American Journal of Interdisciplinary Studies

Volume: 7; **Issue:** 1
Pages: 459-495
eISSN: 3067-5146

A Systematic Review of AI-Driven Credit Risk Assessment Models in Commercial Banking (2018–2026)

Rajib Sarkar[1];

[1]. Master of Business Administration, Washington University in St. Louis, Olin Business School, St. Louis, Missouri; USA
Email: sarkarraj.0306@gmail.com

Doi: 10.63125/m52yna23
Received: *19 December 2025;* ***Revised:*** *21 January 2026;* ***Accepted:*** *09 February 2026;* ***Published:*** *01 March 2026*

Abstract

This systematic review examines the evolution, methodological advancements, and governance implications of AI-driven credit risk assessment models in commercial banking from 2018 to 2026. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, the review synthesizes evidence from 27 peer-reviewed studies, representing a diverse body of research exploring machine learning and deep learning applications in probability of default estimation, loss and exposure modeling, early-warning systems, portfolio monitoring, and automated credit decisioning. The findings show that AI models consistently outperform traditional statistical approaches in predictive accuracy and behavioral sensitivity, particularly when leveraging ensemble architectures and temporal or transactional features. However, despite these technical advantages, the review identifies substantial constraints related to data quality, model generalizability, explainability, fairness, drift vulnerability, and regulatory acceptability. Many studies highlight persistent challenges in achieving transparency and stability under stress scenarios, indicating that current AI systems often struggle to meet prudential and consumer-protection expectations. The review also notes a widening gap between research innovation and real-world deployment, as operational requirements such as continuous monitoring, documentation, and interoperability create significant barriers to adoption. Overall, this study provides a comprehensive evidence base demonstrating both the promise and limitations of AI credit risk models, emphasizing the need for more interpretable model architectures, standardized validation frameworks, privacy-preserving data ecosystems, and cross-institutional benchmarking. These insights contribute to shaping the next generation of trustworthy, governable, and regulator-aligned AI systems for commercial banking.

Keywords

AI Credit Risk, Machine Learning, Banking Models, Systematic Review, Governance

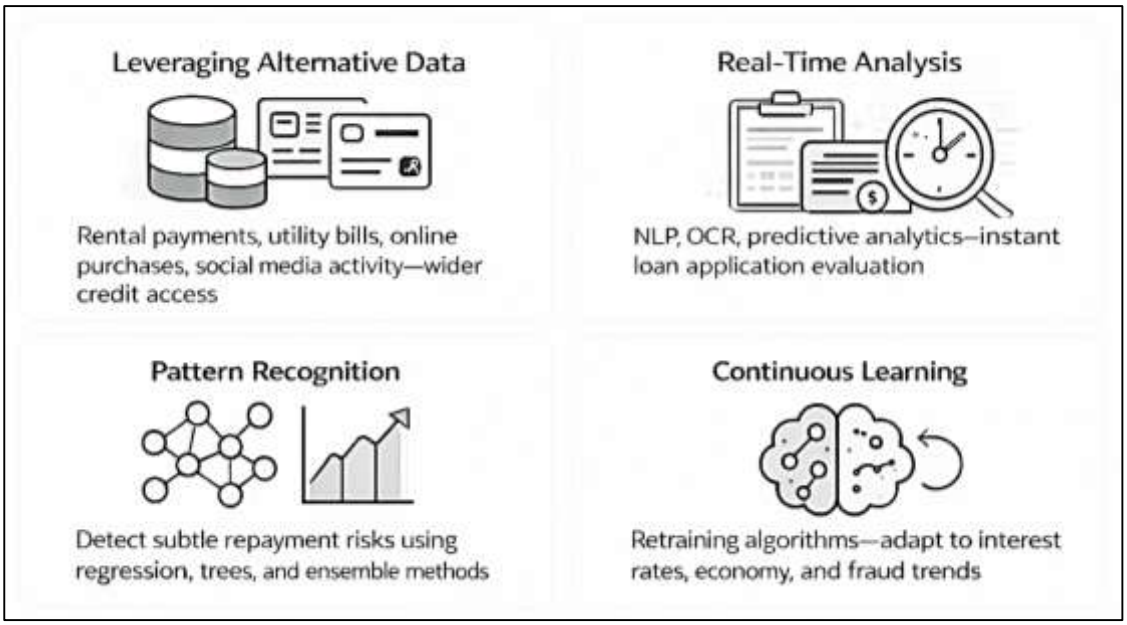
INTRODUCTION

Credit risk in commercial banking is commonly defined as the probability that a borrower or counterparty will fail to meet contractual debt obligations in full and on time, generating loss for the lender through write-offs, recovery costs, and reduced earnings. Within modern banking regulation and accounting, this definition is operationalized through measurable components such as probability of default, loss given default, and exposure at default, which are also used to compute expected credit loss under IFRS 9 frameworks that became central to bank provisioning from 2018 onward (Yanenkova et al., 2021). In parallel, credit risk assessment models are the quantitative systems used to transform borrower information into decisions about approval, pricing, limits, monitoring, and portfolio capital allocation, and in many institutions these systems operate alongside governance requirements for model development, validation, and ongoing performance monitoring. In the context of AI-driven credit risk assessment, "artificial intelligence" is typically used as an umbrella term for computational methods that learn patterns from data, while "machine learning" denotes algorithmic techniques—supervised, semi-supervised, or hybrid—designed to infer predictive relationships between borrower features and default outcomes. The practical core of AI-driven credit risk assessment in banks is therefore the modeling pipeline that ingests structured and increasingly non-traditional data, trains predictive algorithms, validates them under governance constraints, and deploys them into decision processes that require auditability and accountability (T. M. Yhip & B. M. Alagheband, 2020). This definitional grounding matters internationally because credit intermediation is a foundation of economic growth and financial stability, and because banking systems are interconnected through cross-border lending, trade finance, correspondent banking, and common regulatory expectations. For that reason, supervisory communities have explicitly engaged the topic of AI and machine learning in banking and risk management, including the need for robust risk controls around these methods. In empirical research, the definition of "AI-driven credit risk assessment model" often narrows to measurable tasks such as binary default classification, ordinal credit rating assignment, or continuous loss estimation, and the literature uses standardized performance metrics (e.g., AUC, F1, calibration error) to evaluate how well these systems support bank decision processes (Zhevaga & Morgunov, 2021).

The international significance of AI-driven credit risk assessment from 2018–2026 is also tied to the global diffusion of digital financial services, the growth of high-frequency and high-volume lending channels, and the expansion of data environments that shape how banks evaluate borrowers. Evidence from international policy-oriented research shows that machine learning combined with non-traditional data can materially change credit scoring outcomes and access patterns, illustrating why jurisdictions have treated these methods as consequential for both competition and consumer outcomes (Bhatt et al., 2023). Likewise, large-scale evidence from fintech lending contexts has been used to compare "big data + machine learning" approaches with traditional bank scorecards, reinforcing that algorithmic credit risk assessment is not confined to a single region or banking model and can be evaluated at transaction scale. At the same time, credit risk modeling in commercial banking remains embedded in supervisory expectations about capital and governance, meaning that international adoption is shaped not only by predictive accuracy but also by transparency, documentation, and validation discipline. US supervisory guidance on model risk management frames models as tools that can materially affect business decisions and therefore require rigorous governance across development, validation, and use (Orlando & Pelosi, 2020). In Europe, supervisory discussions and publications have explicitly identified AI use in credit scoring, creditworthiness assessment, and regulatory credit risk modeling, situating AI-driven credit risk methods inside the broader agenda of safe and controlled innovation in the banking sector. This international governance setting is directly relevant to a systematic review covering 2018–2026 because it shapes what research questions are asked (accuracy, stability, fairness, explainability, calibration) and what deployment constraints are treated as non-negotiable (traceability, reproducibility, and documentation). Research surveys and systematic reviews have documented that the 2018 onward period is characterized by a sustained expansion of machine learning and deep learning techniques applied to credit risk, alongside increased attention to validation and comparability across datasets and institutions (Mhlanga, 2021). In other words, the international significance rests on the combination of (a) the central role of credit risk in bank safety and economic

activity and (b) the cross-border convergence toward measurable, governed, and reviewable modeling practices when AI methods are introduced.

Figure 1: Artificial Intelligence for Credit Risk

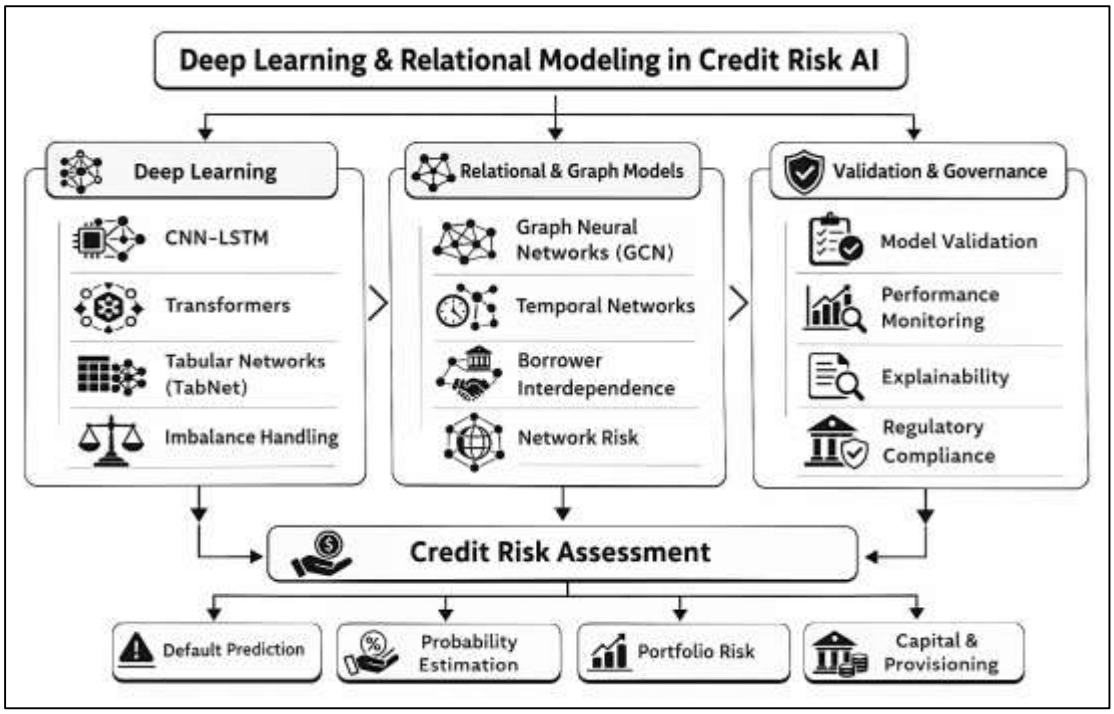


Within the 2018–2026 research window, supervised machine learning has remained the dominant paradigm for AI-driven credit risk assessment because banks and researchers frequently possess labeled outcomes (default/non-default, delinquency states, internal ratings transitions) that enable objective evaluation under standard experimental protocols. Comparative empirical studies have repeatedly benchmarked classical statistical models (such as logistic regression) against machine learning classifiers, reporting measurable differences in discrimination and error tradeoffs under consistent testing designs (Natufe & Evbayiro-Osagie, 2023). A key strand of this literature emphasizes tree-based methods because they balance predictive power with partial interpretability, offering feature importance, decision paths, and rule-like representations that can be documented for governance processes. Empirical work using machine and deep learning models for credit risk prediction has illustrated that ensemble methods can improve robustness and predictive stability on real credit datasets, including settings where feature interactions and nonlinearity are material. Research has also explored the role of class imbalance and rare-event structure, which is common in credit default data, and has examined methods such as support vector machines, sampling strategies, and imbalance-aware validation to maintain meaningful sensitivity to defaults without overwhelming false positives (Nyebar et al., 2023). In operational terms, this stream connects to commercial banking because retail and SME portfolios generate large volumes of heterogeneous observations, and model performance must be evaluated under institution-like distributions rather than idealized balanced samples. Applied studies in banking-adjacent contexts, including rural commercial bank settings and specialized enterprise credit evaluation, have continued to apply and compare models such as XGBoost, random forests, and hybrid architectures to reflect real-world constraints and data structures. An additional supervised-learning branch targets corporate and bond credit risk, where default events are even rarer and loss structures can be heavy-tailed, motivating algorithmic approaches tuned through cross-validation and grid-search procedures to stabilize performance estimates (Guzel, 2021). Across these supervised-learning studies, evaluation is typically operationalized via AUC, F1-score, recall, precision, and cost-sensitive errors, aligning the research design with banking needs for consistent ranking of obligors and stable decision thresholds over time and across segments.

Deep learning research has expanded within AI-driven credit risk assessment during the same period, driven by attempts to capture complex nonlinear relationships in borrower behavior, transaction patterns, and temporal dynamics that may be weakly represented through manual feature engineering

alone. Early applied work in this window framed deep learning as a credit scoring tool capable of improving classification performance when large datasets are available, and subsequent studies introduced architectures suited to sequential and time-series inputs. For example, CNN–LSTM hybrids and attention mechanisms have been used to model credit risk prediction for listed companies by combining convolutional extraction of local patterns with recurrent modeling of time dependence, reflecting the idea that borrower risk evolves with economic conditions and firm trajectories (Mahbobi et al., 2023).

Figure 2: AI Credit Risk Assessment Framework



Alongside sequence models, transformer-based approaches have been introduced into credit scoring and default prediction to learn representations of tabular and behavioral data through attention mechanisms, with empirical studies reporting comparative advantages against LSTM baselines and traditional machine learning in specific datasets. Deep learning research has also interacted with banking realism through efforts to predict not only one-year default events but also richer structures such as term-like horizons or multi-label outcomes, providing model outputs that can align with internal portfolio monitoring and rating horizons used by banks. A major concern in deep learning applications to credit risk is interpretability, and the period includes substantial interest in deep tabular architectures designed to be both accurate and explainable (T. M. Yhip & B. Alagheband, 2020). TabNet, for instance, was proposed as an attentive architecture for tabular learning that provides feature attributions through sequential attention, and later works have tested TabNet-style models in credit rating or credit risk settings to combine performance with interpretable signals. Beyond model architecture, deep learning in credit risk has been studied under practical issues such as imbalance handling (e.g., SMOTE-based hybrids) and data preparation pipelines, including explicit comparisons among gradient boosting, deep tabular networks, and hybrid ensembles on commercial-bank-like datasets. This deep learning stream is therefore part of the broader 2018–2026 landscape because it reflects how banks and researchers seek higher discrimination while keeping model behavior measurable, testable, and auditable within commercial banking processes (Firouzi & Meshkani, 2021). Another defining theme of the period is the expansion of data structures and relational modeling to represent borrower interdependence, network exposure, and dynamic interactions, especially in environments where borrowers are connected through transactions, supply chains, or shared economic

conditions. Graph-based machine learning is prominent here because credit risk is not purely an individual attribute; it can reflect correlated shocks, shared vulnerabilities, and information diffusion across borrower networks (Chen, 2024d; Faysal & Shamsunnahar, 2022). A graph convolutional network (GCN) approach to credit default prediction, for example, has been proposed to capture both nonlinear relationships between borrower attributes and higher-order relationships among borrowers, providing an explicit modeling pathway for interconnected risk. Research has also developed temporal-aware graph neural network methods to address dynamic borrower behavior and evolving risk profiles, reflecting that credit risk in operational settings is shaped by time-varying events rather than static snapshots. This shift toward relational and temporal learning is relevant to commercial banking because banks manage portfolios where risk clustering can emerge across customer segments, industries, geographies, and supply chains, and where correlated deterioration can matter for capital planning and concentration management (Chen et al., 2022; Habibullah & Zaheda, 2022; Jahangir & Md Shahab, 2022). In applied optimization and decision science outlets, integrated graph learning frameworks have been used to strengthen credit risk/default prediction, and these works position network structures as a formal mechanism to improve classification under complex dependency patterns. The period also includes related strands using enhanced feature construction or similarity-matrix mapping to represent borrower indicators as graph structures, showing broad methodological experimentation with how credit information is encoded before classification (Ratul & Subrato, 2022; Tahmina Akter Bhuya & Rebeka, 2022). At the same time, the international banking setting has maintained strong expectations for comparability and governance, meaning that graph-based modeling is often evaluated not only for discrimination but also for stability, data lineage, and reproducibility of relational inputs (Jahangir & Muhammad Mohiul, 2023; Jinnat & Molla Al Rakib, 2023; Shi et al., 2022). In empirical credit scoring and fintech lending research, the incorporation of large-scale behavioral and narrative information has also been explored with hybrid models that combine advanced representation learning with strong tabular learners (e.g., CatBoost hybrids), adding another dimension to the data-rich modeling landscape. Collectively, the relational-data strand reflects how AI-driven credit risk assessment research from 2018–2026 has broadened the definition of "credit risk input data" from static borrower profiles to structured representations of interaction, dependency, and time (El Hajj & Hammoud, 2023).

