Aruleba & Sun, 2025; Gerunov, 2020; Kuyoro et al., 2022; Chi et al., 2021; Zhou et al., 2020); the Australian credit dataset (690 entries) is mobilized in seven studies (Emmanuel et al., 2024; Yu et al., 2025; Pertiwi et al., 2024a; Quan & Sun, 2024; Paz et al., 2025; Aruleba & Sun, 2025; Chi et al., 2021); and the Taiwan credit dataset (30,000 entries) supports six studies (Emmanuel et al., 2024; Sugiyarto et al., 2019; Almajid, 2022; Paz et al., 2025; Ugarte Chacon et al., 2023; Gerunov, 2020). Lending Club data drives 14 additional studies across three configurations: the standard Lending Club extract of 25,535 entries covering 2007–2018 (Nwafor et al., 2024; Liu et al., 2022b; Yu et al., 2025; Pertiwi et al., 2024b; Kim & Cho, 2019; Munsarif et al., 2022; Berhane et al., 2024; Paz et al., 2025; Bahadori & Hassanzadeh, 2026), the much larger P2P Lending Club sample of 2,876,629 loans and 150 features (Trinh, 2024; Mendrofa et al., 2023; Arora & Kaur, 2020; Berhane et al., 2024), and a joint HELOC (10,000 instances) and Lending Club (2.3 million loan applications) configuration (Demajo et al., 2020). The Kaggle platform supplies a further set of seven studies, including the Home Credit dataset of 307,511 rows and 122 columns (Hernes et al., 2023), the Loan Default Prediction/vehicle-loan-default extracts of 255,347 entries (Ukpabi et al., 2025; B. Han, 2024; Rao et al., 2023), and the Credit Risk, Credit Score Classification Clean Data and Give Me Some Credit samples (111,000, 31,695 and 150,000 entries respectively) (Gür et al., 2025; Kanaparthi, 2024; X. Han et al., 2025). The CSMAR database extends the geographic scope to China, with two complementary configurations: 27,032,572 GEM/STAR/SMB samples (Du & Zhang, 2023) and 32,246 data points from 3762 enterprises (C. Shen & Wu, 2025).

Beyond these public benchmarks, several studies leverage private real-world credit registries that strengthen external validity. Such evidence is drawn from the Central Bank of the Republic of North Macedonia credit registry (Doko et al., 2021), a Brazilian retail-banking sample combining 147,000 and 50,000 customers from two institutions over a 12-month window (Bispo & Tessmann, 2025), a credit card dataset of 625 records from UAE banks (Jumaa et al., 2023), 506 SMEs listed on the Shenzhen Stock Exchange between 2016 and 2020 (C. Zhang & Zhou, 2023), Chinese supply-chain credit data on listed enterprises (Yao et al., 2024), 246 family farms in Inner Mongolia (Sun et al., 2025), 610 borrowers from an Indonesian bank between 2015 and 2018 (Putri et al., 2021), and a P2P platform with 136 features reduced to 34 (Mendrofa et al., 2023). The Pramesti et al. (2022) sample of 5199 borrowers from an Indonesian bank, with a 98%/2% imbalance ratio, illustrates the severe class-skew that is typical of credit scoring and motivates the systematic use of SMOTE-family resampling (Pramesti et al., 2022; Maitanmi et al., 2024; Kandi & García-Dopico, 2025; F. Shen et al., 2021; Liu et al., 2022a).

3.4. Accuracy Metrics Analysis

Table 3 reports performance metric ranges across the three model tiers and ten families. Four metrics are presented: accuracy (proportion of correctly classified entries), AUC (area under the ROC curve, measuring the model's ability to discriminate between defaulters

and non-defaulters across all classification thresholds, ranging from 0.5, random to 1.0, perfect), F1-score (harmonic mean of precision and recall, particularly informative under class imbalance), and recall (proportion of actual defaulters correctly identified, critical in credit risk where missed defaults carry high financial cost). Values reflect the observed min–max range across all studies within each family.

Table 3. Reported performance metric ranges by category and model family.

Category	Model Family	N° Studies	AUC	Accuracy	F1-Score	Recall	Precision
Traditional models	Logistic Regression	60	0.5–0.99	29–92%	77–88%	76–97%	15–79%
	Linear Models (LDA)	11	0.73–0.85	81%	-	97%	14–25%
	Decision Trees	33	0.74–0.82	72–96%	73–76%	0–75%	75–82%
	Naive Bayes	8	0.56	79%	-	97%	15%
	K-Nearest Neighbors	18	0.54–0.98	69–77%	-	95%	14–100%
Machine Learning models	Gradient Boosting	58	0.79–0.99	74–97%	58–100%	0–90%	17–99%
	Random Forest	50	0.75–0.92	71–90%	74–86%	74–95%	36–92%
	Support Vector Machine	48	0.61–0.86	80–91%	42–90%	41–97%	26–82%
	Hybrid/Ensemble	8	0.91–0.93	86–97%	79–85%	82%	78–99%
Deep Learning models	Neural Networks (all)	45	0.5–0.98	67–99%	74–85%	74–100%	0–99%

Three tiers emerge with distinct accuracy profiles:

Traditional Models achieve AUC values in the 0.65–0.99 range. Logistic regression, the most tested model (60 studies), systematically achieves lower AUC than ML alternatives but remains competitive in studies using small, well structured datasets. Decision trees (0.74–0.82 AUC) show higher variability depending on depth and pruning. Naive Bayes is reported with 79% accuracy and an AUC = 0.56 in comparable settings (Trinh, 2024), while LDA reaches 81% accuracy with an AUC up to 0.85 and recall near 0.97 (Gür et al., 2025). The logistic regression spread is also configuration-driven: an under-tuned baseline can fall to an AUC = 0.58, with recall of only 22% (Oyeyemi et al., 2025), yet a well-prepared LR reaches an AUC = 0.814 with appropriate preprocessing (Gao et al., 2021) and an AUC = 0.80, with recall of 0.98 and F1 = 0.845, in a balanced setup (Atif & Salmi, 2022).

Machine learning models dominate the accuracy dimension. Gradient boosting methods achieve an AUC up to 0.997 (Wawge, 2026), with XGBoost consistently outperforming in comparative studies. Random forest (0.75–0.92) and SVM (0.61–0.86) are strong secondary choices. Within gradient boosting, LightGBM (AUC = 0.959 (Gao et al., 2021), AUC = 0.937 (Peng, 2022)) and CatBoost (AUC = 0.851 (Z. Wang et al., 2025)) consistently lift performance above the XGBoost baseline (AUC = 0.951 (Gao et al., 2021), 0.928 (Peng, 2022), 0.95 (Wu, 2025)). Hybrid and ensemble configurations, such as the Soft Voting Fusion of LR, SVM, RF, XGBoost and LightGBM (AUC = 0.984 (Gao et al., 2021)); the 1DCNN-XGBoost hybrid reaching 96% accuracy and an AUC = 0.98 (Nwafor et al., 2024); and the stacked classifier evaluated jointly on the Australian, German and Taiwan datasets (AUC = 0.934, 0.944 and 0.870 respectively (Emmanuel et al., 2024)) push accuracy further. Tree-based ensembles also stand out, with the DTE-DSA model reaching an AUC = 0.9016 and a KS statistic of 0.7369 (Yao et al., 2022b), and C4.5 achieving 91.90% accuracy and an AUC = 0.915 (Amrin et al., 2025). Random Forest, when used standalone, reaches an out-of-sample $R^2 = 89.75\%$ (Fauser & Gruener, 2020).

Deep learning models display the widest performance range (AUC 0.5–0.91). Architecture choice is the primary determinant: simple MLP architectures underperform ML baselines in several studies, while advanced attention-based models (DeepCreditRisk: AUC 0.920, (X. Han et al., 2025)) and ensemble TabNet (AUC 0.941, (Z. Wang et al., 2025)) approach or exceed gradient boosting performance. The 0.43 lower bound (MLP under stress conditions (Osei et al., 2021)) underscores the instability risk of DL architectures. The sensitivity to architecture is also visible inside the TabNet family, where successive ensemble refinements lift AUC from 0.878 (vanilla TabNet) to 0.907 (single ensemble) and 0.941 (improved TabNet), with accuracy rising from 0.954 to 0.978 and F1 from 0.752 to 0.821 (Z. Wang et al., 2025). CNN-based architectures span a similarly wide range, from a baseline CNN at an AUC = 0.67, with 75.86% accuracy (Kim & Cho, 2019), to a tuned MLP reaching an AUC = 0.98, with 95% accuracy (Oyeyemi et al., 2025), and a CNN that rises from 67% to 89% accuracy when moving from back testing to stress testing, in mirror image to an MLP that collapses from 91% to 43% under the same shift (Osei et al., 2021). The DeepRisk approach extends this evidence by fusing demographic and behavioural data and reporting gains in precision, recall, F1, AUC and economic loss against the baselines (W. Zhang et al., 2022). Optimization-augmented neural networks (PCA + PSO) reach 82.67% accuracy with an AUC = 0.749 (Sugiyarto et al., 2019), and a multi-layer perceptron tuned on the Taiwan dataset reaches an AUC = 0.7184 (Almajid, 2022).