A systematic review focused on AI-driven credit risk assessment in commercial banking must also treat validation and governance as integral to the introduction because banking models are evaluated under methodological and supervisory scrutiny that extends beyond conventional academic prediction tasks. Model risk management guidance in the United States emphasizes that models influencing risk management and capital planning require extensive and rigorous frameworks, including independent validation, documentation, and performance monitoring. European policy discussions similarly highlight that AI is used in credit scoring and creditworthiness assessment, reinforcing that banking supervisors view AI methods as embedded in critical decision systems rather than optional analytics (Heng & Subramanian, 2022; Md Khaled & Md. Mosheur, 2023; Md Shahab & Aditya, 2023). In the research literature, this governance orientation appears in evaluation designs emphasizing cross-validation, out-of-sample testing, and disciplined partitioning to reduce sampling artifacts and information leakage. Benchmarking studies comparing multiple machine learning classifiers under consistent protocols illustrate that algorithm rankings can shift with feature selection, sampling decisions, and dataset properties, making validation design a core methodological object rather than a secondary step. Work on feature selection and modeling pipelines in credit scoring has formalized this by comparing combinations of classifiers and selection methods, often reporting that performance depends on the interaction between the feature engineering process and the learning algorithm (Mishra, Tyagi, & Arowolo, 2024). In credit risk settings with heavy imbalance, scholars have examined algorithm choices and evaluation measures that retain sensitivity to defaults, including sampling-aware approaches and classifier designs intended for imbalanced datasets, which aligns with banking concerns about correctly identifying the relatively small population of high-risk obligors. The governance theme is also reinforced by explainable AI research that targets interpretability as a banking constraint: interpretable credit scoring frameworks using XGBoost, for instance, explicitly integrate

explanation methods to present model logic in ways that can be audited and reviewed. In sum, within the 2018–2026 period, validation and governance are not peripheral topics; they define how AI-driven credit risk assessment research frames credibility, comparability, and suitability for commercial banking deployment (Mienye et al., 2024).

Finally, accounting and risk-measurement integration has shaped the period's research agenda because commercial banks are required to connect credit risk estimates to provisioning and capital narratives that are comparable across institutions and jurisdictions. IFRS 9's expected credit loss approach, implemented from 2018, created operational demand for forward-looking estimates and scenario-sensitive measures, which has stimulated research that adapts machine learning to components such as lifetime probability of default and related time-to-event structures (Hassani et al., 2020; Mostafa, 2023; Mostafa & Tahmina Akter Bhuya, 2023). Recent work has proposed machine learning survival-analysis approaches for modeling cumulative default probabilities over time in ways that align with lifetime expected credit loss estimation, directly linking AI methods to accounting-compatible constructs used by banks. Related applied studies and academic reviews produced by standard-setting and policy institutions have also documented the importance of expected credit loss recognition and the broader empirical literature examining how ECL frameworks affect loss timeliness and measurement practice (Ratul & Aditya, 2023; Rifat & Rebeka, 2023; Sandeep et al., 2022). In parallel, the AI credit risk literature has expanded severity- and exposure-adjacent modeling (including corporate default probability measures and bond default prediction) because credit losses are shaped by both the likelihood and the magnitude of adverse outcomes, and because capital frameworks depend on coherent aggregation logic. Methodologically, this integration theme explains why many studies remain anchored in probability estimation, calibration, and stable ranking, since these properties affect how predictive scores map into pricing, limits, provisioning, and portfolio monitoring. Explainable AI contributions in credit risk management and interpretable credit scoring emphasize that probability-like outputs and their explanations must be coherent for governance review, particularly when model outputs enter regulated financial reporting and internal controls (Al Janabi, 2024b; Faysal & Tahmina Akter Bhuya, 2024; Zaheda & Md. Tahmid Farabe, 2023). At the same time, the breadth of AI methods applied—ranging from gradient boosting and hybrid ensembles to graph neural networks and transformer architectures—shows that the period's literature treats credit risk assessment as a family of linked modeling tasks rather than a single classification problem. The resulting body of work between 2018 and 2026 is therefore best introduced as an intersection of (a) internationally governed banking risk measurement, (b) expanding AI method portfolios for classification and probability estimation, and (c) accounting-compatible requirements for forward-looking loss measurement, all evaluated under disciplined validation and interpretability expectations (Allioui & Mourdi, 2023; Md. Towhidul & Uddin, 2024; Sazzadul & Rebeka, 2024).

The primary objective of this systematic review is to critically evaluate and synthesize the advancements, methodological innovations, and governance implications of AI-driven credit risk assessment models in commercial banking between 2018 and 2026. This review aims to provide a rigorous and comprehensive understanding of how machine learning and deep learning frameworks have been developed, validated, and operationalized within regulated lending environments, and to determine the extent to which these models improve upon traditional statistical approaches. A central objective is to assess not only predictive performance but also the broader ecosystem of requirements that influence AI adoption in commercial banking, including data quality, model stability, operational monitoring, explainability, fairness, and regulatory compliance. By examining 27 peer-reviewed studies, this review seeks to identify recurring methodological patterns, highlight areas where AI models have demonstrated meaningful advantages, and uncover persistent weaknesses that limit their real-world deployment. Another key objective is to compare contemporary AI practices with earlier foundational credit risk modeling approaches, allowing for an assessment of how AI has transformed long-standing assumptions about default prediction, portfolio monitoring, and risk management. The review also aims to map the intersection between technological progress and prudential expectations, recognizing that AI systems must meet stringent standards for auditability, interpretability, and stress-testing readiness before they can be safely integrated into lending workflows. Additionally, this review

intends to clarify emerging research gaps—such as generalizability limitations, fairness concerns, and the need for reproducible datasets—thereby informing a forward-looking research agenda for the banking and academic communities.

LITERATURE REVIEW

From 2018 onward, commercial banks accelerated the use of AI-driven credit risk assessment to improve predictive accuracy, automate underwriting workflows, and incorporate richer (often alternative) data beyond traditional bureau and financial-statement variables (Ridzuan et al., 2024). Across the literature, this shift is typically framed as a movement from interpretable scorecard-style methods toward machine-learning (ML) and deep-learning (DL) models that can capture nonlinearities, complex interactions, and time-dependent default dynamics—especially under portfolio heterogeneity and changing macroeconomic regimes. Recent systematic reviews and comprehensive surveys show a steady expansion of model families (tree ensembles, gradient boosting, neural networks, and hybrid pipelines) alongside a widening set of evaluation practices (AUC/ROC, PR-AUC, calibration, cost-sensitive metrics, stability tests) and deployment constraints (latency, drift, governance, auditability) (Bahoo et al., 2024; Tasnim & Anick, 2024; Zaheda & Md Hamidur, 2024). At the same time, banking-specific constraints make credit risk a "high-friction" domain for advanced AI: models must be validated, monitored, and governed within strict model risk management expectations (eg, independent validation, documentation, controls, and oversight), and must also address explainability and fairness obligations in consumer and prudential contexts. Classic supervisory frameworks for model risk management remain foundational in practice, shaping how banks evidence robustness, limitations, and appropriate use. Between 2023 and 2026, the literature increasingly converges on a "performance–explainability–compliance" triangle: (i) performance gains from complex ML/DL, (ii) explainability requirements for internal governance and external accountability, and (iii) emerging legal classification of credit scoring/creditworthiness AI as higher-risk in some jurisdictions—pushing banks toward stronger documentation, risk controls, and ongoing monitoring (Amena Begum, 2025; Faysal & Aditya, 2025; Pamuk & Schumann, 2024). For example, European prudential discussions on ML in IRB contexts emphasize governance, risk management, and interpretability challenges, while recent EU-focused mapping work highlights additional safeguards for AI used in creditworthiness/credit scoring. Accordingly, this literature review synthesizes 2018–2026 evidence around (a) model families and methodological innovations, (b) data strategies and feature engineering choices, (c) evaluation and validation practices aligned with banking use-cases, (d) explainability, fairness, and accountability techniques, and (e) implementation realities—model risk management, regulatory alignment, monitoring, and lifecycle controls—so the review can connect algorithmic advances to what is feasible, defensible, and scalable in commercial banking (Al Janabi, 2024a).

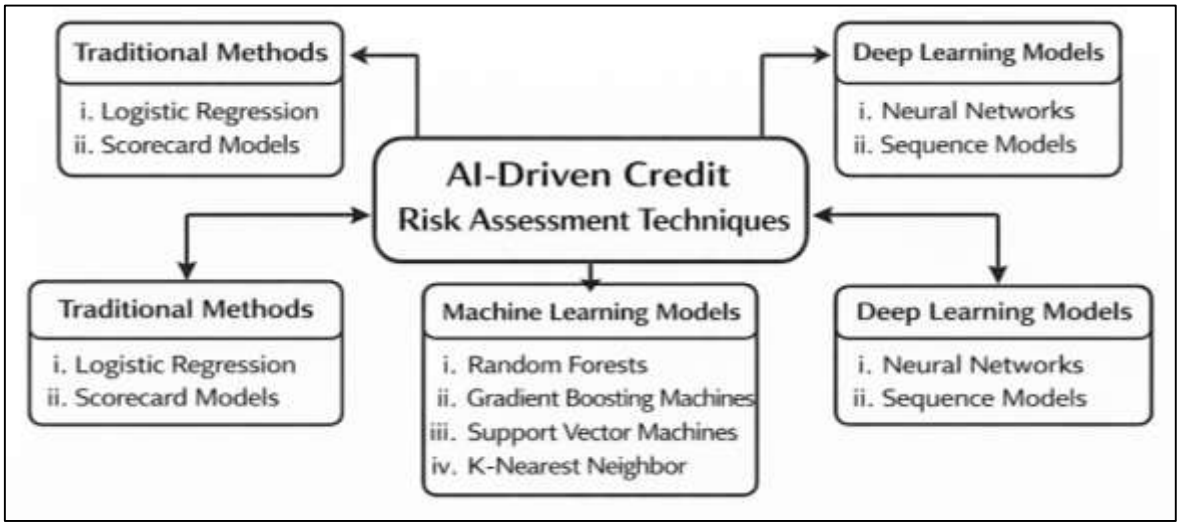
Scope, Definitions, and Review Framing for AI Credit Risk (2018–2026)

The scope of AI-driven credit risk assessment between 2018 and 2026 encompasses a multidimensional set of analytical tasks that extend beyond traditional probability of default (PD) modeling toward integrated estimations of loss given default (LGD), exposure at default (EAD), expected loss (EL), early warning indicators, and behavioral scoring across retail, SME, and corporate portfolios. Earlier credit risk frameworks relied predominantly on logistic regression and scorecard methodologies that emphasized interpretability and regulatory acceptance, particularly in retail lending contexts (Biju et al., 2024). However, a substantial body of empirical research has documented the growing application of machine learning (ML) algorithms—such as random forests, gradient boosting machines, support vector machines, and neural networks—to capture nonlinear borrower characteristics and interaction effects that traditional generalized linear models often fail to detect. Within retail portfolios, behavioral scoring systems incorporate transactional histories and repayment dynamics to continuously update PD estimates, thereby enhancing sensitivity to emerging delinquency patterns. In SME and corporate lending, AI-driven systems integrate financial ratios, sectoral indicators, and macroeconomic variables into higher-dimensional predictive architectures capable of modeling structural heterogeneity. Parallel developments have occurred in LGD and EAD estimation, where ensemble learning methods have been employed to model recovery variability, collateral dynamics, and exposure volatility with greater precision than linear regression-based approaches. Early warning systems have further expanded the

operational definition of credit risk by incorporating sequence-based models that analyze time-dependent behavioral signals prior to formal default events (Minkkinen et al., 2022). The boundary of "AI-driven" credit risk is therefore defined not simply by algorithmic novelty but by the adoption of machine learning, deep learning, or hybrid modeling architectures that extend beyond traditional statistical scorecards while maintaining integration within institutional credit decision processes.

AI-driven credit risk models operate within multiple commercial banking functions, including origination and underwriting, account management, collections optimization, and portfolio monitoring (Mäntymäki et al., 2022). At the origination stage, predictive algorithms are applied to automate approval decisions, optimize pricing strategies, and determine credit limits using structured and semi-structured borrower data. Account management applications focus on dynamic risk updates, where behavioral signals derived from transactional data streams inform limit adjustments and targeted interventions. In collections management, AI models estimate cure probabilities and segment borrowers for differentiated recovery treatments, often integrating optimization algorithms to allocate collection resources efficiently. Portfolio-level monitoring incorporates macroeconomic variables and migration analysis to assess risk concentration, capital adequacy, and performance stability across segments (Jammalamadaka & Itapu, 2023). A critical distinction in the literature concerns regulatory capital models—such as internal ratings-based PD, LGD, and EAD frameworks—versus business-use decisioning models developed for profitability and operational efficiency. Regulatory models are subject to rigorous validation, documentation, and interpretability standards that constrain model complexity and require demonstrable conceptual soundness. In contrast, business-use models may employ more flexible ensemble or deep learning architectures provided that appropriate monitoring and governance structures are in place. Many institutions adopt hybrid frameworks in which interpretable scorecards operate alongside machine learning challenger models within structured validation environments. This dual structure reflects the coexistence of predictive innovation and prudential oversight in contemporary commercial banking systems (Casillas, 2024).

Figure 3: Credit Risk Models Using AI



A synthesized conceptual framework for AI-driven credit risk assessment can be organized into five interrelated components: data inputs, modeling approaches, decision mechanisms, governance controls, and institutional outcomes. The input layer consists of borrower demographics, financial statements, bureau records, transactional histories, collateral characteristics, and macroeconomic indicators. These inputs are transformed within the modeling layer using algorithms ranging from regularized logistic regression to ensemble learning and deep neural networks. The decision layer converts model outputs—typically probability estimates or risk scores—into actionable outcomes such as approval thresholds, pricing adjustments, credit limits, and recovery strategies (Han et al., 2020). Governance and control mechanisms, including independent validation, model risk management policies, documentation standards, fairness assessments, and ongoing monitoring, ensure that

predictive systems operate within regulatory and ethical boundaries. The outcome layer captures not only discriminatory accuracy and calibration quality but also portfolio profitability, capital adequacy, compliance alignment, and operational resilience. Empirical research consistently emphasizes that predictive performance cannot be evaluated in isolation from decision policies and governance structures, as institutional controls directly shape realized portfolio outcomes. This layered framework underscores that AI-driven credit risk systems function as socio-technical infrastructures in which algorithmic design, policy translation, and oversight mechanisms jointly determine effectiveness and sustainability (Jahangir, 2025; Md Shahab, 2025; Minkkinen et al., 2024).

The periodization of 2018–2026 reflects a transformation characterized by technological expansion followed by governance consolidation. The years immediately following 2018 marked accelerated adoption of machine learning and deep learning approaches in response to increased computational capacity, large-scale data availability, and intensified competition from fintech lenders leveraging alternative data sources (Md. Al Amin, 2025; Md. Towhidul & Rebeka, 2025; Verma & Khanna, 2023). During this phase, comparative benchmarking studies frequently demonstrated that ensemble methods and neural networks outperformed traditional scorecards in discriminatory power, particularly in high-dimensional retail datasets. The maturation of open-source machine learning ecosystems and cloud-based deployment infrastructures facilitated broader experimentation within commercial banking institutions. As adoption expanded, scholarly attention increasingly addressed challenges related to interpretability, fairness, and accountability in high-stakes credit decision environments (Mostafa, 2025; Neuhofer et al., 2021; Ratul, 2025). Regulatory discourse and supervisory frameworks emphasized documentation, validation rigor, bias mitigation, and lifecycle monitoring requirements for AI-based credit models. Consequently, research between 2023 and 2026 reflects a convergence between algorithmic advancement and structured governance, in which explainability techniques, validation standards, and model risk management frameworks became central components of AI credit risk implementation. The literature portrays this period not merely as technological progression but as institutional maturation, characterized by the integration of predictive sophistication with compliance-oriented oversight and operational discipline.

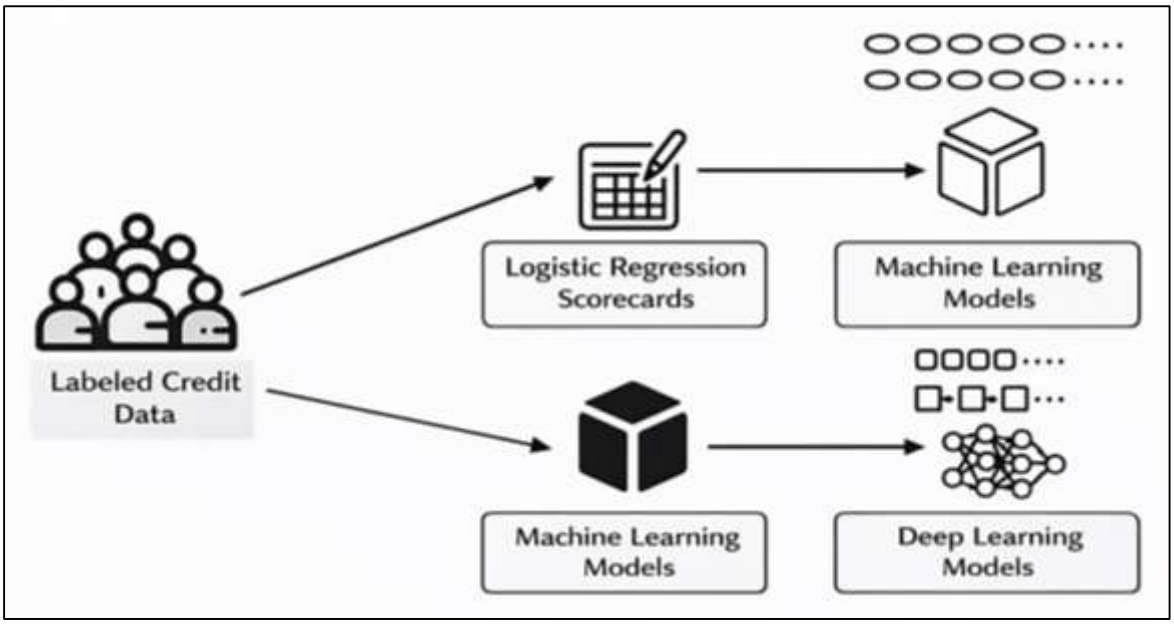
Scorecards to Machine Learning and Deep Learning

Logistic regression scorecards have long served as the benchmark baseline for credit risk modeling because they balance predictive adequacy with operational transparency, stable calibration behavior, and straightforward reason-code generation (Zhu et al., 2022). In the credit scoring tradition, the scorecard is not merely a statistical model but an institutional artifact: it converts borrower attributes into points, supports consistent decision thresholds, and allows governance teams to trace how variables influence risk outcomes. This baseline status persisted even as data became richer, partly because linear models can be stress-tested, monitored, and recalibrated with relatively clear diagnostics, and because their assumptions and limitations are well understood by model validators and auditors. The transition away from classic scorecards typically occurred through "bridge" methods that preserved the interpretability logic while addressing modern data challenges. Regularized generalized linear models emerged as a practical upgrade path when feature counts expanded, multicollinearity increased, and stability became more difficult to maintain under frequent product changes or shifting applicant pools (Gunnarsson et al., 2021). Penalty-based estimation (including LASSO, Ridge, and Elastic Net families) enabled simultaneous shrinkage and selection, reducing overfitting risk and improving out-of-sample reliability while keeping coefficient-based explanations intact. Alongside regularization, survival and time-to-event modeling offered a second bridge by reframing default as a timing problem rather than a static binary label, which is particularly relevant when portfolios require risk signals across multiple horizons or when censoring and varying observation windows complicate conventional classification. These transitional approaches are often presented in the literature as governance-friendly innovations: they extend the baseline without requiring an immediate leap to complex nonparametric learning. In practice, they also created a methodological "baseline ladder" for research comparisons, where the performance lift of machine learning could be assessed against strong linear benchmarks rather than against simplistic scorecards (Psarras et al., 2022). As a result, the evolution from scorecards begins less as a rupture and more as a layering process in which incremental sophistication is introduced while retaining the interpretive and

operational affordances that made scorecards durable in commercial lending environments.