3.5. Top/Worst 10 Studies by Best/Lowest Reported AUC

Top performers: The upper end of the AUC distribution is dominated by gradient boosting, XGBoost in particular, with ensemble configurations and a well-tuned MLP also ranking among the best. The key lesson is that top-tier performance depends as much on feature selection, tuning and data preparation as on the choice of model itself and that the very highest reported scores should be interpreted with caution, as they may reflect dataset-specific or near-linearly separable data. Study-level figures are reported in Table 4.

Table 4. Top 10 studies by best reported AUC.

Study	Best AUC	Best Model	Dataset	Size
(Wawge, 2026)	0.997	XGBoost + RFE	Bank Credit Card dataset	Not mentioned
(Gao et al., 2021)	0.984	Soft Voting Fusion	Not mentioned	373,554 entries
(Oyeyemi et al., 2025)	0.98	MLP Neural Network	Not mentioned	Not mentioned
(Wu, 2025)	0.95	XGBoost	Not mentioned	Not mentioned
(S. Wang & Zhang, 2025)	0.941	Improved TabNet	Give Me Some Credit (Kaagle)	150,000 entries
(Peng, 2022)	0.937	LightGBM	Tianchi competition platform	800,000 entries (initial), 2200 (used)
(X. Han et al., 2025)	0.92	DeepCreditRisk	Kaggle Credit Risk Modeling	111,000 entries
(Amrin et al., 2025)	0.915	C4.5 Decision Tree	Not mentioned	481 vehicle credit records
(Yao et al., 2022b)	0.902	DTE-DSA Ensemble	Not mentioned	Not mentioned
(Vakrani et al., 2025)	0.89	XGBoost	Real-world dataset	1000 entries

Lowest performers: At the bottom of the spectrum, logistic regression and simple tree- or CNN-based baselines recur. The most revealing signal, however, is that the same model, notably XGBoost, appears at both extremes: this is not a model failure but confirmation that data preparation, tuning and evaluation conditions, rather than model family alone,

drive final performance. Robustness under adverse or stress conditions is likewise uneven across models. Individual studies are listed in Table 5.

Table 5. Bottom 10 studies by lowest reported AUC.

Study	Lowest AUC	Worst Model	Dataset	Size
(Oyeyemi et al., 2025)	0.5	Logistic Regression	Not mentioned	Not mentioned
(Amrin et al., 2025)	0.661	CHAID	Not mentioned	481 vehicle credit records
(Kim & Cho, 2019)	0.67	Baseline CNN	Lending Club	143,823 entries
(Mathibela & Maposa, 2026)	0.733	XGBoost (baseline)	German Credit Dataset (UCI)	Not mentioned
(Vakrani et al., 2025)	0.76	XGBoost (stress)	Real-world dataset	1000 entries
(Kanaparthi, 2024)	0.77	Random Forest	Credit risk dataset (Kaggle)	Not mentioned
(Atif & Salmi, 2022)	0.8	Logistic Regression	German credit dataset	Not mentioned
(T. Xu, 2024)	0.81	Random Forest	Not mentioned	Not mentioned
(Gao et al., 2021)	0.814	LR (baseline)	Data from 123 SMEs nationwide	373,554 entries
(Sudjianto & Burakov, 2025)	0.82	LR/XGBoost (low end)	Not mentioned	Not mentioned

3.6. Stability and Robustness Analysis

Stability of scoring models, in contrast, is how robustly a model behaves across datasets and over time. Nevertheless, the stability of prediction is not always that easy to consider as it dramatically depends on the quality and variance of the training data. Model stability is assessed through a variety of complementary approaches in the literature, as summarized in Table 6. Cross-validation (k-fold) and out-of-time validation are the two most widely used techniques: the former evaluates generalization by rotating training and test subsets, while the latter, particularly relevant in credit scoring, tests whether a model holds up on data from a later time period. Bootstrap/resampling extends this logic by repeatedly sampling the dataset to estimate performance variance. On the distributional side, the population stability index (PSI) detects shifts between the training population and the deployment population, and the information value (IV) measures the predictive power of individual variables. Concept drift and data drift monitoring go further by tracking evolving patterns over time. More formal comparisons rely on statistical tests (Wilcoxon, Friedman, *t*-test) to validate whether performance differences between models are significant. Finally, stress testing evaluates models under adverse or crisis scenarios, while multi-dataset robustness and SMOTE/class-balancing address generalizability across populations and class imbalance respectively. Standard errors and confidence intervals remain rarely used, and a basic train/test split, without any stability mechanism, was not reported as a standalone validation strategy in any reviewed study. Concrete applications of these methods are unevenly distributed across the corpus. Stratified or repeated k-fold validation is the most frequent stability evidence, used for instance with 10-fold designs (Naufal & Suseno, 2023; Manik et al., 2022; Y.-Z. Xu et al., 2019), 5-fold designs (Gür et al., 2025; Kim & Cho, 2019; Nur et al., 2019; H. Li et al., 2024; Oliveira & Basso, 2025; X. Li & Li, 2024), and stratified k-fold variants (C. Shen & Wu, 2025; Nguyen et al., 2025), with (Nwafor et al., 2024) additionally complementing cross-validation with the H-measure as a robustness metric incorporating misclassification costs. Out-of-time validation appears in studies that explicitly separate training and testing

periods, for example, training on 2017–2021 data and testing on 2022, where XGBoost maintains high performance, random forest declines slightly, logistic regression is stable but inferior, and SVM deteriorates significantly (Oliveira & Basso, 2025), as well as in the 2007–2012 vs. 2013–2019 comparison of (Trinh, 2024) and in (Z. Wang et al., 2025), where validation results confirm generalization capability on external datasets. Multi-dataset robustness and external validation are reported in (Yadava, 2023), where the accuracy drop under external validation is limited to 5% for a CNN against 12% for a random forest; in (Yu et al., 2025), where the TSC-SVM model maintains consistent performance across Lending Club, Australian and Credit Risk datasets; and in (X. Han et al., 2025), which documents stable performance across diverse credit datasets and time-series lengths. Statistical comparisons rely on the Friedman test combined with McNemar post hoc analysis (Mathibela & Maposa, 2026), on stratified K-fold cross-validation with the Wilcoxon signed-rank test (C. Shen & Wu, 2025), and on the Friedman test together with chi-square and Fisher’s exact tests (Gür et al., 2025). Stress testing and back testing are explicitly implemented in (Osei et al., 2021), which uses an 80%/20% imbalanced split on Home Credit historical data, and in (Vakrani et al., 2025), which reports AUC degradation of about −7% for XGBoost and −11% for logistic regression under macroeconomic shocks; the TSC-SVM model in (Yu et al., 2025) also reports strong robustness on imbalanced datasets. PSI and IV, the standard distributional stability indicators in credit scoring, are jointly reported only in (Sudjianto & Burakov, 2025), which additionally documents standard errors for both metrics and a broader uncertainty-quantification component. Other quantitative stability signals include a Jaccard stability measure (Arora & Kaur, 2020), an out-of-bag score of 0.905 for a random forest (Kuyoro et al., 2022), a 2.6% train/test accuracy gap interpreted as cross-validation variance (X. Zhang et al., 2024a), and the SAFE framework combined with the rank graduation box (Lorenz-Zonoids) as a multi-dimensional stability/sustainability evaluator (Ugarte Chacon et al., 2023). Continuous monitoring of performance over time is formalized in (X. Han et al., 2025) through validation gates, ensemble smoothing, an adaptive learning rate driven by market stability and a performance-stability tracking layer, while (Yadava, 2023) highlights the need for continuous monitoring to detect data drift in operational deployment. Conversely, several review-level papers explicitly flag the absence of PSI/IV in the studies they synthesize, treating stability mainly through a regulatory or qualitative lens (Bahlool et al., 2026; Paz et al., 2025; Bahadori & Hassanzadeh, 2026), with (Bahadori & Hassanzadeh, 2026) in particular identifying SHAP and LIME-stability as an open gap to be quantified in future work.

Table 6. Main stability techniques employed in the reviewed studies.