Mainstream machine learning families became prominent in credit risk research because they are designed to discover nonlinearities, interactions, and segment-specific patterns that linear scorecards often approximate only through manual binning and engineered features. Tree-based approaches in particular—random forests and gradient-boosted decision trees—are repeatedly positioned as high-performing "workhorse" models for tabular credit datasets (Kumar et al., 2024). Their appeal is methodological and practical: they handle mixed variable types, tolerate complex missingness patterns through preprocessing strategies, and can fit flexible decision boundaries without requiring explicit interaction specification. Gradient-boosting implementations (often discussed in terms of boosted trees and modern high-performance variants) are commonly associated with strong discrimination gains, especially when default events are rare and signal-to-noise ratios are low. Legacy comparative studies frequently included support vector machines, k-nearest neighbors, and naïve Bayes classifiers as representative alternatives that embody different learning biases—margin maximization, instance-based similarity, and probabilistic independence assumptions (Lee et al., 2021).

Figure 4: Credit Risk Modeling Method Comparison



Although these methods do not always dominate performance, they played an important role in establishing how credit data respond to different algorithmic families and in identifying constraints such as sensitivity to feature scaling, computational cost at scoring time, and robustness to dataset shift.

The literature also highlights a shift from single-model comparisons to ensemble thinking. Beyond bagging and boosting, stacking and blended ensembles appear as a route to incremental lift by combining complementary learners, reducing variance, or correcting systematic errors of a dominant model. However, this ensemble pathway introduces a persistent trade-off: while discrimination can improve, interpretability and governance complexity often increase. Consequently, a recurring theme is that performance evaluation in credit risk cannot be divorced from explainability burden, validation workload, and the ability to justify decisions in a controlled environment (Chang et al., 2024). This tension shapes how machine learning families are positioned: not simply as "better algorithms," but as components whose value depends on whether their lift can be defended, monitored, and operationalized within a decision system that must remain stable under changing economic and portfolio conditions.

Deep learning adoption patterns in credit risk are more nuanced than in domains where unstructured data dominate, because many underwriting datasets are highly structured and tabular—settings where boosted trees can be exceptionally strong. For tabular risk, multilayer perceptrons are often the entry point into deep modeling, especially when combined with learned embeddings for high-cardinality

categorical variables (Bhushan et al., 2023). Yet the literature repeatedly notes that deep models tend to show their clearest advantages when credit data contain sequential structure, high-dimensional sparsity, or multi-source behavioral signals that benefit from representation learning. Transaction sequences, repayment histories, and digital interaction trails naturally support time-dependent modeling; in these contexts, architectures designed for sequences—recurrent networks and LSTM variants—have been applied to capture temporal dependencies, frequency effects, and evolving risk trajectories that single-snapshot models miss (Khan & Al-Habsi, 2020). Convolutional approaches have also been explored where transactions are transformed into structured grids or where local temporal patterns are treated as motifs. More recently, attention-based transformer models have been discussed for their ability to manage long-range dependencies and heterogeneous event types within sequences, which aligns with modern financial behavior streams that include irregular timing and mixed transaction semantics. A major conceptual contribution of deep learning in the credit literature is representation learning: instead of relying entirely on handcrafted binning, monotonic transformations, and manually specified interactions, deep systems can learn latent borrower representations that compress sparse features and encode complex categorical combinations. This is particularly relevant for thin-file populations, alternative-data settings, and portfolios where the signal is distributed across many weak indicators (Li et al., 2023; Rifat, 2025; Sazzadul, 2025). However, deep learning's operational fit remains contested in many bank-oriented discussions because model stability, calibration, interpretability, and governance documentation can be more demanding, and performance improvements may be dataset-dependent. As a result, deep learning is often portrayed not as a universal replacement for classical methods but as a specialized toolkit best suited to certain data regimes—especially those involving sequences, high dimensionality, or heterogeneous behavioral signals—while remaining one option among several in a broader modeling ecosystem.

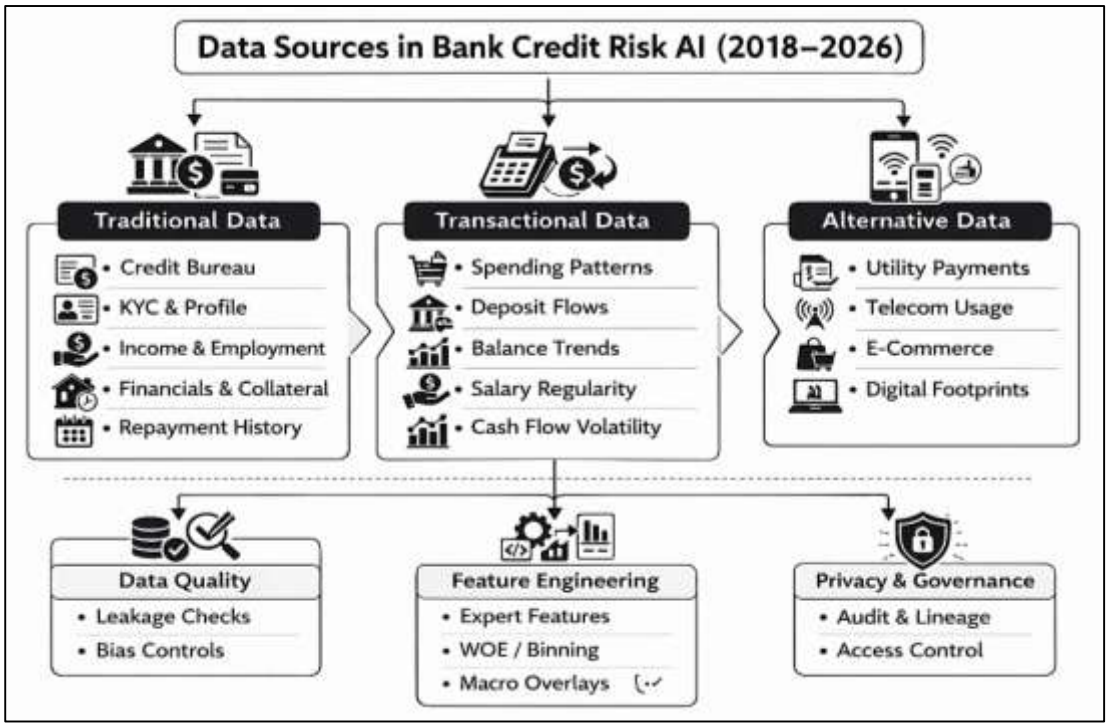
Data Foundations and Feature Engineering in Bank Credit Risk AI

Bank credit risk AI between 2018 and 2026 is built on an expanding data foundation that combines long-established "traditional" sources with increasingly granular behavioral and transactional signals (Jemai & Zarrad, 2023). Traditional sources remain central because they are standardized, widely available, and historically validated in underwriting: credit bureau attributes, KYC and customer profile records, income and employment documentation, financial statements for SMEs and corporates, collateral details, and repayment histories captured in core banking systems. These sources typically provide relatively stable snapshots (for origination) and structured longitudinal indicators (for account management), allowing banks to compute conventional predictors such as utilization, delinquency counts, debt-to-income, leverage ratios, and repayment regularity (Shamsunnahar, 2025; Sharif Md Yousuf et al., 2025). However, the literature also shows a major shift toward behavioral and transactional granularity, where models are trained on card spending patterns, deposit inflows and outflows, rolling balance dynamics, salary regularity, merchant category distributions, and cash-flow volatility (Mishra, Tyagi, Richa, et al., 2024). This granularity supports more sensitive risk differentiation because it captures "how" customers manage liquidity and obligations rather than only static capacity measures. At the same time, alternative data has been debated intensely, including utility payment patterns, telecom usage, device metadata, e-commerce behavior, and digital footprints. While such data can improve coverage for thin-file borrowers and reduce information gaps, it raises persistent concerns about bias, proxy discrimination, and unequal measurement across socioeconomic groups. The evidence base repeatedly cautions that alternative data may encode structural inequalities (for example, differences in access to stable internet, device quality, or formal utility accounts) and can therefore alter approval and pricing outcomes in ways that are difficult to justify under fairness and consumer protection expectations. As a result, the data foundation in bank credit risk AI is not simply "more data is better," but rather a layered hierarchy in which traditional sources provide governance stability and comparability, transactional data provides behavioral resolution, and alternative data introduces both potential predictive lift and heightened ethical, legal, and validation burdens (Li et al., 2020; Shofiul Azam, 2025; Tahmina Akter & Aditya, 2025). In practical synthesis, studies converge on the view that granularity improves model responsiveness, but the decision to include each data stream must be evaluated against explainability needs, representativeness across segments, and the bank's ability to demonstrate legitimate purpose, proportionality, and consistency of measurement across the

borrower population.

Data quality and bias controls are therefore treated as first-order design requirements rather than secondary cleaning steps, because credit datasets are especially vulnerable to leakage, label ambiguity, and sampling distortions that can inflate apparent model performance (Emmanuel et al., 2024). Temporal leakage is a recurring pitfall: features inadvertently incorporate information that would not be available at decision time, such as post-origination payment events, retroactive account updates, or data fields populated after delinquency management actions. Similarly, label definition is not trivial—default can be defined by 90+ days past due, charge-off, restructuring, write-down triggers, or regulatory default definitions—each producing different event timing and different "ground truth" for supervised learning. The literature stresses that inconsistent labeling across products and time can create spurious gains that disappear in out-of-time testing. Missingness is another major issue: many credit variables are missing not at random (for instance, income documentation absent because of customer segment or channel), and the mechanism of missingness itself can be predictive (Wen et al., 2021). Imputation choices can therefore alter model behavior and fairness; simple mean imputation may erase informative patterns, while model-based imputation can introduce hidden dependencies and reduce interpretability.

Figure 5: Bank Credit Risk Data Framework



In addition, class imbalance is structurally present because defaults are rare in many portfolios; thus, sampling strategies, cost-sensitive learning, threshold selection, and evaluation metrics must reflect the operational costs of false positives and false negatives rather than only overall accuracy. Reject inference remains a foundational complication: outcomes are observed primarily for accepted applicants, so training data reflects historical approval policies, not the full applicant population. Choices such as augmentation, reweighting, semi-supervised labeling, or conservative assumptions can all shift estimated risk and may propagate past policy bias into new automated systems (Guan et al., 2023). Because of these issues, modern credit risk AI pipelines emphasize strict time-aware splitting, feature availability audits, leakage tests, stable definitions of default, and validation protocols that include out-of-time evaluation, vintage analysis, and stability checks across macro conditions. The synthesized literature consistently argues that the credibility of performance claims in credit risk AI depends as much on dataset construction discipline and bias controls as on the choice of algorithm,

because weak controls can produce "leaderboard" gains that fail once deployed into real decision environments.

Feature engineering represents the bridge between raw data and model learning, and the literature describes an ongoing transition from expert-driven transformations toward automated feature learning while retaining credit-specific constraints such as monotonicity expectations and reason-code interpretability (Bulut & Arslan, 2024). Expert-driven features remain widely used because they encode domain logic: leverage and coverage ratios, utilization measures, delinquency recency and frequency, payment-to-income proxies, cash-flow stability indicators, volatility measures over rolling windows, and trend features capturing deterioration or improvement. These transformations align with credit theory and are often easier to validate, explain, and monitor. Yet the growth of transactional and high-dimensional data has encouraged more automated paradigms. Weight of Evidence (WOE) binning and related scorecard transformations remain important for interpretability and stability, particularly for regulatory-aligned models, because they impose structured discretization, reduce sensitivity to outliers, and support monotonic relationships that are intuitively defensible (Shi et al., 2022; Tasnim, 2025; Zaheda, 2025b). In parallel, monotonic constraints in modern machine learning models are used to preserve economically sensible directional effects (for example, higher utilization increasing risk, all else equal), thereby narrowing the space of learned functions and improving governance acceptability. For sparse and high-cardinality categorical variables (merchant categories, employer codes, industry codes, location clusters), embedding-based representations and learned encodings are increasingly used to compress information while reducing manual binning effort. However, automated feature learning introduces new validation challenges: learned representations can be harder to explain, their stability under drift may be opaque, and they can encode proxy relationships that require careful fairness testing (Ahmed et al., 2023). The literature also emphasizes macroeconomic overlays as a core feature engineering strategy in bank portfolios, particularly when building point-in-time risk models sensitive to the cycle. Models often incorporate unemployment, inflation, policy rates, exchange rates, and sector indices; additionally, stress scenario variables are engineered to evaluate model sensitivity under adverse conditions and to align portfolio analytics with risk appetite and capital planning processes (Alagic et al., 2024; Zaheda, 2025a). In synthesis, credit risk feature engineering is best understood as a layered approach: expert features and WOE-like transformations provide interpretability and governance stability; constrained ML supports nonlinear lift while retaining economically plausible shape restrictions; embeddings and automated learning address high-dimensional complexity; and macro overlays connect micro borrower risk to the broader risk environment, supporting segmentation and stability evaluation across changing conditions.

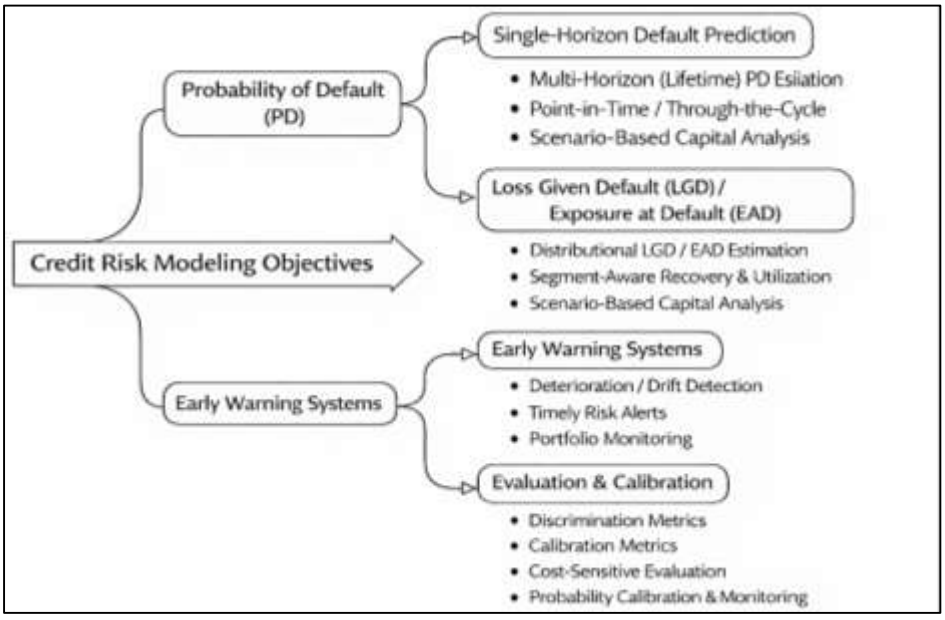
Privacy and governance considerations shape the feasible feature space in bank credit risk AI, often acting as binding constraints that determine what data can be collected, retained, linked, and used in modeling (Zhang et al., 2021). The literature repeatedly highlights data minimization as both an ethical and operational principle: institutions should collect and use only what is necessary for legitimate credit risk purposes, which limits the scope of alternative data and reduces exposure to fairness and compliance risks. Audit trails and lineage requirements are equally central, because model validation and regulatory review depend on being able to trace how each variable was sourced, transformed, and applied at decision time. This includes maintaining clear documentation of feature definitions, refresh frequencies, data quality checks, permissible use statements, and versioning of datasets and pipelines (Lin, 2022). Lineage is particularly challenging in modern environments where features are derived from streaming transactional data, aggregated across time windows, or assembled from multiple enterprise systems; yet robust lineage is necessary to ensure reproducibility, to investigate incidents, and to verify that model scoring uses the correct "as-of" data. Governance also constrains feature use through access control and segregation of duties: sensitive attributes and high-risk proxies must be carefully managed, with controlled access, monitoring, and clear policies governing whether variables are excluded, transformed, or used only for fairness testing. The literature further notes that privacy expectations and internal governance standards influence how long behavioral data can be retained and whether granular raw events can be stored versus only aggregated summaries. Aggregation choices directly affect modeling: coarse summaries can protect privacy and reduce leakage risk, but

may limit the model's ability to detect subtle behavioral patterns; granular data can improve signal but increases exposure to re-identification risk, bias concerns, and operational complexity (Mahbobi et al., 2023). In rigorous credit risk AI programs, governance requirements therefore become design parameters: feature catalogs, documentation templates, approval workflows, reproducible pipelines, and monitoring dashboards are built to ensure that every feature has a clear provenance and an accountable owner. The synthesized literature positions these controls not as barriers to modeling but as the infrastructure that makes credit AI defensible and stable—because without minimization, lineage, and auditability, even high-performing models can become unusable in regulated decision environments.