Stability Method	N° Studies	References
Cross-validation (k-fold)	18	(Nwafor et al., 2024; Gür et al., 2025; Trinh, 2024; Kim & Cho, 2019; Vakrani et al., 2025; Ukpabi et al., 2025; C. Shen & Wu, 2025; Naufal & Suseno, 2023; H. Li et al., 2024; Manik et al., 2022; X. Zhang et al., 2024b; Y. Li et al., 2022; Y.-Z. Xu et al., 2019; Oliveira & Basso, 2025; Yadava, 2023; Q. Zhang et al., 2024; Nguyen et al., 2025; X. Li & Li, 2024)
Statistical tests (Wilcoxon, Friedman, <i>t</i> -test)	4	(Gür et al., 2025; Mathibela & Maposa, 2026; C. Shen & Wu, 2025; M. Yang et al., 2023)
Out-of-time validation	3	(Z. Wang et al., 2025; Oliveira & Basso, 2025; Yadava, 2023)
PSI	1	(Sudjianto & Burakov, 2025)
IV (information value)	1	(Sudjianto & Burakov, 2025)

Operational definition. For this review, stability is defined as the degree to which a scoring model preserves both its discriminatory performance and its score distribution when the evaluation conditions change, specifically across (i) resampling of the same data (k-fold cross-validation, bootstrap), (ii) out-of-time samples drawn from later periods (temporal robustness and concept drift), and (iii) shifts in population or macroeconomic regime (stress conditions). Operationally, it is captured by the consistency of AUC and accuracy across folds and time windows and by dedicated drift indicators such as the population stability index (PSI) and the information value (IV). This definition makes explicit the dimension that, as the corpus shows, is most often left unmeasured.

Critical Finding: More than 82.1% of studies (96/117) report no stability or robustness metric whatsoever. This represents the most significant methodological gap identified in the entire corpus.

Stability evidence is reinforced by several configuration-based and ensemble-based studies. A 5-fold cross-validation is used to compare SVM, NN and NB on a single dataset in (Nur et al., 2019), where the neural network reaches an AUC = 0.9158 at the cost of the longest execution time (3.058 s). The split-sensitivity analysis of (Nasution et al., 2023) reports CatBoost AUC moving from 0.781066 (70/30) to 0.789583 (80/20) and 0.80329 (90/10), confirming the dependence of out-of-sample performance on training-set size. Multi-layer deep networks are reported as stable under data noise in (Z. Wang et al., 2025), while the CNN–LSTM–ATT architecture of (L. Zhang, 2022) is reported as more robust than LR, RF, CNN and LSTM baselines. The microfinance neural network of (Yaroshchuk et al., 2019) reports an accuracy of 95.4% on the test sample, with a loan-decision error of 0.00469 (issue) and 0.00406 (prohibition) over 64 training epochs. A deep learning ensemble combining LSTM and AdaBoost is statistically validated through the Wilcoxon test (F. Shen et al., 2021), while a soft-attention SVM polynomial kernel is reported as the most stable kernel configuration, exhibiting the smallest FPR and FNR (Putri et al., 2021).

3.7. Interpretability Analysis

Interpretability is a fundamental concern for financial regulators in the governance of AI, notably the Consumer Financial Protection Bureau (CFPB) in the United States and the European Banking Authority (EBA) in Europe. In this context, interpretability is defined as the degree to which a human can comprehend the factors driving a model prediction. It is important to note that interpretability does not automatically mean fairness and a lack of interpretability does not necessarily mean unfairness. There also exist non-interpretable models that preserve fairness in their decision-making process. Table 7 summarizes the main interpretability approaches employed in the reviewed studies.

The interpretability landscape maps directly onto the three-tier structure.

Intrinsic transparency versus post hoc explanation. Two fundamentally different routes to interpretability must be distinguished. Intrinsically transparent (white-box) models, logistic regression coefficients, decision tree rules, and scorecards, are self-explanatory: the model is the explanation, and no external library is required. Post hoc XAI tools, SHAP, LIME, permutation feature importance, and partial dependence, are applied after training to approximate the behavior of opaque (black-box) models, and their explanations are themselves approximations that may vary across runs. This distinction is decisive under financial supervision: where regulation requires directly auditable, stable and reproducible reason codes (adverse-action notices under the CFPB, model documentation under Basel III and IFRS 9, and the transparency expectations of the EU AI Act for high-risk systems), intrinsic transparency offers the strongest guarantees, whereas post hoc explanations are admissible but require validation of their own stability. The mapping of each route to these regulatory scenarios is consolidated in the decision framework of Section 4.

Table 7. Frequency of main interpretability techniques.

Interpretability Method	N° Studies	References
Feature importance	35	(Nwafor et al., 2024; Hernes et al., 2023; Emmanuel et al., 2024; Zong, 2025; Mendrofa et al., 2023; Munsarif et al., 2022; Du & Zhang, 2023; Pertiwi et al., 2024a; Wu, 2025; Peng, 2022; Yao et al., 2022b; Vakrani et al., 2025; C. Shen & Wu, 2025; H. Li et al., 2024; Y. Li et al., 2022; Y.-Z. Xu et al., 2019; Yadava, 2023; X. Zhang et al., 2024b; Ying et al., 2025; Reja Sweet et al., 2024; Rahman, 2024; Yao et al., 2022a; Torrent et al., 2020; Alagic et al., 2024; B. Han, 2024; Muhammad et al., 2024; Uddin et al., 2022; Gerunov, 2020; Kuyoro et al., 2022; Z. Yang et al., 2025; Rao et al., 2023; Z. Wang et al., 2025; Chairunnisa, 2023; H. Wang & Liu, 2025; Zhou et al., 2020)
SHAP	18	(Nwafor et al., 2024; Hernes et al., 2023; Mathibela & Maposa, 2026; Bahloul et al., 2026; Oyeyemi et al., 2025; Peng, 2022; Vakrani et al., 2025; Ukpabi et al., 2025; Oliveira & Basso, 2025; Yadava, 2023; Ying et al., 2025; Nguyen et al., 2025; Ugarte Chacon et al., 2023; Bahadori & Hassanzadeh, 2026; Zheng et al., 2026; Torrent et al., 2020; Liu et al., 2024; Gafsi, 2025)
LIME	8	(Bahloul et al., 2026; Paz et al., 2025; X. Han et al., 2025; Oyeyemi et al., 2025; Yadava, 2023; Nguyen et al., 2025; Bahadori & Hassanzadeh, 2026; Liu et al., 2024)
PDP	1	(Y. Li et al., 2022)
ARSLRP	1	(Pavitha & Sugave, 2024)

Traditional models retain an inherent advantage: decision-tree rules and logistic regression coefficients require no post hoc tools—the model itself is the explanation. This transparency remains the primary argument for Traditional models in regulatory contexts where documentation must be auditable without dependency on third-party libraries. In the cited studies, interpretability is mainly achieved through intrinsic mechanisms: logistic regression coefficients and odds ratios are explicitly used as the interpretability channel (Hernes et al., 2023; Du & Zhang, 2023; Atif & Salmi, 2022; Pramesti et al., 2022; Reja Sweet et al., 2024; R. Zhang et al., 2025; Gerunov, 2020; Bispo & Tessmann, 2025), including within the penalized logistic tree regression (PLTR) of Dumitrescu et al. (2022), which couples them with decision-tree rules to deliver a hybrid intrinsic explanation. Decision-tree rules themselves serve as the interpretability vehicle in multiple works (Hernes et al., 2023; Zong, 2025; Demajo et al., 2020; Y.-Z. Xu et al., 2019; Reja Sweet et al., 2024; Gerunov, 2020). Ceteris-paribus charts are additionally reported in Hernes et al. (2023) to trace how a single variable change affects the prediction, and L1 regularization on logistic regression provides an embedded feature-selection form of interpretability in Atif and Salmi (2022). Beyond traditional models, attention mechanisms act as a built-in interpretability channel in the CNN–LSTM–ATT model (L. Zhang, 2022), while the IAIBS framework emphasizes self-interpretable features by combining an ARPD algorithm and agglomerative clustering with a logistic regression sub-model to keep predictions traceable across four public datasets (R. Zhang et al., 2025). Transparency-accountability indicators (Z. Wang et al., 2025), weighted by analytic hierarchy process, similarly provide a structured, auditable explanation layer that complements the deep architecture.