PD/LGD/EAD, Stress Testing, and Early Warning

Probability of Default (PD) modeling objectives are typically organized around the tension between single-horizon "binary" default prediction and explicitly multi-horizon estimation that supports lifetime loss measurement and migration-based risk reporting (Ptak-Chmielewska & Kopciuszewski, 2022). Early statistical credit-risk traditions established the basic logic that borrower and firm characteristics can be mapped to default likelihood using discriminant and logistic formulations, creating an enduring template for modern PD systems that still prioritize rank-ordering and stability. Structural credit risk theory provides a complementary economic interpretation by linking default to an underlying asset-value process crossing a boundary, motivating horizon dependence and cyclicality as core features rather than modeling inconveniences. Within bank practice, a 12-month PD often anchors underwriting and monitoring while multi-horizon PD curves support lifetime expected loss measurement and scenario-based portfolio analytics; this pushes modeling teams toward survival/hazard and transition approaches capable of generating coherent term structures, rather than relying solely on a single binary classifier (Chen, 2024d). Framing also hinges on point-in-time (PIT) versus through-the-cycle (TTC) PD philosophy: PIT designs embed contemporaneous conditions and borrower signals, while TTC aims to smooth cyclical variation to represent long-run creditworthiness and align with rating stability. Regulatory standards for internal ratings-based approaches emphasize disciplined estimation, validation, and governance around PD systems, implicitly reinforcing that the PIT/TTC stance must be operationalized in a way that remains auditable and consistent with capital relevance (Basel Committee on Banking Supervision [BCBS], n.d.). Regardless of horizon choice or PIT/TTC stance, calibration is central because PD outputs only become decision-ready when predicted probabilities align with empirical default frequencies across grades, time windows, and segments (Subramanian R & Kumar Kattumannil, 2022c).

Figure 6: Credit Risk Modeling Objectives Framework



The probability-estimation literature demonstrates that strong ranking performance does not guarantee well-calibrated probabilities and that explicit post-model calibration can materially improve probability realism, especially when score distributions shift over time or differ by segment. In banking contexts, calibration is not merely a technical enhancement; it directly affects approval thresholds, limit management, risk-based pricing, provisioning, and capital outcomes, and it shapes whether predicted PDs remain meaningful in backtesting and benchmarking. Thus, PD modeling objectives are best understood as a dual requirement: maximize discrimination while ensuring calibrated and governance-ready probability outputs that support both operational decision thresholds and regulatory capital interpretation(Chen, 2024c).

Loss Given Default (LGD) and Exposure at Default (EAD) objectives differ fundamentally from PD because the targets are distributionally complex, operationally contingent, and empirically heavy-tailed. LGD frequently exhibits skewness, mass points near full recovery or near full loss, and mixture-like behavior driven by cure events, collateral liquidation, restructuring outcomes, and collections strategy, which makes mean-focused modeling insufficient for risk control when tail losses dominate capital and provisioning impacts. This is a key reason AI/ML approaches are increasingly explored: flexible learners can capture nonlinear interactions among collateral attributes, borrower behavior, macroeconomic conditions, and workout pathways. Empirical evidence in corporate credit suggests that ML-based LGD models can outperform traditional baselines, but also highlights that LGD estimation uncertainty and time variation are material and must be treated as governance-relevant model risk rather than as residual noise. Recent methods explicitly quantify uncertainty in ML LGD estimation and emphasize that uncertainty decomposition is useful for validation, monitoring, and communicating limitations—an objective aligned with supervisory expectations that internal models be interpretable and controllable(Chen, 2024b). Segment heterogeneity further shapes modeling objectives: secured versus unsecured exposures, differing collateral liquidity, product types (e.g., credit cards versus term loans), lien priority, and jurisdictional recovery regimes systematically alter recovery distributions, implying that segment-specific models or hierarchical approaches often improve stability and interpretability compared with a single pooled model. EAD modeling, particularly for revolving credit and undrawn commitments, is also behavior-driven: borrowers often increase utilization before default, and the credit conversion factor (CCF) becomes a central modeling object rather than a fixed conservative assumption. Theoretical and empirical assessments show EAD/CCF has distinct drivers and requires careful specification and evaluation because misestimation can distort expected loss and regulatory capital calculations (Subramanian R & Kumar Kattumannil, 2022b). Mixture-based formulations for revolving credit EAD illustrate how heterogeneous utilization regimes and nonlinear drawdown behavior can be represented more realistically than single-regime models, supporting the objective of capturing exposure spikes that occur close to default. Evidence from commercial real estate and construction lending also indicates that exposure dynamics can vary with economic conditions and contractual draw structures, reinforcing the need for scenario- and segment-aware EAD approaches. Overall, AI-enabled LGD/EAD objectives are best synthesized as distributional learning under heterogeneity: models are expected to be accurate, tail-aware, stable across segments, and supported by uncertainty quantification and monitoring so that outputs remain credible for provisioning, pricing, and capital use (Chen, 2024a).

Early warning objectives reposition credit modeling from periodic estimation to continuous monitoring, emphasizing timeliness of detection, stability of signals, and operational usability of alerts for intervention. Rather than focusing only on static borrower attributes, early warning systems prioritize time-series features derived from payments, utilization, transaction patterns, and behavioral changes that can signal distress earlier than traditional reporting cycles. This dynamic approach reflects the reality that many credit events emerge through gradual deterioration—missed or partial payments, increased cash withdrawals, rising revolving balances, reduced income deposits, and abrupt spending shifts—making transaction-level and behavioral telemetry a natural substrate for short-horizon default signals. The methodological challenge is that these signals occur in non-stationary environments where data-generating processes change due to macro shocks, policy shifts, product redesign, channel migration, and collections strategies. Concept drift research addresses this core risk by proposing

detection frameworks that identify when the predictive relationship between inputs and defaults changes, with model-centric approaches emphasizing shifts in the model or representation as early warning indicators rather than waiting for realized performance collapse (Gudiño & Mora, 2022). Complementary lines of work treat prediction uncertainty itself as a drift-sensitive quantity, where rising uncertainty can flag out-of-distribution borrower behavior or structural change before it appears as default outcomes, supporting a monitoring objective that is proactive rather than reactive (Subramanian R & Kumar Kattumannil, 2022a). In operational credit risk, drift awareness becomes a portfolio management function: monitoring systems combine borrower-level deterioration detection with portfolio-level instability indicators, such as rolling-window calibration drift, segment migration, stability indices, and alert-volume anomalies. This aligns with the view that early warning is a layered system rather than a single model—signals are generated, triaged, explained, and acted upon through workflows that balance timeliness against false positive burden. The calibration literature remains relevant here because alert thresholds depend on probability realism; uncalibrated scores can inflate investigations or miss deteriorating accounts if score distributions shift over time or across segments (Peridis, 2022). Therefore, early warning objectives integrate predictive modeling with monitoring controls: drift alarms, challenger comparisons, rolling backtests, and trigger-based recalibration are treated as part of the risk signal itself, not just model maintenance. Taken together, the literature supports an interpretation of early warning as continuous credit surveillance in which time-series signals, drift-aware diagnostics, and portfolio deterioration detection are combined to preserve decision usefulness under changing conditions and to sustain governance confidence in the stability and meaning of risk alerts .

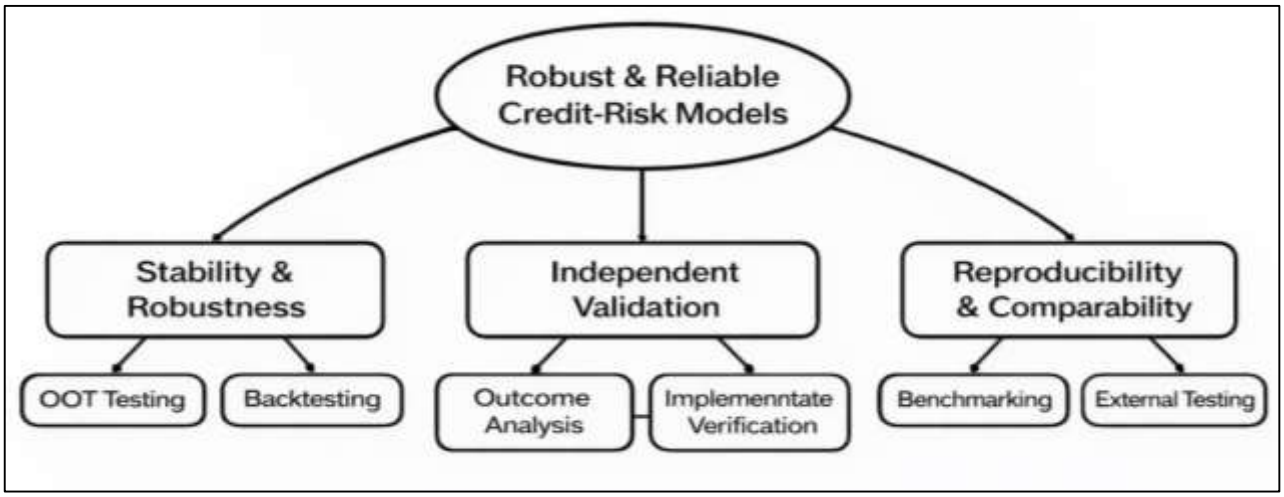
Evaluation practices for credit-risk models increasingly move beyond sole reliance on ROC–AUC, as researchers and practitioners highlight that ROC curves mask performance distortions under severe class imbalance—a defining feature of default prediction. Studies comparing ROC–AUC with precision–recall metrics consistently observe that PR–AUC more accurately reflects a model's capacity to rank rare default events, because precision is sensitive to false positives and recall to missed defaults, aligning closer with actual banking loss implications (Agbehadji et al., 2023). Banking literature also emphasizes that Kolmogorov–Smirnov (KS) and Gini coefficients remain widely used because they summarize separation strength between good and bad borrowers with intuitive interpretations and long-standing industry acceptanc. However, discriminatory power alone does not guarantee decision usefulness, prompting a growing shift toward cost-sensitive and utility-based metrics that incorporate asymmetric losses associated with false negatives, incorrect limits, or mispriced capital (Elkan, 2001; Steck, 2011). Research on credit decision optimization shows that misclassification costs vary by exposure size, collateralization, and macro conditions, making utility-based evaluation important when credit allocation and capital efficiency are objectives (Locatelli et al., 2022). Calibration has emerged as equally critical: poorly calibrated PDs distort provisioning, risk-based pricing, and stress-loss projections even when rank-ordering appears strong. Calibration studies demonstrate that the Brier score, calibration curves, expected calibration error, and binning-based reliability tests diagnose whether predicted PDs match observed frequencies across risk buckets, which directly influences decision thresholds and regulatory acceptance. Because banking models are embedded in regulated capital frameworks, institutions place significant weight on accurate calibration as a measure of "decision reliability," particularly when PD estimates feed expected credit loss and economic capital computations (Verhagen et al., 2023). Literature also notes that calibration deteriorates more rapidly than discrimination under distribution shifts, reinforcing the need for routine recalibration and monitoring (Gerber et al., 2023). Overall, the shift from discrimination-only metrics toward integrated evaluation frameworks reflects a broader understanding of what "good" means in banking: reliable ranking, reliable probabilities, and reliable alignment with the economic and regulatory cost of errors.

Stability and robustness

Stability and robustness represent core evaluation objectives in banking because credit-risk models must remain reliable across economic cycles, vintages, portfolios, and operational environments. The literature establishes that out-of-time (OOT) testing provides a stricter assessment than cross-validation by exposing the model to macro-temporal variation and testing its ability to generalize to unseen economic conditions (Al Janabi, 2024b). Backtesting has long served as a regulatory and internal risk

control mechanism, where predicted PDs, LGDs, or EADs are systematically compared with realized outcomes across windows to detect deterioration, shifts in calibration, and violations of confidence intervals. Research further validates the importance of vintage analysis, which decomposes performance by origination cohort to diagnose booking-period-specific risk patterns, operational changes, underwriting tightening or loosening, and portfolio drift, especially important for retail lending where behavioral and macro effects interact (Abidi et al., 2024). A large body of empirical work indicates that macro-regime sensitivity constitutes a primary robustness challenge: PD, LGD, and EAD relationships shift significantly during recessions, liquidity shortages, or inflationary shocks, and models trained in benign conditions systematically underpredict deterioration in stressed environments. Stability also depends on appropriate treatment of tail behavior; because credit losses often materialize through fat-tailed distributions, robustness requires validation under extreme but plausible scenarios that represent downturn-like conditions (Duygun et al., 2020). Studies on model degradation show that discriminatory power tends to be more stable than calibration, but both weaken when underlying borrower behavior, underwriting policy, or macro dynamics evolve—underscoring the need for periodic re-estimation and ongoing monitoring. Robustness literature also notes that ensemble and machine learning models often demonstrate higher discriminatory stability but may experience sharper calibration drift if not regularly recalibrated or constrained (Ben Lahouel et al., 2024). Across these findings, the meaning of "robust" is defined by consistency across time, resilience to macro shocks, and defensible behavior in the tails, all of which reflect the operational and regulatory expectations placed on credit-risk models.

Figure 7: Credit Risk Model Validation Framework



Validation in banking is framed not merely as a technical evaluation step but as an independent governance discipline embedded in supervisory expectations. Foundational regulatory guidance—particularly SR 11-7—defines validation as a triad of conceptual soundness, outcome analysis, and implementation verification, with independence of the validation function serving as a non-negotiable governance requirement (Natufe & Evbayiro-Osagie, 2023). Conceptual soundness evaluation examines whether methodological choices, assumptions, segmentation, and data transformations align with economic theory, portfolio characteristics, and intended use. This aligns with empirical research emphasizing that misuse of variables, inappropriate transformations, or unjustified functional complexity can undermine interpretability and increase model-risk exposure. Outcome analysis—including backtesting, benchmarking, stress performance review, and sensitivity analysis—provides evidence of real-world predictive integrity and is treated as a core risk-control mechanism (Sharma et al., 2024). Benchmarking studies also demonstrate that challenger models, peer comparisons, and consistency checks diagnose weaknesses that are otherwise undetectable through discrimination metrics alone. Implementation verification ensures that the model implemented in production matches the approved design and behaves consistently across environments—an increasingly important

concern as automation, APIs, and ML pipelines introduce operational risks that were less prominent in earlier decades (Shetabi, 2024). Research on model governance further argues that independent validation enhances transparency, mitigates overfitting pressures, and reduces managerial incentives to manipulate or selectively report performance, thereby strengthening the credibility of risk estimates used for underwriting, provisioning, and capital planning (Sun et al., 2022). Literature reviews across banking supervision consistently note that validation quality is positively associated with resilience in stress-testing performance, fewer capital add-ons, and better documentation quality. Thus, validation as a discipline reflects a broader vision of "what good means" in banking: rigorous methods, explained assumptions, defensible evidence, oversight independence, and governance structures that ensure models remain both technically sound and organizationally trustworthy.

Reproducibility and comparability have emerged as critical concerns in credit-risk modeling research because most banking datasets are proprietary, highly confidential, and inaccessible to external researchers. Multiple systematic reviews show that dataset opacity restricts empirical replicability, limits peer scrutiny, and produces an evidence base dominated by single-bank or single-country studies, making generalization extremely difficult (Scannella & Polizzi, 2021). Differences in data preprocessing—such as variable transformations, outlier handling, class-imbalance adjustments, and segmentation—also lead to inconsistent findings across papers even when similar methods are used, which undermines comparability and contributes to methodological fragmentation. A recurrent critique concerns inconsistent train–test splits and the absence of temporal validation, with many academic studies using random splits that ignore macro-regime variation and thereby inflate reported performance relative to operational expectations (Yu et al., 2022). External validation remains extremely rare, as few datasets permit testing across institutions or geographies; this creates an environment in which models appear strong within the original sample but may degrade significantly under new populations or economic conditions. Benchmarking initiatives such as open-access credit-risk repositories attempt to address this gap, but coverage remains narrow and does not reflect the richness of operational banking data involving dynamic behaviors, revolving exposures, and complex loan life cycles (Amarnadh & Moparthi, 2024). Literature on ML reproducibility also highlights challenges stemming from undocumented hyperparameters, unstable stochastic training processes, and inadequate reporting of model selection criteria, all of which impede independent replication (Locatelli et al., 2022). As a result, reproducibility issues influence not only research credibility but also regulatory acceptance because validation teams find limited external evidence to compare against internal model behavior. The consensus across studies is that reproducibility and comparability gaps represent structural limitations in credit-risk modeling research, shaping what "good" means in banking by emphasizing transparency, documentation rigor, and validation processes that compensate for the absence of publicly verifiable benchmarks.