Machine Learning models rely primarily on post hoc tools: feature importance (35 studies) is the most widely used, leveraging the native impurity-based or permutation-based importance available in tree models. SHAP (18 studies) has emerged as the gold-standard post hoc explanation method, offering both global (summary plot) and local (force plot) explanations that satisfy many regulatory explanation requirements. LIME (eight studies) is the second most frequent post hoc explainer applied to ML models. Reported SHAP

findings include: in a stacked ensemble, current account status ($|SHAP| = 0.153$), loan duration (0.064) and savings account (0.063) as the top drivers, with SHAP applied at both base-learner and meta-learner levels (Mathibela & Maposa, 2026); credit amount, loan duration and age as the main global drivers, older borrowers being associated with lower default risk (Vakrani et al., 2025); basic information and lending behavior information dominating credit information for both XGBoost and LightGBM (Peng, 2022); and a feature-selection use of SHAP to retain the 20 most important variables (Torrent et al., 2020). Tree SHAP is used in (Muhammad et al., 2024) to rank features for non-performing loans and loan-loss provisions, while (Zheng et al., 2026) applies SHAP at the model-contribution level, quantifying how much each base learner (LDA, LR, SVM, RF, XGBoost) contributes to the meta-learner's prediction. SHAP is also embedded within broader XAI frameworks: (Ugarte Chacon et al., 2023) couples SHAP with the SAFE framework and the rank graduation box (Lorenz-Zonoids) for a sustainability/accuracy/fairness/explainability validation, and (Bahadori & Hassanzadeh, 2026) positions SHAP as more stable than LIME but computationally heavy for large-scale repeated application. Other ML-side interpretability evidence includes partial dependence plots applied to XGBoost (Y. Li et al., 2022), Bayesian networks providing what-if analyses (Liu et al., 2024), recursive feature elimination as an intrinsic feature-ranking method (Wawge, 2026; B. Han, 2024), and rule- or instance-based explanations (Anchors, ProtoDash) evaluated for completeness, effectiveness, understandability, trustworthiness and detail sufficiency in (Demajo et al., 2020). Filter- or wrapper-based feature-selection methods (information gain in (Emmanuel et al., 2024), Garson's measure for a neural network in (Muhammad et al., 2024), relative variable importance in (Uddin et al., 2022), Gini-Index ranking in (Kuyoro et al., 2022), and relief/maximum information coefficient/quasi-separable selection in (Rao et al., 2023)) provide additional, model-specific feature-importance signals.

Permutation Feature Importance. Among global, model-agnostic attribution methods, permutation feature importance (PFI) warrants explicit mention. PFI quantifies the importance of a variable by measuring the degradation in a performance metric (e.g., AUC or F1) when that variable's values are randomly permuted, thereby corrupting its information. Unlike impurity-based importance, which is intrinsic to tree ensembles, PFI is model-agnostic and applies equally to traditional, ML and DL models, acting as a bridge between coefficient-based interpretation and complex learners. In this review, PFI is subsumed within the feature importance category of Table 7 (which aggregates impurity- and permutation-based variants). We make its role explicit here given its foundational status in global interpretability.

Deep learning models use adaptive relevance scaling for layer-wise relevance propagation (ARSLRP) and PDP as architecture-specific attribution methods and attention mechanisms as built-in interpretability channels. Critically, 65% of all studies use no formal interpretability technique, a persistent gap that limits the operational deployability of many published models. More specifically, ARSLRP is applied to a multistage 1D CNN ensemble classifier to provide meaningful predictions for credit risk (Pavitha & Sugave, 2024), while attention mechanisms are mobilized both as a hybrid component of a CNN-LSTM-ATT architecture (L. Zhang, 2022) and as the core interpretability channel of the improved TabNet, where a serialized attention mechanism and a dynamic feature-weight allocation module deliver both instance-level and global explainability (Z. Wang et al., 2025). The DeepCreditRisk model combines an attention-based feature importance method with LIME for local feature contribution analysis, with risk-factor interaction maps and with time-series contribution tracking (X. Han et al., 2025). In temporal convolutional network backbones, C. Shen and Wu (2025) introduces a modified channel attention mechanism and a multi-scale dilated attention (MSDA) module within DilateFormer for joint tempo-

ral/channel attribution. Post hoc XAI is also applied to deep architectures: SHAP and LIME are used together in (Yadava, 2023) to compare feature contributions across CNN and classical models under external validation, in (Oyeyemi et al., 2025) to interpret an MLP that reaches an AUC = 0.98, and in (Nguyen et al., 2025) to evaluate model stability and prevent overfitting. Finally, Ying et al. (2025) foreshadows a SHAP- and SAFE-based explainability evaluation of an attention-enhanced hybrid boosting model.

Several studies frame the accuracy–interpretability tension explicitly. Ke (2025) reports that logistic regression preserves transparency but underperforms on non-linear patterns and imbalance, that random forest improves stability at the cost of interpretability, and that XGBoost and LightGBM deliver superior predictive ability and efficiency but with lower transparency. Bispo and Tessmann (2025) reach a similar conclusion in a Brazilian banking setting, where MLP and gradient boosting dominate logistic regression on every accuracy metric (Acc up to 94.00%, AUC up to 94.10%), but the authors recommend future research on hybrid models to recover the interpretability required by regulators. The penalized logistic tree regression (Dumitrescu et al., 2022) is positioned precisely on this frontier: it combines decision-tree effects and logistic-regression coefficients to deliver competitive accuracy with intrinsic interpretability and is explicitly recommended for regulatory environments. The XGBoost-based selection in (Shinde & Kale, 2023) illustrates a related operational trade-off, where recall is deliberately prioritized over precision to avoid missing default instances.

4. Analysis and Discussion

This section reads the 117 studies retained by the systematic review through the three lenses that, together, decide whether a credit-scoring model is actually usable in production: accuracy, stability and interpretability. The intent is not to crown a single best model but to give practitioners and researchers a working framework, a way to read the evidence and, from there, to choose the model that fits their own constraints. Following PRISMA reporting, the discussion stays strictly within what the corpus reports.

Interpreting cross-study comparisons. Because the reviewed studies differ markedly in sample size, feature sets, default definitions and validation protocols, the reported metric ranges should be read as descriptive indications rather than head-to-head measurements. In particular, near-perfect AUC values obtained on small public benchmarks are not directly comparable with the lower but more realistic figures reported on large operational datasets. We therefore frame comparisons by subgroup, small public benchmarks versus large real-world samples, and by model family and refrain from aggregating effect sizes; a formal meta-analysis would require the homogeneous effect-and-variance reporting that the corpus does not yet provide and is identified as future work.

Read together, the studies retained in this systematic review tell a coherent story about the state of credit-scoring research, and that story has more to do with how the field evaluates its models than with which model is, in the abstract, the best one. Three dimensions structure the evidence and must be considered jointly: accuracy, stability and interpretability. Accuracy has been the dominant lens for years, and the corpus reflects that machine learning approaches, particularly gradient boosting, sit at the upper end of the performance distribution; traditional models, such as logistic regression, remain competitive on smaller, well-structured datasets; and deep learning architectures show a wide range of results that depend strongly on tuning. The more important point, however, is that strong accuracy figures are conditional, not intrinsic: the same algorithm can sit at the top of one ranking and near the bottom of another depending on data preparation, feature engineering and validation choices, and reading the literature as a contest of headline scores misses what actually drives operational performance. Interpretability has matured visibly:

a small number of post hoc methods now carry most of the field, with SHAP emerging as the de facto standard because it offers both global and local explanations across very different model families, while feature importance remains the most accessible entry point on tree-based models, LIME and partial-dependence approaches act as useful cross-checks, and architecture-specific tools such as LRP cover the deep learning side. Even so, a substantial share of published models is still released without any formal interpretability technique, which limits how readily they can be deployed in regulated contexts where decisions must be explainable. Stability is, by far, the dimension the literature has overlooked: most studies report no stability or robustness metric at all, and there is no dedicated stream of research focusing solely on stability in credit scoring; it is consistently treated as an ancillary concern, when it is mentioned at all. This is the most important methodological gap surfaced by this review because stability, the capacity of a model to behave consistently across populations, datasets and time, is precisely what determines whether the accuracy reported in a paper survives contact with deployment. It is also the clearest opportunity for future work: a structured research program on stability, its definitions, its metrics and its trade-offs with the other two dimensions, has yet to be built. Taken as a whole, this review is best understood as a framework rather than a verdict. It does not point to a universally optimal model, and it should not be read that way. Instead, it gives practitioners and researchers a structured way to navigate the trade-offs: when auditability is the binding constraint, traditional models remain the safer choice, and the loss in raw performance is a price knowingly paid; when predictive performance on tabular financial data is the priority, gradient boosting paired with a robust post hoc explanation method is the strongest evidence-based option; when the data structure genuinely calls for deep learning, attention-based or tabular-specific architectures combined with serious stability testing and architecture-aware interpretability are the most defensible route. In every case, the choice is a deliberate trade-off, and the responsibility for arbitrating between accuracy, stability and interpretability stays with the reader, who alone knows the regulatory environment, the dataset characteristics, the deployment horizon and the level of explanation that their stakeholders will accept. The contribution of this work is to make that arbitration explicit, to show that interpretability is now well served by a maturing post hoc ecosystem, and, above all, to flag stability as the missing pillar of the credit-scoring evidence base, the dimension on which the next round of meaningful research is most needed.