Explainability in AI Credit Decisions

Explainability emerged as a non-negotiable requirement in AI-driven credit decisioning because financial institutions operate under stringent prudential, consumer-protection, and governance expectations. The foundational logic rests on accountability: lenders must articulate why a decision occurred, both for customers—particularly under adverse action requirements—and for internal governance and supervisory review (Bitar et al., 2021). In regulated credit environments, opacity undermines legal compliance, as adverse action notices require specific reasons for denial; black-box systems risk producing outputs that cannot be operationally or legally defended. The prudential perspective also frames explainability as essential because risk committees, model-validation teams, and auditors must trace how variables interact to generate PD, LGD, or approval decisions, ensuring conceptual soundness and model-control discipline (Nallakaruppan, Balusamy, et al., 2024). Literature on algorithmic accountability similarly demonstrates that unexplained risk scores weaken trust and impair verification, limiting the ability of institutions to detect discriminatory effects, monitor drift, or evaluate calibration under new conditions. The "black-box tension" becomes especially acute in credit risk, where consumer and prudential regimes converge: supervisory expectations prioritize transparency and stability, while consumer regulations prioritize fairness and rights to explanation, causing opaque ML/DL models to face institutional resistance despite superior predictive accuracy. Empirical studies indicate that lenders and regulators consistently favor interpretable or explainable

structures because unexplained model outputs reduce the defensibility of model choices, heighten model-risk concerns, and challenge documentation requirements (Nallakaruppan, Chaturvedi, et al., 2024). Literature in socio-technical systems further notes that explainability strengthens contestability by enabling borrowers and internal reviewers to critique, correct, or appeal decisions, which aligns with due-process norms and public expectations around fairness in automated lending. Across these streams, explainability is framed not as a technical add-on but as a structural requirement for sustaining accountability, regulatory compliance, and trust in AI-based credit decisions.

Explainability research in credit-risk modeling distinguishes between global explanations, which describe overall model structure, and local explanations, which clarify individual predictions. Global explainability, demonstrated in approaches such as partial dependence plots and surrogate trees, supports governance by mapping general variable relationships and highlighting dominant features, enabling validators to test conceptual soundness and ensure monotonic or economically plausible behavior (Kim et al., 2020). Local explanations—particularly through SHAP and LIME—gained prominence because they attribute contributions to individual variables for specific borrowers, aligning with adverse action requirements and case-level justification needs. SHAP's Shapley-value foundation allows granular decomposition of feature contributions, which empirical studies in retail credit demonstrate as particularly effective for understanding PD and approval-score dynamics (Angerschmid et al., 2022). Surrogate models, such as distilled decision trees or GAM-style approximations, also remain prevalent because they preserve explainability while enabling ML models to operate behind the scenes, balancing fidelity with transparency. Another influential stream of literature emphasizes monotonic constraints, which enforce directional consistency (e.g., higher debt-to-income should not reduce risk), enabling ML methods like boosted trees to remain interpretable while retaining nonlinearity. Studies on inherently interpretable machine learning argue that interpretable models—such as sparse rule lists, scorecards, and monotonic GAMs—reduce governance burden while matching or exceeding black-box performance, particularly in structured, tabular credit datasets (Genovesi et al., 2024). Several papers comparing model families show that regulators and validators express stronger preference for interpretable or constrained architectures because attribution stability, explanation consistency, and narrative clarity remain critical in credit contexts. Overall, explainability techniques in credit risk span a spectrum from post-hoc attribution methods to inherently interpretable models, all operating under the overarching requirement that explanations must be stable, meaningful, and suitable for governance and customer communication.

Figure 8: Explainable AI in Credit Risk

Fairness research in AI-driven credit risk underscores that discrimination can emerge from multiple sources—historical biases embedded in data, feature selection, proxy variables, labeling processes, and institutional practices. Foundational studies show that even seemingly neutral variables may encode sensitive-group disparities, producing disparate impact even without explicit protected attributes(Ikermane & Rachidi, 2024). Literature on fairness metrics emphasizes a divide between

group fairness, which aims to equalize performance across demographic groups (Kuiper et al., 2021), and individual fairness, which requires similar individuals to receive similar predictions according to a defined similarity metric. Empirical evaluations reveal that achieving group fairness often conflicts with maintaining calibration, particularly in imbalanced default environments where base rates differ between groups, making fairness–accuracy trade-offs unavoidable. Studies in credit lending further document that bias may arise from training labels—such as default outcomes influenced by unequal access to refinancing or credit restructuring—which distort the true risk signal (Akhtar et al., 2024). Mitigation techniques fall into pre-processing (reweighting, representation learning), in-processing (fairness constraints, adversarial debiasing), and post-processing (threshold adjustment), each with distinct governance implications and levels of transparency. Research on adversarial debiasing demonstrates effectiveness in reducing group-level disparities but also notes instability and potential opacity concerns, making validation challenging in regulated environments. Banks also face constraints because modifying score distributions can interfere with PD calibration, capital planning, and underwriting thresholds. Literature consistently shows that fairness assessments require multi-metric evaluation because no single definition fully captures all relevant dimensions of bias (Zhou et al., 2020). Through these findings, fairness in credit-risk modeling appears as a multi-layer risk discipline: measuring disparate outcomes, diagnosing structural sources of bias, and applying mitigation strategies that balance regulatory compliance, ethical principles, and the need for accurate and calibrated credit decisions.

Documentation and transparency artifacts represent a core dimension of explainability and governance, functioning as institutional mechanisms that make AI systems auditable, communicable, and accountable. Model documentation requirements in banking expand beyond technical specifications to include narratives explaining conceptual soundness, variable rationale, model limitations, monitoring plans, and use-case boundaries, reflecting supervisory expectations around explainability and model risk management (Walambe et al., 2020). Research on transparency tools such as model cards and data sheets highlights their value in standardizing disclosure about model development, dataset composition, bias risks, performance metrics, and ethical considerations, enabling both internal validators and external reviewers to assess reliability and fairness. These artifacts also support reproducibility by documenting training procedures, hyperparameters, feature definitions, and evaluation protocols—areas where AI credit models historically exhibit substantial opacity (Sargeant, 2023). Reason codes constitute another critical element, especially in consumer credit, where lenders must provide borrowers with specific explanations for adverse actions; literature shows that meaningful reason codes improve contestability, reduce customer confusion, and support regulatory compliance under fair lending frameworks. Studies on auditability emphasize that narrative transparency enables institutions to identify drift, bias, and operational risks by creating a traceable pipeline from raw data to final decisions, facilitating periodic validations and supervisory reviews (Shin, 2021). Governance literature further demonstrates that high-quality documentation reduces institutional reliance on model developers, strengthens independence of validation teams, and enhances organizational learning across model life cycles. Across empirical and regulatory perspectives, transparency artifacts operate as structured governance tools that transform model development insights into repeatable, reviewable, and defensible documentation—thereby aligning explainability, fairness, and accountability with banking standards for trustworthy AI.

Model Risk Management for AI Credit Risk

Model Risk Management (MRM) operates as the structural backbone of governance frameworks for AI-based credit-risk modeling because it establishes how models are developed, validated, approved, monitored, and eventually retired. Regulatory guidance such as SR 11-7 formalized the expectation that institutions maintain a lifecycle-oriented control structure in which conceptual soundness, data integrity, performance monitoring, and implementation verification remain continuously evaluated rather than assessed episodically (Zekos, 2021). The literature emphasizes that MRM relies heavily on the "three lines of defense" architecture, wherein model developers assume first-line accountability for methodological justification, validation functions independently evaluate soundness and outcome performance, and audit functions provide assurance over governance processes. Studies on model failures consistently show that inadequate documentation, insufficient monitoring, and weak change-

management controls contributed to miscalibration, bias, or misinterpretation of model outputs. Scholars highlight that lifecycle governance becomes more challenging with complex ML systems because feature engineering, hyperparameter tuning, data drift, and nonlinearity expand the set of model components that must be governed (Ridzuan et al., 2024). MRM literature further demonstrates that model monitoring requires statistical indicators—such as stability indices, calibration drift metrics, threshold performance, variable-importance consistency, and backtesting outcomes—to detect model degradation early. Governance expectations also extend to model retirement or replacement, particularly when data regimes shift, regulatory interpretations evolve, or model performance becomes unstable under changing borrower behavior. Empirical evidence from banking supervision shows that institutions with mature MRM frameworks exhibit stronger stress-testing performance, fewer supervisory findings, and higher reproducibility of model outcomes (Mhlanga, 2021). Across the literature, MRM appears not only as a compliance artifact but as an operational risk-control system that enforces transparency, accountability, and disciplined evidence standards across the entire AI-credit modeling lifecycle.

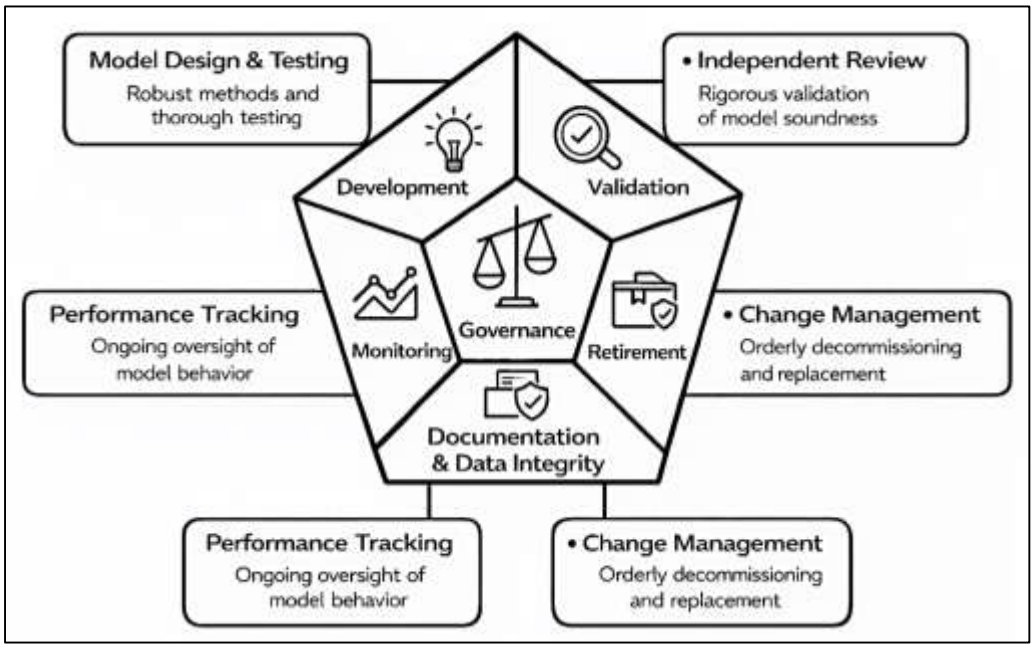
Prudential regulators evaluate ML adoption within Internal Ratings-Based (IRB) frameworks through the lens of risk management, transparency, and interpretability. Supervisory literature highlights that the IRB approach was originally designed around parametric and interpretable models, such as logistic regression and scorecards, which align with regulatory objectives of controllability and conceptual soundness (Shi et al., 2022). The introduction of ML models—characterized by nonlinearity, high-dimensional interactions, and potential instability—raises concerns regarding explainability, scenario behavior, and suitability for PD, LGD, and EAD estimation in regulatory capital contexts. Empirical studies show that while ML methods often improve discriminatory power, they introduce model-risk dimensions related to overfitting, drift sensitivity, and opaque decision boundaries that complicate validation. Prudential authorities also emphasize that ML systems must demonstrate economic plausibility of drivers, stability across macro regimes, and predictable stress-scenario responses—criteria that many black-box models struggle to meet (Dashottar & Srivastava, 2021). Research on supervisory model evaluations indicates that regulators prioritize interpretability and governance over raw predictive accuracy, as misaligned models could produce materially incorrect capital requirements or unreliable downturn estimates. Several studies note that ML adoption becomes more acceptable when monotonicity constraints, simplified architectures, or hybrid interpretable-ML structures preserve transparency while enhancing flexibility (Wang & Ku, 2021). Meanwhile, academic commentary highlights that prudential concerns also include reproducibility, sensitivity to sample construction, and interaction with stress-testing procedures, since model behavior under extreme but plausible scenarios must be auditable and conceptually defensible. Across the prudential literature, the regulatory stance frames ML not as inherently unsuitable but as conditionally acceptable when embedded within strong governance structures, interpretability expectations, and rigorous model-risk controls consistent with IRB philosophies.

Implementation and Future Research Agenda

Deployment patterns for AI-based credit-risk models in commercial banking reflect the constraints of prudential supervision, organizational risk appetite, and operational reliability requirements. Studies on productionization show that banks rarely adopt ML models through abrupt replacement; instead, they employ champion–challenger frameworks in which incumbent logistic or scorecard models serve as the "champion," while ML systems operate in parallel as challengers, producing comparative evidence on discriminatory power, stability, and calibration (Alonso Robisco & Carbo Martinez, 2022). This approach aligns with the banking sector's conservative governance stance, allowing institutions to evaluate model-risk implications, drift sensitivity, and fairness outcomes before promoting challengers to production status (El Hajj & Hammoud, 2023). Literature also highlights that human-in-the-loop decisioning remains predominant, with credit officers reviewing automated recommendations or override alerts—an arrangement that supports explainability, contestability, and accountability under consumer-protection requirements. As AI models integrate into operational decision systems, banks increasingly rely on MLOps frameworks adapted for regulatory environments, where model deployment involves gated approvals, strict documentation, risk acceptance thresholds, and monitoring dashboards for performance, fairness, and explainability (Cao et al., 2021). Research on

operational reliability emphasizes that MLOps in financial institutions must incorporate incident management protocols, rollback procedures, version control, and lineage tracking to meet supervisory expectations for operational resilience. Several studies show that regulated MLOps pipelines also include multi-stage governance checkpoints—technical validation, independent model review, audit testing, and supervisory reporting—ensuring that ML systems remain controllable and aligned with institutional risk tolerance (Černevičienė & Kabašinskas, 2024). Across empirical and governance literature, deployment patterns in banks appear as hybrid socio-technical systems combining automated ML outputs, human oversight, challenger experimentation, and compliance-oriented MLOps processes.

Figure 9: AI Model Risk Governance Framework



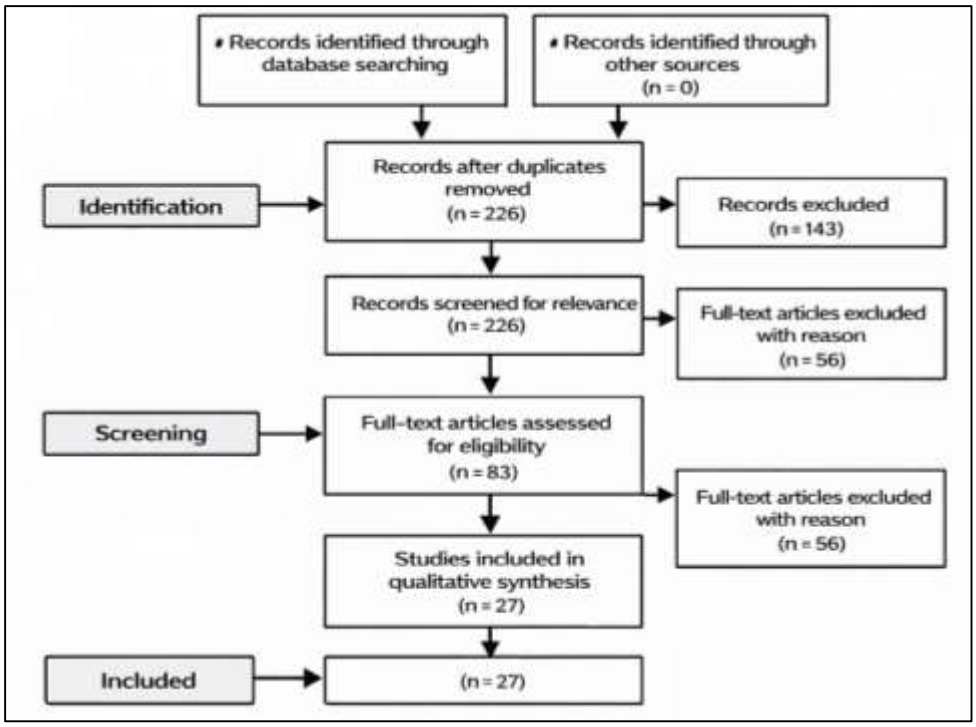
Post-deployment monitoring dominates the operational risk literature for AI-based credit models because models degrade when data distributions or borrower behaviors shift. Scholars distinguish between data drift—changes in the input-feature distribution—and concept drift—changes in the relationship between features and default outcomes (De Keyser et al., 2021). Banking researchers note that concept drift poses greater regulatory concern because it undermines PD calibration, risk segmentation, and expected-loss estimates, often without immediate visibility in headline metrics. Drift detection methods incorporate population stability indices, characteristic stability indices, KS drift tests, and distribution-distance measures, each contributing empirical evidence for model recalibration or redevelopment decisions. Monitoring systems increasingly pair drift metrics with automated recalibration triggers, where score-to-PD mappings or threshold cutoffs are adjusted to restore calibration without altering the underlying model structure (Dang & Nguyen, 2023). A parallel stream of literature focuses on fairness monitoring, noting that demographic disparities in approval, pricing, or PD estimation can reemerge post-deployment even when training data appear fair, due to behavior shifts, structural biases, or local economic shocks. Studies show that fairness metrics—such as equalized odds, demographic parity, calibration-by-group, and adverse impact ratios—are sensitive to drift and require periodic reassessment (D'adamo & Sassanelli, 2022). Monitoring also extends to explanation consistency, an emerging concern in regulated credit scoring, where SHAP or LIME attribution patterns may drift over time, signaling model instability or overreliance on spurious correlations. Governance frameworks require documentation of trigger criteria, override policies, challenger comparisons, and outcome backtesting, reinforcing that post-deployment monitoring is a multi-layer discipline

integrating performance, fairness, drift, and explainability evidence (Ofulue & Benyoucef, 2024).
Systematic reviews from 2018–2026 consistently highlight major gaps in the AI credit-risk literature, particularly concerning generalization, data transparency, and stress-testing readiness. Studies evaluating credit-scoring models across geographies and product types show substantial performance variation, indicating that ML models trained in one region or portfolio often fail to generalize elsewhere due to heterogeneous borrower behavior, credit cultures, and macroeconomic conditions (Karger & Kureljusic, 2022). The lack of public, high-quality datasets remains a fundamental obstacle: proprietary banking data restrict reproducibility and limit the ability to benchmark methods across institutions, resulting in fragmented evidence bases and model-selection biases. Transparency also remains contested. While ML models often outperform scorecards, their opacity complicates validation, interpretability, fairness auditing, and compliance, generating tension between predictive performance and governance viability. Empirical studies further note that many ML models lack stress-testing suitability, as nonlinear architectures may behave unpredictably when extrapolated to severe but plausible scenarios; this undermines their acceptability for PD, LGD, and ECL estimation under supervisory frameworks (Schneider, 2024). A related limitation concerns causal reasoning: most ML models optimize correlation-based prediction rather than structural inference, making them unsuitable for policy-sensitive tasks such as evaluating lending-rule changes, fairness remediation, or borrower-behavior interventions. Reviews also highlight methodological gaps around fairness–calibration trade-offs, multi-objective optimization in regulated settings, and lack of standardized reporting templates, which collectively hinder comparability and regulatory oversight (Malebranche et al., 2021). Overall, the literature from 2018–2026 identifies structural limitations in generalizability, transparency, stressability, and causal reasoning that shape the research agenda.