From the trilemma to a model-selection framework. The evidence does not crown a single best model; it reveals a trilemma in which gains in one dimension (accuracy, stability, interpretability) are typically paid for by another. To convert this into actionable guidance, Table 8 maps recurring operational, regulatory and data scenarios to the model class best supported by the reviewed evidence, with the corresponding interpretability route and the stability safeguards required.

Read through this framework, the three dimensions cease to be isolated classification labels and become coupled decision axes: moving toward higher accuracy (ML/DL) raises the interpretability cost and, frequently, the stability risk, which must then be bought back through post hoc explanation and explicit robustness testing. The framework is intended to be reusable and extensible; each row can be refined as primary studies begin to report stability and interpretability as systematically as they already report accuracy.

Positioning against prior reviews. Our findings both confirm and extend earlier syntheses. The long-standing benchmark of Lessmann et al. (2015) established that advanced classifiers, particularly heterogeneous ensembles, outperform logistic regression on raw discrimination; our corpus corroborates this for gradient boosting while showing that the gap narrows on small, well-structured datasets and is frequently offset by interpretability and stability costs. Relative to the recent systematic review of Bahloul et al. (2026),

which centers on performance, fairness and explainability, our contribution is the explicit and joint treatment of stability as a third axis and the finding that more than 82,1% of studies report no stability metric. Compared with the AI-credit-scoring review of Bahadori and Hassanzadeh (2026), which emphasizes transparency, alternative data and real-time learning, we add a quantified mapping of interpretability techniques (Table 7) and an evidence-based model-selection framework (Table 8).

Table 8. Evidence-based framework for credit-scoring model selection.

Scenario/Context	Recommended Model Class	Interpretability Route	Required Stability Safeguards
Binding auditability (Basel III, IFRS 9, EU AI Act high-risk)	Traditional—LR scorecard, Decision Tree	Intrinsic (coefficients/rules)	Out-of-time validation + PSI monitoring
Predictive performance is the priority, moderate transparency needs	Gradient Boosting (XG-Boost/LightGBM/CatBoost)	Post hoc (SHAP: global + local)	Cross-validation + SHAP-stability checks across retraining
Large, structured tabular or sequential data	Deep Learning (TabNet, attention-based)	Architecture-aware (attention, ARS-LRP) + post hoc	Mandatory stress/out-of-time testing (high DL variance)
Small, well-structured dataset	Traditional (LR/LDA) remains competitive	Intrinsic	k-fold cross-validation
Severe class imbalance	ML ensembles with resampling (SMOTE) or cost-sensitive learning	Post hoc	Validate imbalance handling against temporal robustness
Undisclosed or heterogeneous data provenance	Prefer transparent baselines; treat results cautiously	Intrinsic first	Independent out-of-time replication

Algorithmic fairness in ML-based credit scoring.

The fairness imperative. Accuracy, stability and interpretability, the three lenses that structure this review, do not by themselves guarantee that a scoring model is fit for responsible deployment. A fourth dimension, algorithmic fairness, has become one of the most consequential challenges of AI-driven credit scoring: a model is unfair when its error rates or score distributions systematically disadvantage groups defined by protected attributes such as gender, race, age or national origin. Within the 117 studies retained in this study, fairness is engaged by only a small minority (e.g., Bahloul et al., 2026; Nwafor et al., 2024; Z. Yang et al., 2025), a notable gap in the literature otherwise focused on predictive performance.

Sources of bias and financial inclusion. Bias enters credit models through two main channels: historical labels that encode past discriminatory lending and proxy variables (postal code, occupation, transactional or device data) that remain correlated with protected attributes even after those attributes are formally excluded, producing disparate impact (Barocas & Selbst, 2016; Hurley & Adebayo, 2016). Machine learning and deep learning models are especially exposed because they exploit complex, high-order feature interactions. The stakes are direct for financial inclusion: thin-file and minority applicants can be systematically underserved, although well-designed ML pipelines can also extend credit to populations excluded by traditional scorecards (Z. Yang et al., 2025).

Fairness metrics, trade-offs and regulation. Fairness is not a single quantity. Competing formal criteria, including demographic parity, equal opportunity and equalized odds (Hardt et al., 2016) as well as calibration within groups, cannot generally be satisfied simultaneously (Kleinberg et al., 2017), so selecting a fairness definition is an irreducibly normative

choice rather than a purely technical one. In credit scoring, imposing such constraints carries a measurable but often modest cost in profit or predictive power (Mehrabani et al., 2021; Kozodoi et al., 2022). Fairness is at the same time a legal requirement: in the United States, the Equal Credit Opportunity Act and fair-lending rules prohibit discrimination and mandate adverse-action reasons, while in the European Union, Article 22 of the GDPR and the EU AI Act classify credit scoring as a high-risk application subject to bias monitoring, documentation and human oversight. Transparency is the operational prerequisite for these obligations, since the post hoc explanation tools discussed in Section 3.7 (SHAP, LIME) and intrinsically interpretable models (Rudin, 2019) are what allow unfair decisions to be detected, audited and contested.

These considerations position fairness as a fourth evaluation dimension, alongside accuracy, stability and interpretability, and motivate the research directions set out in Section 5.

5. Future Work

The gaps surfaced by this review point to five research directions that are concrete, urgent, and directly connected to what it takes to deploy a credit-scoring model responsibly.

The most pressing need is stability. The field has spent years refining accuracy benchmarks while largely ignoring whether those benchmarks hold up over time. Future work should treat out-of-time validation, PSI monitoring, and concept drift detection as mandatory evaluation and develop shared protocols that make stability results comparable across studies.

A second priority is moving interpretability from demonstration to validation. SHAP and LIME are now well established, but key questions remain open: do SHAP explanations stay consistent across retraining cycles? Do they satisfy the specific explanation requirements of the European Union Artificial Intelligence Act (EU AI Act) or the Consumer Financial Protection Bureau's (CFPB) adverse action notice standards in practice? That gap between tool availability and regulatory fitness is where future research is most needed.

Dataset diversity is a third concern. With 64.1% of studies omitting their data source and the corpus heavily concentrated on German and US markets, published findings generalize poorly to other credit environments. Curated, publicly available benchmarks spanning diverse geographies, borrower profiles, and economic cycles would meaningfully strengthen the evidence base.

Fourth, the interaction between class imbalance handling and model stability deserves dedicated attention. Synthetic minority oversampling technique (SMOTE) and cost-sensitive learning are widely used but rarely evaluated for their effect on temporal robustness or explanation consistency, a gap with direct operational consequences.

A sixth priority is algorithmic fairness. Because the reviewed corpus almost never measures it, fairness should be treated as a distinct evaluation dimension rather than an afterthought: future benchmarking should report group-wise performance together with at least one fairness metric (e.g., demographic parity or equalized odds) and assess fairness-aware algorithms across the preprocessing, in-processing and post-processing families. This is essential both for financial inclusion and for compliance with emerging regulation, such as the EU AI Act.

Finally, the field needs more industrial evidence. Most reviewed studies live in controlled academic settings. Replication artifacts, integration cost assessments, and real-world deployment reports are rare. Closing that gap is what will ultimately determine whether the accuracy gains documented in this corpus translate into systems that banks, auditors, and regulators can actually trust.

Two methodological extensions would further sharpen the evidence base. First, a bibliometric layer: co-citation and keyword co-occurrence mapping (e.g., via VOSviewer or the Bibliometrix R package) would complement the manual thematic synthesis with an objective view of the field's intellectual structure and emerging hotspots. Second, the agenda above should be operationalized into standardized sub-studies with shared protocols: a PROBAST-based quality appraisal of the corpus, a benchmarked out-of-time and PSI stability protocol applied to a common dataset, and a controlled comparison of intrinsic versus post hoc interpretability against specific regulatory explanation requirements (CFPB adverse-action notices, EU AI Act). Such templated designs would turn broad recommendations into reproducible, comparable research.

6. Conclusions

This systematic review consolidates evidence on traditional, machine learning and deep learning approaches to credit risk scoring across accuracy, stability and interpretability and finds no universally optimal model but a genuine trilemma in which a gain in one dimension is consistently paid for by another.

Gradient boosting, led by XGBoost, dominates accuracy; logistic regression remains competitive on small, well-structured datasets; and deep learning results vary the most, driven more by tuning and data preparation than by architecture. On interpretability, SHAP has become the de facto post hoc standard, while traditional models keep a structural edge through inherent transparency, yet a substantial share of studies still apply no formal interpretability technique. Stability is where the literature falls most conspicuously short; most studies report no robustness metric, and out-of-time validation remains the exception, even though it is what determines whether performance obtained in a controlled experiment survives live deployment.