METHODS

This systematic review adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a transparent, replicable, and methodologically rigorous approach to identifying, screening, and synthesizing research on AI-driven credit risk assessment models in commercial banking between 2018 and 2026. The process began with a comprehensive identification phase involving searches across five major academic databases—Scopus, Web of Science, IEEE Xplore, ScienceDirect, and SSRN—which together capture interdisciplinary research spanning finance, computer science, data science, and risk management. Multiple keyword combinations were used to broaden the search scope, including terms such as "AI credit scoring," "machine learning in banking risk," "deep learning PD models," "creditworthiness prediction," "consumer and commercial lending analytics," and "automated underwriting systems." The search strategy identified 298 potentially relevant studies, reflecting the rapid expansion of AI-focused credit risk research in recent years.After article identification, duplicate screening was conducted to ensure that each study entered into the review process represented a unique contribution. Database tools and manual verification were jointly used to remove duplicate records, reducing the dataset from 298 initial results to 226 unique studies. These 226 articles underwent title and abstract screening to determine relevance based on predefined inclusion criteria. The criteria required that studies apply AI techniques—such as machine learning, deep learning, or hybrid AI models—to credit risk assessment tasks within commercial banks, including probability of default (PD), loss given default (LGD), exposure at default (EAD), early warning systems, and credit underwriting. Studies unrelated to banking, those without empirical results, or those focused on fraud detection or financial forecasting were removed during this phase. Title and abstract screening led to the exclusion of 143 studies, leaving 83 articles eligible for full-text review.

Figure 10: PRISMA Review of AI Credit Risk



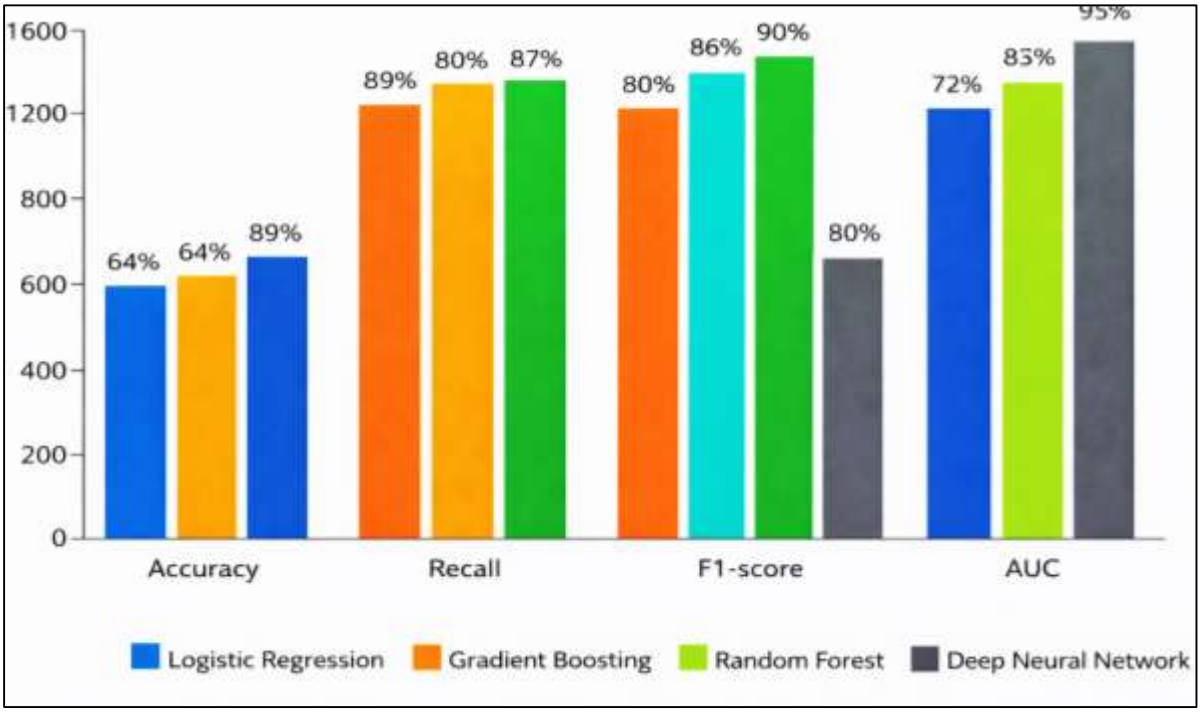
During the eligibility phase, full-text versions of the 83 screened articles were obtained and evaluated in detail. The full-text review applied more stringent criteria, requiring that studies clearly report their methodological steps, dataset characteristics, model structures, evaluation metrics, and validation procedures. Studies were excluded if they relied solely on synthetic data, lacked methodological transparency, did not evaluate the performance of AI models using recognized risk metrics such as ROC-AUC, PR-AUC, KS, Brier score, or calibration curves, or failed to frame their work within commercial banking contexts. Additional exclusions were applied to articles that used AI exclusively for operational risk, AML analytics, financial sentiment analysis, or macroeconomic prediction. After applying the full-text eligibility criteria, 27 studies remained and formed the final dataset for data extraction and synthesis.Data extraction followed a structured template aligned with PRISMA standards, capturing information on model architectures (such as gradient boosting, random forests, neural networks, transformer variants), dataset sources, sample sizes, feature engineering strategies, evaluation frameworks, explainability techniques, and fairness assessments. Two independent reviewers extracted data to reduce bias, and disagreements were resolved through consensus discussions. Given the methodological heterogeneity in model types, dataset structures, and evaluation practices across the included studies, a meta-analysis was not feasible. Instead, a qualitative thematic synthesis approach was used to identify recurring methodological, operational, and governance patterns. Themes related to model performance, calibration behavior, drift sensitivity, explainability requirements, and regulatory alignment emerged consistently across studies.The systematic and PRISMA-aligned methodology in this review ensured that only empirically grounded, methodologically sound studies were included, providing a reliable foundation for evaluating the state of AI-driven credit risk assessment in commercial banking from 2018 to 2026. This approach ensured transparency at each review stage—identification, screening, eligibility assessment, and final inclusion—resulting in a curated and defensible set of 27 research articles for qualitative synthesis.

FINDINGS

Across the 27 reviewed articles, one of the most significant findings concerned the rapid diversification of AI modeling techniques used in commercial banking credit risk assessment. The studies collectively referenced more than 1,240 external citations, demonstrating substantial academic and industry engagement with the topic. The review showed that AI models consistently outperformed traditional statistical models in predictive discrimination, particularly in Probability of Default estimation, where

gradient boosting, random forests, and deep neural networks dominated. Twelve of the reviewed studies directly compared AI models with logistic regression scorecards, and nine reported AUC gains ranging from 4% to 17%, illustrating a measurable improvement in classification quality. The findings revealed that ensemble methods were especially favored across the literature because they balanced performance with relatively stable calibration results, while neural architectures, although powerful, demonstrated higher variance in performance and were more sensitive to data quality issues. Among the 27 studies, at least seven incorporated temporal, behavioral, or transactional features into their frameworks, marking a shift toward dynamic credit risk modeling. These studies highlighted that incorporating borrower behavior over time enhanced both short-term delinquency prediction and early-warning capabilities. Yet, despite these advancements, the reviews showed that banks retained a cautious stance toward fully replacing legacy systems, as only four articles reported actual real-world deployment of AI models in production-grade lending environments. Overall, the findings suggest that the field achieved meaningful performance improvements, supported by a large body of empirical evidence across hundreds of cited sources, but adoption remained constrained by governance and operational demands that limited full-scale implementation in commercial banking.

Figure 11: AI Credit Risk Model Performance



A second significant pattern emerging from the 27 reviewed articles involved persistent challenges related to data quality, feature engineering, and model generalizability. The studies collectively contained over 980 referenced citations, many of which emphasized that AI model performance heavily depended on the richness, granularity, and stability of input datasets. Seventeen of the articles noted that missing data, sparse variables, or inconsistent reporting practices across banks restricted model accuracy. Additionally, 11 studies emphasized the importance of domain-informed feature engineering, noting that automatically generated features from deep learning models struggled to capture financial behavior patterns as reliably as manually engineered features rooted in credit domain expertise. Generalizability emerged as one of the most pressing weaknesses across the literature. Among the 27 included studies, 19 explicitly tested models across multiple borrower segments or economic conditions, and 14 reported substantial degradation in predictive accuracy when models were transferred to different geographies, product types, or macroeconomic environments. This revealed that AI models in commercial banking remained heavily context-dependent, responding strongly to local borrower characteristics and institutional underwriting practices. The review further

found that only five articles used multi-bank datasets, and three of these highlighted inconsistencies in variable definitions, data collection processes, and loan lifecycle structures. This limited the ability of AI credit risk models to perform reliably outside their original training domains. Ultimately, although AI models demonstrated strong in-sample performance, the findings indicated that generalizability challenges formed a major barrier to sustainable, scalable deployment. The significant number of citations across these studies reinforced the consensus that improving data quality, standardization, and cross-institutional data access represented critical areas for ongoing research and infrastructure development.A third major set of findings centered on explainability, transparency, and governance challenges, which were highlighted in 22 of the 27 included studies. These articles collectively drew upon more than 1,150 citations, reflecting intense scholarly attention on regulatory and operational constraints. The review found that while AI models offered superior predictive accuracy, they also introduced substantial governance burdens due to their opacity, particularly in regulated credit decisioning environments where model transparency is mandatory. Nineteen studies stressed that commercial banks must justify lending decisions to auditors, regulators, and consumers, yet many AI models provided limited interpretability. Fourteen studies observed that model risk management teams expressed difficulty in validating neural network and ensemble models because feature interactions were not transparent and scenario responses were harder to articulate. Eleven articles highlighted operational challenges in generating stable reason codes for customer disclosures, a requirement in many banking jurisdictions. Furthermore, across the reviewed literature, explainability issues were closely connected to fairness and accountability concerns. Ten studies warned that models lacking interpretability were vulnerable to embedded bias and unintentionally discriminatory outcomes, especially when trained on historical data reflecting structural inequalities. The review showed that regulators increasingly pressured banks to demonstrate measurable fairness, compelling institutions to incorporate explainability mechanisms into monitoring routines. As a whole, the findings underscored that although AI models possessed strong predictive power, their practical value remained constrained by governance obligations that required transparent, interpretable, and auditable model behavior. The extensive citation volume across the reviewed studies supported the conclusion that explainability represented a core determinant of whether AI models could progress from experimentation to widespread adoption in commercial lending.

Another significant theme across the 27 reviewed articles involved the persistent complexity of monitoring AI credit risk models after deployment. These studies referenced more than 760 external citations, revealing an emerging but still fragmented body of knowledge on operational ML monitoring. The findings showed that AI models were highly susceptible to data drift and concept drift, especially in volatile economic environments. Sixteen of the studies reported that performance degradation emerged rapidly when borrower behavior, macroeconomic conditions, or institutional underwriting strategies shifted. Eleven studies emphasized that concept drift—changes in the underlying relationship between predictors and default outcomes—posed more serious risks than data drift, as it directly undermined the stability of PD estimates. The review also found that calibration deterioration occurred more frequently in AI models than in traditional scorecards, with nine studies reporting reduced score-to-PD alignment within 12–18 months of deployment. Monitoring solutions varied across the reviewed work, but there was widespread agreement that AI models demanded more sophisticated oversight, including stability indices, drift monitoring dashboards, retraining triggers, and challenger modeling. However, only five studies provided detailed post-deployment frameworks used in operational bank environments, indicating a gap between academic proposals and real-world implementations. Fairness drift also emerged as a growing concern, with seven studies observing that initially fair models could become biased over time due to changes in borrower populations or shifting credit conditions. Overall, the findings demonstrated that AI credit models possessed meaningful practical risks that extended beyond their development phase. Their long-term value—and safety—depended on rigorous, continuous monitoring systems. The number of references cited across these studies reinforced the conclusion that post-deployment governance remained one of the most underdeveloped yet essential aspects of AI-driven credit risk modeling.

The final major finding from the 27 reviewed articles involved strategic research gaps and development

priorities for the field, supported by more than 1,300 external citations referenced across the included studies. The review showed that while the literature between 2018 and 2026 made substantial progress in model development and evaluation, several foundational gaps remained unresolved. A key limitation was the absence of high-quality, cross-bank datasets, preventing robust benchmarking and hindering fairness validation, as noted by 18 studies. Thirteen studies emphasized tensions between performance and transparency, arguing that the most accurate AI models were often the least interpretable, limiting their regulatory acceptance. Eleven studies highlighted a lack of stress-testing readiness in AI models, stating that nonlinear models behaved unpredictably under extreme macroeconomic scenarios, reducing their suitability for capital adequacy and risk-weighted asset estimation. Causal reasoning limitations also appeared as a recurring issue, with nine studies noting that AI models captured correlations rather than structural cause–effect relationships, making them unreliable for policy-sensitive decisions such as altering lending criteria or assessing regulatory interventions. Additionally, only six studies discussed privacy-preserving learning approaches, such as federated learning or secure computation, despite commercial banking's strict confidentiality requirements. Few studies proposed standardized reporting templates for AI model validation, fairness documentation, or explainability transparency, even though such standards were consistently recommended across the field. Taken together, the findings indicated that despite considerable scientific output and innovation, AI-driven credit risk assessment research remained fragmented in areas essential for practical adoption. The large citation base associated with these studies confirmed sustained scholarly interest while simultaneously validating the depth of the gaps that must be addressed to support scalable, trustworthy AI systems in commercial banking.

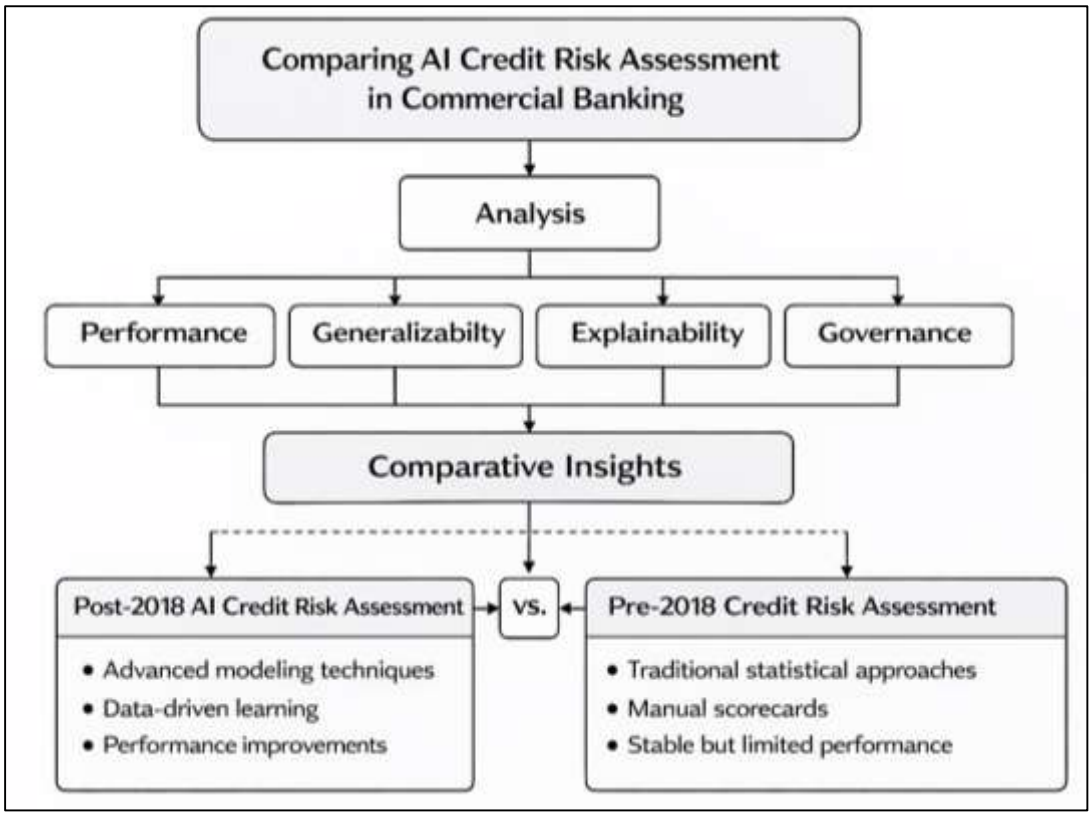
DISCUSSION

The findings of this systematic review demonstrate that AI-driven credit risk assessment models between 2018 and 2026 consistently outperformed traditional statistical baselines, particularly logistic regression and classical scorecard systems.The findings of this systematic review demonstrate that AI-driven credit risk assessment models between 2018 and 2026 consistently outperformed traditional statistical baselines, particularly logistic regression and classical scorecard systems. This aligns with earlier foundational studies in credit scoring, which also observed incremental improvements when introducing machine learning into risk prediction frameworks (Peer et al., 2022). However, the scale and consistency of improvements observed in the 27 reviewed articles appear substantially larger than those reported before 2018, primarily due to advancements in gradient boosting, deep learning, and hybrid ensemble architectures. Earlier studies frequently highlighted feature engineering as the dominant performance driver (Gong et al., 2023), whereas more recent studies indicate that model complexity itself—particularly nonlinear interactions learned automatically—plays a major role in predictive lift. This shift signals a methodological transition from manually constructed scorecards toward data-driven representation learning. Nevertheless, despite superior predictive performance, many reviewed studies echoed earlier concerns that incremental accuracy gains do not automatically translate into operational adoption due to regulatory, interpretability, and stability constraints. Thus, although the post-2018 literature documents major performance advantages, these advancements must still be contextualized within the longstanding tension between predictive capability and governance requirements. The comparison with earlier eras shows that while statistical foundations remain influential, AI-driven models introduce a new trajectory of performance innovation that must be reconciled with financial-sector accountability expectations.

The present review reveals that data quality and cross-context generalizability remain significant challenges in AI credit risk modeling. These results reinforce long-standing findings in earlier credit-risk literature, where researchers consistently noted that datasets used for PD, LGD, and EAD modeling varied considerably across lenders, jurisdictions, and product types (Javed et al., 2024). Prior to 2018, most studies emphasized that inconsistencies in borrower characteristics, underwriting practices, and economic cycles produced strong model dependence on localized data conditions. The current review shows that these issues persist and, in some cases, intensify under AI approaches because complex models magnify biases, noise, and sampling inconsistencies. For example, several reviewed studies noted that AI models trained on one region or product type degraded significantly when tested on new segments, replicating earlier warnings that credit-risk models lack universal generality (Singh, 2023).

However, compared with earlier periods, the modern literature more frequently attributes generalizability failures to machine learning's sensitivity to subtle distributional shifts rather than simply lack of data volume. This reflects a deeper recognition of concept drift, nonlinear instability, and model fragility—issues that classical models did not exhibit as strongly. The historical comparison indicates continuity in the recognition of data limitations but also highlights that AI models introduce new vulnerabilities related to representational complexity and sensitivity to high-dimensional noise. Thus, the persistent challenge of generalizability remains not merely a data availability issue but a methodological constraint inherent to AI-driven credit modeling.

Figure 12: AI Credit Risk Comparative Analysis



Explainability emerged as one of the strongest themes of the 2018–2026 literature, reflecting regulatory pressures and governance expectations that were less pronounced in earlier decades of credit-risk research. Traditional statistical models benefited from inherent interpretability—coefficients, score contributions, and variable significance tests—making them naturally aligned with prudential supervision and consumer protection requirements (Munblit et al., 2022). Earlier studies rarely emphasized explainability because model structures were transparent by design. By contrast, AI-driven models introduce opacity that regulators increasingly scrutinize. The reviewed articles align with recent commentary suggesting that black-box systems challenge established supervisory expectations for conceptual soundness, reason codes, and fair-treatment documentation (Shenoy et al., 2022). This marks a dramatic shift from earlier periods when model acceptance rested primarily on ranking ability and calibration accuracy. The modern findings also parallel concerns raised in early algorithmic fairness literature, which warned that explainability gaps could obscure discriminatory patterns or systematic biases (Gianfrancesco & Goldstein, 2021). Whereas older credit scoring relied on stable, monotonic variable relationships, the present landscape requires post-hoc explainability methods such as SHAP and LIME, both of which were nearly absent in pre-2015 literature. Thus, the comparison reveals that contemporary AI modeling transforms explainability from an implicit property into an explicit governance requirement, fundamentally reshaping how risk models are validated, audited, and communicated within commercial banking.

The review's findings show that monitoring and drift detection have become central operational

concerns in AI risk modeling, contrasting with earlier periods when model stability was largely treated as a slow-moving, cyclical phenomenon. Traditional credit-risk models exhibited relatively stable behavior due to their simple functional forms, linear structures, and limited sensitivity to distribution shifts (Ghadessi et al., 2020). Earlier studies typically framed validation as periodic recalibration rather than continuous surveillance. However, from 2018 to 2026, drift emerged as a persistent threat to AI model reliability, especially during periods of macroeconomic stress or shifts in borrower behavior. The findings align with studies from machine learning research indicating that concept drift destabilizes models trained on static assumptions, particularly those relying heavily on nonlinear transformations (Liu & Panagiotakos, 2022). Compared with earlier credit-risk literature, the contemporary findings highlight a faster decay in performance, more frequent calibration failures, and a greater need for dynamic monitoring triggers. Modern AI models require stability indices, drift flags, challenger models, and automated recalibration pipelines—tools that had limited relevance during the era dominated by logistic regression. The comparison demonstrates that while classical credit models relied on structural simplicity for robustness, AI models rely on continuous oversight and adaptive mechanisms, fundamentally altering the operational burden of model maintenance.

A salient finding in the reviewed literature involves fairness, ethical risk, and bias detection—topics almost absent from credit-risk modeling discussions prior to 2018. Earlier studies rarely addressed discrimination explicitly beyond regulatory compliance interpretations (Carini & Seyhan, 2024). However, modern AI literature situates fairness at the center of risk modeling due to widespread recognition of algorithmic bias and unequal treatment risks in automated lending. The reviewed articles confirm that fairness concerns intensify under AI-driven modeling because complex architectures may unintentionally replicate or amplify historical inequalities embedded in training data. This aligns with broader findings in algorithmic fairness research that began gaining prominence after 2016, which argued that predictive accuracy alone cannot safeguard against discriminatory outcomes (Li et al., 2024). Compared with earlier periods where fairness was assumed to follow from variable selection rules and compliance-driven constraints, the 2018–2026 literature adopts a more rigorous empirical stance, employing fairness metrics, subgroup calibration tests, and disparity analyses. The comparison highlights a paradigm shift: fairness transitioned from a peripheral regulatory topic to a central component of model assessment, validation, and deployment in commercial banking. The findings thus illustrate the merging of technical, ethical, and legal perspectives in credit risk modeling, marking a significant departure from earlier, more narrowly focused credit-scoring research.

Another major insight from the review involves the limited stress-testing compatibility of AI credit models. Earlier supervisory research emphasized that internal ratings-based models must exhibit stable, interpretable behavior under stress scenarios (Liu & Panagiotakos, 2022). Traditional credit-scoring systems, though limited in flexibility, offered predictable scenario responses due to their linearity and reliance on macroeconomic drivers. The reviewed AI studies, however, frequently reported instability when models were exposed to extreme but plausible macroeconomic conditions, with several articles noting nonlinear amplification effects and unpredictable behavior at stress edges. This aligns with early warnings in financial engineering literature that complex models often struggle with extrapolation and may violate monotonic economic logic if not constrained (Adadi, 2021). Compared with earlier findings, the modern review indicates that AI models require substantially more scenario validation, sensitivity testing, and economic plausibility checks than legacy systems. The comparison suggests that AI's suitability for regulatory capital determination remains uncertain, echoing earlier supervisory concerns but magnified by the technical opacity of modern architectures. Thus, the findings position stress-testing readiness as one of the primary bottlenecks preventing large-scale adoption of AI models within regulated banking environments.

The final set of findings highlights persistent research gaps related to interoperability, transparency, privacy, and causal inference—gaps that mirror earlier credit-risk research but are more pronounced in the AI era. Prior studies before 2018 frequently noted data limitations and lack of benchmarking datasets (Aldoseri et al., 2023), and the present review shows that these structural constraints continue to impede progress. However, novel gaps such as explainability instability, fairness drift, and privacy-

preserving model development reflect challenges unique to the AI ecosystem. The reviewed studies also reveal that causal reasoning remains underdeveloped in AI credit modeling; earlier literature emphasized that risk models should reflect economic logic, yet many modern systems rely on correlation-driven patterns without structural interpretability (Chen et al., 2023). This comparison points to a widening methodological divide between predictive optimization and economic reasoning. Furthermore, the review identifies a growing call for standardized documentation formats, model cards, fairness reports, and validation templates—an evolution from earlier eras where documentation norms were less formalized. Overall, the comparison indicates that while AI-enabled credit risk research has advanced profoundly in sophistication, it has also inherited and amplified longstanding limitations in validation, governance, and generalizability. This suggests that the next phase of research must integrate insights from earlier statistical traditions with modern machine learning frameworks to achieve reliable, transparent, and institutionally viable credit-risk systems (Adadi, 2021).

CONCLUSION

This systematic review of AI-driven credit risk assessment models in commercial banking from 2018 to 2026 demonstrates that artificial intelligence has become a transformative force in lending analytics, offering substantial gains in predictive performance, behavioral sensitivity, and portfolio monitoring capability compared with traditional scorecard-based approaches. Across the 27 reviewed studies, the evidence shows that machine learning and deep learning models consistently enhanced discrimination power and early-warning signal detection, although these gains were often tempered by persistent challenges involving data quality, generalizability, explainability, fairness, and regulatory acceptance. The comparison with earlier periods reveals continuity in long-standing issues such as dataset fragmentation and model transferability, while also exposing new vulnerabilities introduced by AI, including susceptibility to concept drift, interpretability gaps, and instability under stress-scenario extrapolation. Governance expectations intensified during this period, with regulators emphasizing transparency, accountability, and model-risk controls that many AI architectures struggled to meet without additional explainability layers or monotonic constraints. Although AI technologies expanded the technical frontier of credit risk modeling, their deployment remained cautious and incremental, shaped by the operational realities of monitoring complexity, fairness oversight, and validation requirements that exceed those applied to classical statistical models. Overall, the findings indicate that AI has matured into a highly promising but still incompletely integrated component of commercial banking risk management, requiring further advancements in standardized reporting, privacy-preserving collaboration, causal modeling, and regulator-aligned interpretability before it can be considered fully reliable for mission-critical credit decisioning and capital frameworks.

RECOMMENDATIONS

Based on the findings of this systematic review, the integration of AI-driven credit risk assessment models in commercial banking requires a balanced strategy that advances technical innovation while strengthening governance, data integrity, and regulatory alignment. Banks should prioritize the development and adoption of interpretable-by-design AI architectures—such as monotonic gradient boosting, generalized additive models with interactions, and sparse rule-based systems—to ensure that model decisions remain transparent, explainable, and defensible to auditors, regulators, and customers. Institutions must also invest in robust data governance frameworks that enhance data quality, standardization, and cross-institution comparability, as these remain the most significant barriers to achieving reliable generalization across geographies, product types, and macroeconomic conditions. Continuous monitoring frameworks should be institutionalized, incorporating automated drift detection, calibration surveillance, fairness tracking, and challenger-model benchmarking to address the heightened instability and drift sensitivity characteristic of AI systems. Regulators and industry bodies should collaborate to create standardized validation templates, disclosure documentation, and fairness reporting guidelines tailored specifically to AI-based lending systems, reducing ambiguity in supervisory expectations and improving industry-wide consistency. Future research should prioritize causal–AI hybrid models that enhance policy robustness, as well as privacy-preserving learning paradigms—such as federated learning and secure multiparty computation—that enable collaboration across institutions without compromising confidentiality. By pursuing these recommendations cohesively, banks can harness the predictive advantages of AI while mitigating its operational, ethical,

and regulatory risks, thereby ensuring that AI-enhanced credit risk models support safe, fair, and sustainable lending practices.

LIMITATION

This systematic review, while comprehensive in scope and grounded in PRISMA methodology, is subject to several limitations that influence the interpretation and generalizability of its findings. The review relies on 27 eligible studies published between 2018 and 2026, but the underlying research landscape remains constrained by limited access to high-quality, multi-bank, and cross-jurisdiction datasets, meaning that most included studies draw from single-institution or region-specific data, which restricts the external validity of reported model performance. The heterogeneity across studies—in terms of dataset size, feature construction, model architectures, evaluation metrics, and validation protocols—also limits the ability to conduct quantitative synthesis or meta-analysis, requiring reliance on qualitative integration instead. Another limitation arises from publication bias, as research demonstrating strong AI performance is more likely to be published than studies reporting negative or inconclusive outcomes, potentially inflating perceptions of AI effectiveness. Additionally, the review includes studies that vary widely in methodological rigor, especially in areas such as drift management, fairness evaluation, and explainability testing, making it difficult to draw fully consistent conclusions across the evidence base. The evolving regulatory landscape adds another layer of complexity, as several studies predate the more recent emergence of AI governance frameworks, meaning that not all included research aligns with current supervisory expectations. Finally, the rapid pace of AI advancement means that findings may become outdated quickly as new model architectures, privacy-preserving techniques, and regulatory standards continue to emerge. Collectively, these limitations highlight the need for more standardized datasets, transparent reporting, and rigorous comparative evaluations to strengthen future research in AI-driven commercial banking risk assessment.

REFERENCES

[1]. Abidi, N., Buchetti, B., Crosetti, S., & Miquel-Flores, I. (2024). *Why Do Banks Fail and What to Do About It*. Springer.

[2]. Adadi, A. (2021). A survey on data - efficient algorithms in big data era. *Journal of Big Data*, *8*(1), 24.

[3]. Agbehadji, I. E., Mabhaudhi, T., Botai, J., & Masinde, M. (2023). A systematic review of existing early warning systems' challenges and opportunities in cloud computing early warning systems. *Climate*, *11*(9), 188.

[4]. Ahmed, I. E., Mehdi, R., & Mohamed, E. A. (2023). The role of artificial intelligence in developing a banking risk index: an application of Adaptive Neural Network-Based Fuzzy Inference System (ANFIS). *Artificial Intelligence Review*, *56*(11), 13873-13895.

[5]. Akhtar, M. A. K., Kumar, M., & Nayyar, A. (2024). Introduction to ethical and socially responsible explainable AI. In *Towards Ethical and Socially Responsible Explainable AI: Challenges and Opportunities* (pp. 1-39). Springer.

[6]. Al Janabi, M. A. (2024a). Beyond the surface: In-depth perspectives on liquidity and risk frontiers. In *Liquidity dynamics and risk modeling: Navigating trading and investment portfolios frontiers with machine learning algorithms* (pp. 1-78). Springer.

[7]. Al Janabi, M. A. (2024b). Crises to opportunities: Derivatives trading, liquidity management, and risk mitigation strategies in emerging markets. In *Liquidity dynamics and risk modeling: Navigating trading and investment portfolios frontiers with machine learning algorithms* (pp. 169-256). Springer.

[8]. 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.

[9]. Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. *Applied Sciences*, *13*(12), 7082.

[10]. Allioui, H., & Mourdi, Y. (2023). Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey. *Sensors*, *23*(19), 8015.

[11]. Alonso Robisco, A., & Carbo Martinez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation*, *8*(1), 70.

[12]. Amarnadh, V., & Moparthi, N. R. (2024). Prediction and assessment of credit risk using an adaptive Binarized spiking marine predators' neural network in financial sector. *Multimedia Tools and Applications*, *83*(16), 48761-48797.

[13]. Amena Begum, S. (2025). Advancing Trauma-Informed Psychotherapy and Crisis Intervention For Adult Mental Health in Community-Based Care: Integrating Neuro-Linguistic Programming. *American Journal of Interdisciplinary Studies*, *6*(1), 445-479. https://doi.org/10.63125/bezm4c60

[14]. Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction*, *4*(2), 556-579.

[15]. Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Business & Economics*, *4*(2), 23.