In practice, the appropriate choice is scenario dependent: traditional models when auditability is the binding constraint under Basel III, IFRS 9 or the EU AI Act; gradient boosting paired with SHAP when predictive performance is the priority; and attention-based or tabular-specific deep architectures, such as TabNet, when the data structure calls for them, provided they are backed by rigorous stability testing and architecture-aware interpretability.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jrfm19070526/s1>, File S1: PRISMA Checklist.

Author Contributions: Conceptualization, R.A., K.S. and N.E.H.C.; methodology, R.A. and N.E.H.C.; formal analysis, R.A.; investigation, R.A. and K.S.; data curation, R.A.; validation, K.S. and N.E.H.C.; supervision, K.S. and N.E.H.C.; visualization, R.A.; writing—original draft preparation, R.A.; writing—review and editing, R.A. and K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this article and its Supplementary Materials. Additional information can be obtained from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used AI-assisted writing tools, specifically Microsoft Copilot (2026 version) and OpenAI ChatGPT (GPT-4.1), solely to support language editing. No results, analyses, or interpretations were generated by AI. All content was

reviewed, verified, and edited by the authors, who take full responsibility for the accuracy and integrity of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., & Selmanovic, E. (2024). Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. *Machine Learning and Knowledge Extraction*, 6(1), 53–77. [\[CrossRef\]](#)
- Almajid, A. S. (2022). Multilayer perceptron optimization on imbalanced data using SVM-SMOTE and One-Hot encoding for credit card default prediction. *Journal of Advances in Information Systems and Technology*, 3(2), 67–74. [\[CrossRef\]](#)
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. [\[CrossRef\]](#)
- Amrin, A., Pahlevi, O., & Rianto, H. (2025). Analisa komparasi model data mining algoritma C4.5, CHAID, dan random forest untuk penilaian kelayakan kredit. *Computer Science (CO-SCIENCE)*, 5(1), 49–57. [\[CrossRef\]](#)
- Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, 86, 105936. [\[CrossRef\]](#)
- Aruleba, I., & Sun, Y. (2025). An improved ensemble method with data resampling for credit risk prediction. *IEEE Access*, 13, 71275–71287. [\[CrossRef\]](#)
- Atif, D., & Salmi, M. (2022). Feature selection for credit risk classification. In A. Bennour, T. Ensari, Y. Kessentini, & S. Eom (Eds.), *Intelligent systems and pattern recognition* (Vol. 1589, pp. 165–179). Springer International Publishing. [\[CrossRef\]](#)
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *The Journal of the Operational Research Society*, 54(6), 627–635. [\[CrossRef\]](#)
- Bahadori, S., & Hassanzadeh, A. (2026). A comprehensive literature review on ai-based credit assessment in digital lending: Exploring models, transparency, alternative data, and real-time learning. *Expert Systems with Applications*, 318, 131872. [\[CrossRef\]](#)
- Bahloul, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. *Journal of Risk and Financial Management*, 19(2), 104. [\[CrossRef\]](#)
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. [\[CrossRef\]](#)
- Berhane, T., Melese, T., & Seid, A. M. (2024). Performance evaluation of hybrid machine learning algorithms for online lending credit risk prediction. *Applied Artificial Intelligence*, 38(1), 2358661. [\[CrossRef\]](#)
- Bispo, U. A., & Tessmann, M. S. (2025). Does deep learning with multilayer perceptron perform well in predicting credit risk? *Journal of Applied Finance & Banking*, 15(4), 1–12. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chairunnisa, N. (2023). Prediksi kemampuan pembayaran klien home credit menggunakan model random forest, decision tree, dan logistic regression. *Jurnal Elektronika Dan Teknik Informatika Terapan (JENTIK)*, 1(3), 140–147. [\[CrossRef\]](#)
- Chi, G., Uddin, M., Habib, T., Zhou, Y., Islam, R., & Chowdhury, A. (2021). A hybrid model for credit risk assessment: Empirical validation by real-world credit data. *The Journal of Risk Model Validation*, 14(4), 1–39. [\[CrossRef\]](#)
- Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447–1465. [\[CrossRef\]](#)
- Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable AI for interpretable credit scoring. *Computer Science & Information Technology (CS & IT)*, 185–203. [\[CrossRef\]](#)
- Doko, F., Kalajdziski, S., & Mishkovski, I. (2021). Credit risk model based on central bank credit registry data. *Journal of Risk and Financial Management*, 14(3), 138. [\[CrossRef\]](#)
- Du, M., & Zhang, Z. (2023). Modeling selection for credit risk measurement: Based on meta path features. *Tehnicky Vjesnik-Technical Gazette*, 30(2), 545–554. [\[CrossRef\]](#)
- Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178–1192. [\[CrossRef\]](#)
- Emmanuel, I., Sun, Y., & Wang, Z. (2024). A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data*, 11(1), 23. [\[CrossRef\]](#)
- Fausser, D. V., & Gruener, A. (2020). Corporate social irresponsibility and credit risk prediction: A machine learning approach. *Credit and Capital Markets—Kredit Und Kapital*, 53(4), 513–554. [\[CrossRef\]](#)
- Gafsi, N. (2025). Machine learning approaches to credit risk: Comparative evidence from participation and conventional banks in the UK. *Journal of Risk and Financial Management*, 18(7), 345. [\[CrossRef\]](#)
- Gao, G., Wang, H., & Gao, P. (2021). Establishing a credit risk evaluation system for SMEs using the soft voting fusion model. *Risks*, 9(11), 202. [\[CrossRef\]](#)
- Gerunov, A. (2020). Classification algorithms for modeling economic choice. *Economic Thought Journal*, 65(2), 45–67. [\[CrossRef\]](#)

- Gür, Y. E., Toğaçar, M., & Solak, B. (2025). Integration of CNN models and machine learning methods in credit score classification: 2D image transformation and feature extraction. *Computational Economics*, 65(5), 2991–3035. [\[CrossRef\]](#)
- Han, B. (2024). Evaluating machine learning techniques for credit risk management: An algorithmic comparison. *Applied and Computational Engineering*, 112(1), 29–34. [\[CrossRef\]](#)
- Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. *Symmetry*, 17(3), 341. [\[CrossRef\]](#)
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (Vol. 29, pp. 3315–3323). Curran Associates, Inc.
- Hernes, M., Adaszyński, J., & Tutak, P. (2023). Credit risk modeling using interpreted XGBoost. *European Management Studies*, 2023(3), 46–70. [\[CrossRef\]](#)
- Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4), 847–856. [\[CrossRef\]](#)
- Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, 18(1), 148–216.
- James, C. (2021). Explainable AI in credit scoring: Improving transparency and accountability. *Journal of Intelligent Systems and Emerging Management*, 3(2), 45–60.
- Jumaa, M., Saqib, M., & Attar, A. (2023). Improving credit risk assessment through deep learning-based consumer loan default prediction model. *International Journal of Finance & Banking Studies*, 12(1), 85–92. [\[CrossRef\]](#)
- Kanaparthi, V. (2024). AI-based personalization and trust in digital finance (version 1). *arXiv*. [\[CrossRef\]](#)
- Kandi, K., & García-Dopico, A. (2025). Enhancing performance of credit card model by utilizing LSTM networks and XGBoost algorithms. *Machine Learning and Knowledge Extraction*, 7(1), 20. [\[CrossRef\]](#)
- Ke, C. (2025). Exploring the trade-off between accuracy and transparency in credit risk prediction models. *Theoretical and Natural Science*, 142(1), 119–126. [\[CrossRef\]](#)
- Kim, J.-Y., & Cho, S.-B. (2019). Towards repayment prediction in peer-to-peer social lending using deep learning. *Mathematics*, 7(11), 1041. [\[CrossRef\]](#)
- Klein, T., & Walther, T. (2024). Advances in explainable artificial intelligence (xAI) in finance. *Finance Research Letters*, 70, 106358. [\[CrossRef\]](#)
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th innovations in theoretical computer science conference (ITCS 2017)* (pp. 43:1–43:23). Association for Computing Machinery. [\[CrossRef\]](#)
- Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3), 1083–1094. [\[CrossRef\]](#)
- Kuyoro, A. O., Ogunyolu, O. A., Ayanwola, T. G., & Ayankoya, F. Y. (2022). Dynamic effectiveness of random forest algorithm in financial credit risk management for improving output accuracy and loan classification prediction. *Ingénierie des Systèmes d'Information*, 27(5), 815–821. [\[CrossRef\]](#)
- Kwon, S., Jang, J., & Kim, C. O. (2025). Credit scoring using multi-task Siamese neural network for improving prediction performance and stability. *Expert Systems with Applications*, 259, 125327. [\[CrossRef\]](#)
- 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(1), 124–136. [\[CrossRef\]](#)
- Li, H., Lan, Q., & Xiong, Q. (2024). Methodology for smooth transition from experience-based to data-driven credit risk assessment modeling under data scarcity. *Mathematics*, 12(15), 2410. [\[CrossRef\]](#)
- Li, X., & Li, L. (2024). A deep learning model-based approach to financial risk assessment and prediction. *Applied Mathematics and Nonlinear Sciences*, 9(1), 20230489. [\[CrossRef\]](#)
- Li, Y., Stasinakis, C., & Yeo, W. M. (2022). A hybrid XGBoost-MLP model for credit risk assessment on digital supply chain finance. *Forecasting*, 4(1), 184–207. [\[CrossRef\]](#)
- Liu, J., Liu, S., Li, J., & Li, J. (2022a). Financial credit risk assessment of online supply chain in construction industry with a hybrid model chain. *International Journal of Intelligent Systems*, 37(11), 8790–8813. [\[CrossRef\]](#)
- Liu, J., Zhang, S., & Fan, H. (2022b). A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. *Expert Systems with Applications*, 195, 116624. [\[CrossRef\]](#)
- Liu, J., Zhang, X., & Xiong, H. (2024). Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation. *Journal of Forecasting*, 43(5), 1625–1660. [\[CrossRef\]](#)
- Maitanmi, O. S., Ogunyolu, O. A., & Kuyoro, A. O. (2024). Evaluation of financial credit risk management models based on gradient descent and meta-heuristic algorithms. *Ingénierie des Systèmes d'Information*, 29(4), 1441–1452. [\[CrossRef\]](#)
- Manik, A., Nababan, E. B., & Tulus, T. (2022). Improved support vector machine performance using particle swarm optimization in credit risk classification. *Jurnal Teknik Informatika (Jutif)*, 3(6), 1739–1746. [\[CrossRef\]](#)
- Mathibela, M. R., & Maposa, D. (2026). Predictive modelling of credit default risk using machine learning and ensemble techniques. *Mathematical and Computational Applications*, 31(2), 45. [\[CrossRef\]](#)

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. [\[CrossRef\]](#)
- Mendrofa, R. D., Siallagan, M. H., Amalia, J., & Pakpahan, D. P. (2023). Credit risk analysis with extreme gradient boosting and adaptive boosting algorithm. *Journal of Information System, Graphics, Hospitality and Technology*, 5(1), 1–7. [\[CrossRef\]](#)
- Montevechi, A. A., Miranda, R. D. C., Medeiros, A. L., & Montevechi, J. A. B. (2024). Advancing credit risk modelling with machine learning: A comprehensive review of the state-of-the-art. *Engineering Applications of Artificial Intelligence*, 137, 109082. [\[CrossRef\]](#)
- Muhammad, A. B., Olawoyin, I. O., Yahaya, A., Gulumbe, S. U., Muhammad, A. A., & Salisu, I. A. (2024). Credit risk analysis: An assessment of the performance of six machine learning techniques in credit scoring modelling. *Fudma Journal of Sciences*, 8(6), 163–173. [\[CrossRef\]](#)
- Munsarif, M., Sam'an, M., & Safuan, S. (2022). Peer to peer lending risk analysis based on embedded technique and stacking ensemble learning. *Bulletin of Electrical Engineering and Informatics*, 11(6), 3483–3489. [\[CrossRef\]](#)
- Nasution, F. A., Saadah, S., & Yunanto, P. E. (2023). Credit risk detection in peer-to-peer lending using CatBoost. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(5), 1056–1062. [\[CrossRef\]](#)
- Naufal, A. R., & Suseno, A. T. (2023). Penerapan fitur seleksi dan particle swarm optimization pada algoritma support vector machine untuk analisis credit scoring. *Journal of Computer System and Informatics (JoSYC)*, 5(1), 184–195. [\[CrossRef\]](#)
- Nguyen, Q. G., Nguyen, L. H., Hosen, M. M., Rasel, M., Shorna, J. F., Mia, M. S., & Khan, S. I. (2025). Enhancing credit risk management with machine learning: A comparative study of predictive models for credit default prediction. *The American Journal of Applied Sciences*, 7(1), 21–30. [\[CrossRef\]](#)
- Nur, I. T. A., Setiawan, N. Y., & Bachtiar, F. A. (2019). Perbandingan performa metode klasifikasi SVM, neural network, dan naive bayes untuk mendeteksi kualitas pengajuan kredit di koperasi simpan pinjam. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 6(4), 444. [\[CrossRef\]](#)
- Nwafor, C. N., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach. *Scientific Reports*, 14(1), 25174. [\[CrossRef\]](#) [\[PubMed\]](#)
- Oliveira, N. A. D., & Basso, L. F. C. (2025). Explaining corporate ratings transitions and defaults through machine learning. *Algorithms*, 18(10), 608. [\[CrossRef\]](#)
- Osei, S., Mpinda, B. N., Sadefo-Kamdem, J., & Fadugba, J. (2021). *Accuracies of some learning or scoring models for credit risk measurement*. Working Papers hal-03194081. HAL.
- Oyeyemi, D. O., Okosieme, O. O., Idowu-Kunlere, T., Okosieme, O., Moussa, A. H., & Julius, E. A. (2025). AI-driven credit risk models for small-scale lending: A business analytics framework for predictive performance and responsible deployment. *International Journal of Science and Research Archive*, 17(1), 168–184. [\[CrossRef\]](#)
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021b). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88, 105906. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pavitha, N., & Sugave, S. (2024). Explainable multistage ensemble 1D convolutional neural network for trust worthy credit decision. *International Journal of Advanced Computer Science and Applications*, 15(2), 351. [\[CrossRef\]](#)
- Paz, Á., Crawford, B., Monfroy, E., Barrera-García, J., Peña Fritz, Á., Soto, R., Cisternas-Caneo, F., & Yáñez, A. (2025). Machine learning and metaheuristics approach for individual credit risk assessment: A systematic literature review. *Biomimetics*, 10(5), 326. [\[CrossRef\]](#) [\[PubMed\]](#)
- Peng, X. (2022). Research on personal credit risk assessment based on combination weight and shap interpretable machine learning. *Academic Journal of Computing & Information Science*, 5(10), 54–59. [\[CrossRef\]](#)
- Pertiwi, D. A. A., Ahmad, K., Salahudin, S. N., Annegrat, A. M., & Muslim, M. A. (2024a). Using genetic algorithm feature selection to optimize XGBoost performance in Australian credit. *Journal of Soft Computing Exploration*, 5(1), 92–98. [\[CrossRef\]](#)
- Pertiwi, D. A. A., Ahmad, K., Unjung, J., & Muslim, M. A. (2024b). A performance comparison of data balancing model to improve credit risk prediction in P2P lending. *Scientific Journal of Informatics*, 11(4), 881–890. [\[CrossRef\]](#)
- Pramesti, S. H., Indahwati, I., & Syafitri, U. D. (2022). Analisis regresi logistik dan cart untuk credit scoring dengan penanganan kelas tak seimbang. *Xplore: Journal of Statistics*, 11(3), 226–237. [\[CrossRef\]](#)
- Putri, N. H., Fatekurohman, M., & Tirta, I. M. (2021). Credit risk analysis using support vector machines algorithm. *Journal of Physics: Conference Series*, 1836(1), 012039. [\[CrossRef\]](#)
- Quan, J., & Sun, X. (2024). Credit risk assessment using the factorization machine model with feature interactions. *Humanities and Social Sciences Communications*, 11(1), 234. [\[CrossRef\]](#)

- Rahman, I. (2024). Prediction of creditworthiness of borrowers for mortgage lending using random forest algorithm. *International Journal of Scientific Research in Engineering and Management*, 8(5), 1–5. [CrossRef]
- Rao, C., Liu, Y., & Goh, M. (2023). Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost model. *Complex & Intelligent Systems*, 9(2), 1391–1414. [CrossRef]
- Reja Sweet, M. M., Arif, M., Uddin, A., Sharif, K. S., Tusher, M. I., Devi, S., Ahmed, M. P., Bhuiyan, M., Rahman, M. H., Mamun, A. A., Rahman, T., Asaduzzaman, M., Ahmmed, M. J., & Islam Sarkar, M. A. (2024). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *International Journal on Computational Engineering*, 1(3), 62–67. [CrossRef]
- 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. [CrossRef] [PubMed]
- Seo, J. Y. (2020). Machine learning in consumer credit risk analysis: A review. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 6440–6445. [CrossRef]
- Shen, C., & Wu, J. (2025). Research on credit risk of listed companies: A hybrid model based on TCN and DilateFormer. *Scientific Reports*, 15(1), 2599. [CrossRef] [PubMed]
- Shen, F., Zhao, X., Kou, G., & Alsaadi, F. E. (2021). A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, 98, 106852. [CrossRef]
- Shinde, S., & Kale, S. (2023). Unleashing the power of machine learning: A comparative study of classification algorithms for credit risk assessment. *International Journal of Advanced Research in Science, Communication and Technology*, 3(16), 552–556. [CrossRef]
- Siddiqi, N. (Ed.). (2012). *Credit risk scorecards: Developing and implementing intelligent credit scoring* (1st ed.). Wiley. [CrossRef]
- Sudjianto, A., & Burakov, D. (2025). An information-theoretic framework for credit risk modeling: Unifying industry practice with statistical theory for fair and interpretable scorecards (version 1). *arXiv*. [CrossRef]
- Sugiyarto, I., Sudarsono, B., & Faddillah, U. (2019). Performance comparison of data mining algorithm to predict approval of credit card. *Sinkron*, 4(1), 149. [CrossRef]
- Sun, P., Jia, Y., Shi, Y., Ren, J., Li, Z., & Li, X. (2025). Artificial intelligence credit risk assessment model based on MLP-hybrid clustering. *Complexity*, 2025(1), 3308222. [CrossRef]
- Thomas, L. C., Edelman, D., & Crook, J. N. (2002). *Credit scoring and its applications*. Society for Industrial and Applied Mathematics. [CrossRef]
- Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. [CrossRef] [PubMed]
- Torrent, N. L., Visani, G., & Bagli, E. (2020). PSD2 explainable AI model for credit scoring (version 3). *arXiv*. [CrossRef]
- Trinh, L. T. (2024). A comparative analysis of consumer credit risk models in peer-to-peer lending. *Journal of Economics, Finance and Administrative Science*, 29(58), 346–365. [CrossRef]
- Uddin, M. S., Chi, G., Al Janabi, M. A. M., & Habib, T. (2022). Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *International Journal of Finance & Economics*, 27(3), 3713–3729. [CrossRef]
- Ugarte Chacon, D., Lee, S., & Park, J. (2023). *An explainable machine learning model for consumer credit scoring in Mexico*. KDI School of Pub Policy & Management Paper No. 23-09. Available online: <https://ssrn.com/abstract=4540099> (accessed on 14 June 2026). [CrossRef]
- Ukpabi, K. H., Gambo, F. L., Abdullahi, A., & Ibrahim, A. M. (2025). Development and evaluation of machine learning model for loan default risk assessment. *International Journal of Nature and Science Advance Research*, 9(9). [CrossRef]
- Vakrani, D. S., Padhye, P. S., Gupta, S. K., Roy, J. K., Nerlekar, V. S., & Parashar, N. (2025). Evaluating AI-driven credit scoring models versus traditional statistical techniques. *Discover Artificial Intelligence*, 6(1), 72. [CrossRef]
- Wang, H., & Liu, M. (2025). Credit risk assessment of green supply chain finance for SMEs based on multi-source information fusion. *Sustainability*, 17(4), 1590. [CrossRef]
- Wang, S., & Zhang, X. (2025). Credit rating model based on improved TabNet. *Mathematics*, 13(9), 1473. [CrossRef]
- Wang, Z., Wang, C., Bai, Z., & Song, S. (2025). Green credit risk identification and anti-corruption measures under the application of the multi-layer deep network. *Humanities and Social Sciences Communications*, 12(1), 1311. [CrossRef]
- Wawge, S. J. (2026). Predicting credit scoring in the fintech era: A study of machine learning algorithms. In A. Iglesias, J. Shin, N. Bhatt, & A. Joshi (Eds.), *Information systems for intelligent systems* (Vol. 1748, pp. 64–74). Springer Nature. [CrossRef]
- Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *The Journal of Financial and Quantitative Analysis*, 15(3), 757. [CrossRef]
- Wu, H. (2025). From scores to decisions: Comparing logistic regression, random forest, and XGBoost for calibrated, cost-sensitive credit default prediction. *Advances in Economics, Management and Political Sciences*, 240(1), 70–79. [CrossRef]
- Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, 93, 182–199. [CrossRef]

- Xu, T. (2024). Credit risk assessment using a combined approach of supervised and unsupervised learning. *Journal of Computational Methods in Engineering Applications*, 4(1), 1–12. [\[CrossRef\]](#)
- Xu, Y.-Z., Zhang, J.-L., Hua, Y., & Wang, L.-Y. (2019). Dynamic credit risk evaluation method for E-commerce sellers based on a hybrid artificial intelligence model. *Sustainability*, 11(19), 5521. [\[CrossRef\]](#)
- Yadava, A. (2023). AI-driven credit risk assessment: Enhancing financial decision-making in SME lending using deep learning algorithms. *International Journal of Innovative Research in Computer and Communication Engineering*, 11(06), 8443–8454. [\[CrossRef\]](#)
- Yang, M., Lim, M. K., Qu, Y., Li, X., & Ni, D. (2023). Deep neural networks with L1 and L2 regularization for high dimensional corporate credit risk prediction. *Expert Systems with Applications*, 213, 118873. [\[CrossRef\]](#)
- Yang, Z., Lin, J., Qian, Y., & Shu, L. (2025). Machine learning and financial inclusion: Evidence from credit risk assessment of small-business loans in China. *The Singapore Economic Review*, 1–32. [\[CrossRef\]](#)
- Yao, G., Hu, X., Song, P., Zhou, T., Zhang, Y., Yasir, A., & Luo, S. (2024). AdaFNDFS: An AdaBoost ensemble model with fast nondominated feature selection for predicting enterprise credit risk in the supply chain. *International Journal of Intelligent Systems*, 2024(1), 5529847. [\[CrossRef\]](#)
- Yao, G., Hu, X., & Wang, G. (2022a). A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. *Expert Systems with Applications*, 200, 117002. [\[CrossRef\]](#)
- Yao, G., Hu, X., Zhou, T., & Zhang, Y. (2022b). Enterprise credit risk prediction using supply chain information: A decision tree ensemble model based on the differential sampling rate, synthetic minority oversampling technique and AdaBoost. *Expert Systems*, 39(6), e12953. [\[CrossRef\]](#)
- Yaroshchuk, S. O., Shapovalova, N. N., Striuk, A. M., Rybalchenko, O. H., Dotsenko, I. O., & Bilashenko, S. V. (2019, November 29). *Credit scoring model for microfinance organizations*. Student Workshop on Computer Science & Software Engineering (pp. 115–127), Kryvyi Rih, Ukraine.
- Ying, C., Shi, A., & Li, X. (2025). Hybrid boosted attention-based LightGBM framework for enhanced credit risk assessment in digital finance. *Humanities and Social Sciences Communications*, 12(1), 1036. [\[CrossRef\]](#)
- Yu, T., Huang, W., Tang, X., & Zheng, D. (2025). A hybrid unsupervised machine learning model with spectral clustering and semi-supervised support vector machine for credit risk assessment. *PLoS ONE*, 20(1), e0316557. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, C., & Zhou, X. (2023). Forecasting credit risk of SMEs in supply chain finance using bayesian optimization and XGBoost. *Mathematical Problems in Engineering*, 2023(1), 5609996. [\[CrossRef\]](#)
- Zhang, L. (2022). The evaluation on the credit risk of enterprises with the CNN-LSTM-ATT model. *Computational Intelligence and Neuroscience*, 2022, 6826573. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, Q., Zhang, C., & Zhao, X. (2024). Credit risk classification prediction based on optimised adaboost algorithm with long short-term memory neural network (LSTM). *Advances in Economics, Management and Political Sciences*, 87(1), 152–158. [\[CrossRef\]](#)
- Zhang, R., Li, I., & Ding, Z. (2025). An interpretable credit risk assessment model with boundary sample identification. *PeerJ Computer Science*, 11, e2988. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, W., Yan, S., Li, J., Tian, X., & Yoshida, T. (2022). Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E: Logistics and Transportation Review*, 158, 102611. [\[CrossRef\]](#)
- Zhang, X., Ma, Y., & Wang, M. (2024a). An attention-based Logistic-CNN-BiLSTM hybrid neural network for credit risk prediction of listed real estate enterprises. *Expert Systems*, 41(2), e13299. [\[CrossRef\]](#)
- Zhang, X., Xu, L., Li, N., & Zou, J. (2024b). Research on credit risk assessment optimization based on machine learning. *Applied and Computational Engineering*, 69(1), 173–178. [\[CrossRef\]](#)
- Zheng, C., Weng, F., Zhu, M., & Zhang, Z. (2026). Credit scoring modeling based on an improved two-stage method with fragmentary data. *Journal of Big Data*, 13(1), 32. [\[CrossRef\]](#)
- Zhou, X., Zhang, W., & Jiang, Y. (2020). Personal credit default prediction model based on convolution neural network. *Mathematical Problems in Engineering*, 2020, 5608392. [\[CrossRef\]](#)
- Zong, Z. (2025). Credit risk assessment model based on XGBoost. *Advances in Economics, Management and Political Sciences*, 193(1), 171–180. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.