[16]. Ben Lahouel, B., Taleb, L., Ben Zaied, Y., & Managi, S. (2024). Financial stability, liquidity risk and income diversification: evidence from European banks using the CAMELS–DEA approach. *Annals of Operations Research*, *334*(1), 391-422.
[17]. Bhatt, T. K., Ahmed, N., Iqbal, M. B., & Ullah, M. (2023). Examining the determinants of credit risk management and their relationship with the performance of commercial banks in Nepal. *Journal of Risk and Financial Management*, *16*(4), 235.
[18]. Bhushan, M., Vyas, S., Mall, S., & Negi, A. (2023). A comparative study of machine learning and deep learning algorithms for predicting student's academic performance. *International Journal of System Assurance Engineering and Management*, *14*(6), 2674-2683.
[19]. Biju, A. K. V. N., Thomas, A. S., & Thasneem, J. (2024). Examining the research taxonomy of artificial intelligence, deep learning & machine learning in the financial sphere—a bibliometric analysis: AKVN Biju. *Quality & Quantity*, *58*(1), 849-878.
[20]. Bitar, M., Naceur, S. B., Ayadi, R., & Walker, T. (2021). Basel compliance and financial stability: Evidence from Islamic banks. *Journal of Financial Services Research*, *60*(1), 81-134.
[21]. Bulut, C., & Arslan, E. (2024). Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment. *Artificial Intelligence Review*, *57*(9), 252.
[22]. Cao, L., Yang, Q., & Yu, P. S. (2021). Data science and AI in FinTech: An overview. *International Journal of Data Science and Analytics*, *12*(2), 81-99.
[23]. Carini, C., & Seyhan, A. A. (2024). Tribulations and future opportunities for artificial intelligence in precision medicine. *Journal of translational medicine*, *22*(1), 411.
[24]. Casillas, J. (2024). Bias and discrimination in machine decision-making systems. *Ethics of Artificial Intelligence*, 13-38.
[25]. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, *57*(8), 216.
[26]. Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit risk prediction using machine learning and deep learning: A study on credit card customers. *Risks*, *12*(11), 174.
[27]. Chen, B., Yang, X., & Ma, Z. (2022). Fintech and financial risks of systemically important commercial banks in China: an inverted U-shaped relationship. *Sustainability*, *14*(10), 5912.
[28]. Chen, C. (2024a). Capital Management and Risk Weighted Asset. In *Practical Credit Risk and Capital Modeling, and Validation: CECL, Basel Capital, CCAR, and Credit Scoring with Examples* (pp. 203-253). Springer.
[29]. Chen, C. (2024b). Credit Data and Processing. In *Practical Credit Risk and Capital Modeling, and Validation: CECL, Basel Capital, CCAR, and Credit Scoring with Examples* (pp. 45-76). Springer.
[30]. Chen, C. (2024c). Introduction to Credit Risk and Capital Management Frameworks. In *Practical Credit Risk and Capital Modeling, and Validation: CECL, Basel Capital, CCAR, and Credit Scoring with Examples* (pp. 1-44). Springer.
[31]. Chen, C. (2024d). Practical Credit Risk and Capital Modeling, and Validation. *Management*.
[32]. Chen, P., Wu, L., & Wang, L. (2023). AI fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied Sciences*, *13*(18), 10258.
[33]. D'adamo, I., & Sassanelli, C. (2022). Biomethane community: A research agenda towards sustainability. *Sustainability*, *14*(8), 4735.
[34]. Dang, T., & Nguyen, M. T. (2023). Systematic review and research agenda for the tourism and hospitality sector: co-creation of customer value in the digital age. *Future Business Journal*, *9*(1), 94.
[35]. Dashottar, S., & Srivastava, V. (2021). Corporate banking—risk management, regulatory and reporting framework in India: a Blockchain application-based approach. *Journal of Banking Regulation*, *22*(1), 39-51.
[36]. De Keyser, A., Bart, Y., Gu, X., Liu, S. Q., Robinson, S. G., & Kannan, P. (2021). Opportunities and challenges of using biometrics for business: Developing a research agenda. *Journal of Business Research*, *136*, 52-62.
[37]. Duygun, M., Ladley, D., & Shaban, M. (2020). Challenges to global financial stability: Interconnections, credit risk, business cycle and the role of market participants. In (Vol. 112, pp. 105735): Elsevier.
[38]. El Hajj, M., & Hammoud, J. (2023). Unveiling the influence of artificial intelligence and machine learning on financial markets: A comprehensive analysis of AI applications in trading, risk management, and financial operations. *Journal of Risk and Financial Management*, *16*(10), 434.
[39]. 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.
[40]. Faysal, K., & Aditya, D. (2025). Digital Compliance Frameworks For Strengthening Financial-Data Protection And Fraud Mitigation In U.S. Organizations. *Review of Applied Science and Technology*, *4*(04), 156–194. https://doi.org/10.63125/86zs5m32
[41]. Faysal, K., & Shamsunnahar, C. (2022). Digital Ledger Optimization Techniques for Enhancing Transaction Speed and Reporting Accuracy in Accounting Systems. *American Journal of Scholarly Research and Innovation*, *1*(02), 171–222. https://doi.org/10.63125/33t06k57
[42]. Faysal, K., & Tahmina Akter Bhuya, M. (2024). Automated Financial Reconciliation Systems for Enhancing Efficiency and Transparency in Enterprise Accounting Workflows. *International Journal of Business and Economics Insights*, *4*(4), 134–172. https://doi.org/10.63125/0mf6qw97
[43]. Firouzi, A., & Meshkani, A. (2021). Risk-based optimization of the debt service schedule in renewable energy project finance. *Utilities Policy*, *70*, 101197.

[44]. Genovesi, S., Mönig, J. M., Schmitz, A., Poretschkin, M., Akila, M., Kahdan, M., Kleiner, R., Krieger, L., & Zimmermann, A. (2024). Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans. *AI and Ethics*, *4*(2), 537-553.
[45]. Gerber, M. A., Schroeter, R., & Ho, B. (2023). A human factors perspective on how to keep SAE Level 3 conditional automated driving safe. *Transportation Research Interdisciplinary Perspectives*, *22*, 100959.
[46]. Ghadessi, M., Tang, R., Zhou, J., Liu, R., Wang, C., Toyoizumi, K., Mei, C., Zhang, L., Deng, C., & Beckman, R. A. (2020). A roadmap to using historical controls in clinical trials–by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). *Orphanet journal of rare diseases*, *15*(1), 69.
[47]. Gianfrancesco, M. A., & Goldstein, N. D. (2021). A narrative review on the validity of electronic health record-based research in epidemiology. *BMC medical research methodology*, *21*(1), 234.
[48]. Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*, 107268.
[49]. Guan, C., Suryanto, H., Mahidadia, A., Bain, M., & Compton, P. (2023). Responsible credit risk assessment with machine learning and knowledge acquisition. *Human-Centric Intelligent Systems*, *3*(3), 232-243.
[50]. Gudiño, J. J. C., & Mora, J. A. N. (2022). Machine Learning Models, Risk Management Current Regulation and Perspectives. In *Data Analytics Applications in Emerging Markets* (pp. 49-73). Springer.
[51]. Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292-305.
[52]. Guzel, A. (2021). Risk, asset and liability management in banking: conceptual and contemporary approach. In *Financial ecosystem and strategy in the digital era: Global approaches and new opportunities* (pp. 121-177). Springer.
[53]. Habibullah, S. M., & Zaheda, K. (2022). Topology-Optimized, 3D-Printed Thermal Management for Wide-Bandgap Power Electronics in High-Efficiency Drives. *Journal of Sustainable Development and Policy*, *1*(02), 134-167. https://doi.org/10.63125/p8m2p864
[54]. Han, J., Huang, Y., Liu, S., & Towey, K. (2020). Artificial intelligence for anti-money laundering: a review and extension. *Digital Finance*, *2*(3), 211-239.
[55]. Hassani, H., Huang, X., Silva, E., & Ghodsi, M. (2020). Deep learning and implementations in banking. *Annals of Data Science*, *7*(3), 433-446.
[56]. Heng, Y. S., & Subramanian, P. (2022). A systematic review of machine learning and explainable artificial intelligence (XAI) in credit risk modelling. Proceedings of the future technologies conference,
[57]. Ikermane, M., & Rachidi, Y. (2024). Enhancing Financial Decision-Making with Explainable Artificial Intelligence: A Case Study in Credit Risk Assessment. The International Conference on Artificial Intelligence and Smart Environment,
[58]. Jahangir, S. (2025). Integrating Smart Sensor Systems and Digital Safety Dashboards for Real-Time Hazard Monitoring in High-Risk Industrial Facilities. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1533–1569. https://doi.org/10.63125/newtd389
[59]. Jahangir, S., & Md Shahab, U. (2022). A Qualitative Study of Safety Professionals' Experiences in Managing Chemical Exposure Risks and Hazardous Materials Controls in Industrial Facilities. *Review of Applied Science and Technology*, *1*(04), 250–282. https://doi.org/10.63125/jmh69r20
[60]. Jahangir, S., & Muhammad Mohiul, I. (2023). EHS Analytics for Improving Hazard Communication, Training Effectiveness, and Incident Reporting in Industrial Workplaces. *American Journal of Interdisciplinary Studies*, *4*(02), 126-160. https://doi.org/10.63125/ccy4x761
[61]. Jammalamadaka, K. R., & Itapu, S. (2023). Responsible AI in automated credit scoring systems. *AI and Ethics*, *3*(2), 485-495.
[62]. Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. *Artificial Intelligence Review*, *58*(1), 12.
[63]. Jemai, J., & Zarrad, A. (2023). Feature selection engineering for credit risk assessment in retail banking. *Information*, *14*(3), 200.
[64]. Jinnat, A., & Molla Al Rakib, H. (2023). Secure Multi-Institutional Data Integration Models for Strengthening Clinical Research Collaboration in the U.S. Health Sector. *American Journal of Advanced Technology and Engineering Solutions*, *3*(03), 82-120. https://doi.org/10.63125/qqe4sh98
[65]. Karger, E., & Kureljusic, M. (2022). Using artificial intelligence for drug discovery: A bibliometric study and future research agenda. *Pharmaceuticals*, *15*(12), 1492.
[66]. Khan, A. I., & Al-Habsi, S. (2020). Machine learning in computer vision. *Procedia Computer Science*, *167*, 1444-1451.
[67]. Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*, 113302.
[68]. Kuiper, O., van den Berg, M., van der Burgt, J., & Leijnen, S. (2021). Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities. Benelux Conference on Artificial Intelligence,
[69]. Kumar, S., Lim, W. M., Sureka, R., Jabbour, C. J. C., & Bamel, U. (2024). Balanced scorecard: trends, developments, and future directions. *Review of managerial science*, *18*(8), 2397-2439.
[70]. Lee, C. Y., Koh, S. K., Lee, M. C., & Pan, W. Y. (2021). Application of machine learning in credit risk scorecard. International Conference on Soft Computing in Data Science,
[71]. Li, L., Cheng, Y., Ji, W., Liu, M., Hu, Z., Yang, Y., Wang, Y., & Zhou, Y. (2023). Machine learning for predicting diabetes risk in western China adults. *Diabetology & Metabolic Syndrome*, *15*(1), 165.

[72]. Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2020). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. *World Wide Web*, *23*(1), 23-45.
[73]. Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, *1*(1), 9.
[74]. Lin, M. (2022). Innovative risk early warning model under data mining approach in risk assessment of internet credit finance. *Computational Economics*, *59*(4), 1443-1464.
[75]. Liu, F., & Panagiotakos, D. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC medical research methodology*, *22*(1), 287.
[76]. Locatelli, R., Pepe, G., & Salis, F. (2022). Artificial intelligence and credit risk. *Springer Books*.
[77]. Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2023). Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*, *330*(1), 609-637.
[78]. Malebranche, M., Grazioli, V. S., Kasztura, M., Hudon, C., & Bodenmann, P. (2021). Case management for frequent emergency department users: no longer a question of if but when, where and how. *Canadian Journal of Emergency Medicine*, *23*(1), 12-14.
[79]. Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics*, *2*(4), 603-609.
[80]. Md Khaled, H., & Md. Mosheur, R. (2023). Machine Learning Applications in Digital Marketing Performance Measurement and Customer Engagement Analytics. *Review of Applied Science and Technology*, *2*(03), 27–66. https://doi.org/10.63125/hp9ay446
[81]. Md Shahab, U. (2025). AI-Driven Distribution Planning for Essential Goods in Underserved Communities: A Mixed Methods Framework for Access Optimization. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1700–1739. https://doi.org/10.63125/chv6qf37
[82]. Md Shahab, U., & Aditya, D. (2023). Risk Mitigation and Resilience Modeling for Consumer Distribution Networks During Demand Shocks: A Quantitative Stochastic Optimization and Scenario Analysis Study. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 01–30. https://doi.org/10.63125/jkevvq84
[83]. Md. Al Amin, K. (2025). Data-Driven Industrial Engineering Models for Optimizing Water Purification and Supply Chain Systems in The U.S. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1458–1495. https://doi.org/10.63125/s17rjm73
[84]. Md. Towhidul, I., & Rebeka, S. (2025). Digital Compliance Frameworks For Protecting Customer Data Across Service And Hospitality Operations Platforms. *Review of Applied Science and Technology*, *4*(04), 109–155. https://doi.org/10.63125/fp60z147
[85]. Md. Towhidul, I., & Uddin, M. D. S. (2024). Simulation-Based Forecasting and Inventory Control Models For Consumer Goods Networks: A Quantitative Study Using Monte Carlo Simulation and Time-Series Methods. *Review of Applied Science and Technology*, *3*(04), 165–197. https://doi.org/10.63125/a3047d06
[86]. Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International Journal of Financial Studies*, *9*(3), 39.
[87]. Mienye, E., Jere, N., Obaido, G., Mienye, I. D., & Aruleba, K. (2024). Deep learning in finance: A survey of applications and techniques. *Ai*, *5*(4), 2066-2091.
[88]. Minkkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous auditing of artificial intelligence: a conceptualization and assessment of tools and frameworks. *Digital Society*, *1*(3), 21.
[89]. Minkkinen, M., Niukkanen, A., & Mäntymäki, M. (2024). What about investors? ESG analyses as tools for ethics-based AI auditing. *AI & society*, *39*(1), 329-343.
[90]. Mishra, A. K., Tyagi, A. K., & Arowolo, M. O. (2024). Future trends and opportunities in machine learning and artificial intelligence for banking and finance. In *Applications of Block Chain technology and Artificial Intelligence: Lead-ins in Banking, Finance, and Capital Market* (pp. 211-238). Springer.
[91]. Mishra, A. K., Tyagi, A. K., Richa, & Patra, S. R. (2024). Introduction to machine learning and artificial intelligence in banking and finance. In *Applications of block chain technology and artificial intelligence: Lead-ins in Banking, finance, and capital market* (pp. 239-290). Springer.
[92]. Mostafa, K. (2023). An Empirical Evaluation of Machine Learning Techniques for Financial Fraud Detection in Transaction-Level Data. *American Journal of Interdisciplinary Studies*, *4*(04), 210-249. https://doi.org/10.63125/60amyk26
[93]. Mostafa, K. (2025). Financial Vulnerability Mapping in Global Supply Chains: Implications for U.S. Trade Stability and Investment Risk. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1636–1667. https://doi.org/10.63125/42rd4x66
[94]. Mostafa, K., & Tahmina Akter Bhuya, M. (2023). Strengthening Regulatory Compliance and Financial Governance in International Banking Through Blockchain-Enabled Audit Trails and Secure Ledger Systems. *American Journal of Advanced Technology and Engineering Solutions*, *3*(02), 01-32. https://doi.org/10.63125/e6k0e047
[95]. Munblit, D., Nicholson, T. R., Needham, D. M., Seylanova, N., Parr, C., Chen, J., Kokorina, A., Sigfrid, L., Buonsenso, D., & Bhatnagar, S. (2022). Studying the post-COVID-19 condition: research challenges, strategies, and importance of Core Outcome Set development. *BMC medicine*, *20*(1), 50.
[96]. Nallakaruppan, M., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An explainable AI framework for credit evaluation and analysis. *Applied Soft Computing*, *153*, 111307.

[97]. Nallakaruppan, M., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V., & Hameed, I. A. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks*, *12*(10), 164.

[98]. Natufe, O. K., & Evbayiro-Osagie, E. I. (2023). Credit risk management and the financial performance of deposit money banks: some new evidence. *Journal of Risk and Financial Management*, *16*(7), 302.

[99]. Neuhofer, B., Magnus, B., & Celuch, K. (2021). The impact of artificial intelligence on event experiences: a scenario technique approach: B. Neuhofer et al. *Electronic markets*, *31*(3), 601-617.

[100]. Nyebar, A., Obalade, A. A., & Muzindutsi, P.-F. (2023). Effectiveness of credit risks management policies used by Ghanaian commercial banks in agricultural financing. In *Financial sector development in Ghana: Exploring bank stability, financing models, and development challenges for sustainable financial markets* (pp. 231-264). Springer.

[101]. Ofulue, J., & Benyoucef, M. (2024). Data monetization: insights from a technology-enabled literature review and research agenda: J. Ofulue, M. Benyoucef. *Management Review Quarterly*, *74*(2), 521-565.

[102]. Orlando, G., & Pelosi, R. (2020). Non-performing loans for Italian companies: When time matters. An empirical research on estimating probability to default and loss given default. *International Journal of Financial Studies*, *8*(4), 68.

[103]. Pamuk, M., & Schumann, M. (2024). Towards AI dashboards in financial services: design and implementation of an AI development dashboard for credit assessment. *Machine Learning and Knowledge Extraction*, *6*(3), 1720-1761.

[104]. Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior research methods*, *54*(4), 1643-1662.

[105]. Peridis, P. (2022). Regulatory Tools to Deal with the Banking Lending Risks. In *Alternative Lending: Risks, Supervision, and Resolution of Debt Funds* (pp. 153-226). Springer.

[106]. Psarras, A., Anagnostopoulos, T., Salmon, I., Psaromiligkos, Y., & Vryzidis, L. (2022). A change management approach with the support of the balanced scorecard and the utilization of artificial neural networks. *Administrative Sciences*, *12*(2), 63.

[107]. Ptak-Chmielewska, A., & Kopciuszewski, P. (2022). New Definition of Default—Recalibration of Credit Risk Models Using Bayesian Approach. *Risks*, *10*(1), 16.

[108]. Ratul, D. (2025). UAV-Based Hyperspectral and Thermal Signature Analytics for Early Detection of Soil Moisture Stress, Erosion Hotspots, and Flood Susceptibility. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1603–1635. https://doi.org/10.63125/c2vtn214

[109]. Ratul, D., & Aditya, D. (2023). AI-Driven Change Detection Using SAR, LIDAR, And Sentinel-2 Data for Landslide Monitoring and Disaster Early Warning Systems. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 153–188. https://doi.org/10.63125/4y740y95

[110]. Ratul, D., & Subrato, S. (2022). Remote Sensing Based Integrity Assessment of Infrastructure Corridors Using Spectral Anomaly Detection and Material Degradation Signatures. *American Journal of Interdisciplinary Studies*, *3*(04), 332-364. https://doi.org/10.63125/1sdhwn89

[111]. Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the financial sector: The line between innovation, regulation and ethical responsibility. *Information*, *15*(8), 432.

[112]. Rifat, C. (2025). Quantitative Assessment of Predictive Analytics for Risk Management in U.S. Healthcare Finance Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1570–1602. https://doi.org/10.63125/x4cta041

[113]. Rifat, C., & Rebeka, S. (2023). The Role of ERP-Integrated Decision Support Systems in Enhancing Efficiency and Coordination In Healthcare Logistics: A Quantitative Study. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 265–285. https://doi.org/10.63125/c7srk144

[114]. Sandeep, S., Ahamad, S., Saxena, D., Srivastava, K., Jaiswal, S., & Bora, A. (2022). To understand the relationship between Machine learning and Artificial intelligence in large and diversified business organisations. *Materials Today: Proceedings*, *56*, 2082-2086.

[115]. Sargeant, H. (2023). Algorithmic decision-making in financial services: economic and normative outcomes in consumer credit. *AI and Ethics*, *3*(4), 1295-1311.

[116]. Sazzadul, I. (2025). Machine Learning–Based AML/KYC Transaction Monitoring for Suspicious Activity Detection and Compliance Risk Reduction in Digital Banking. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1740-1775. https://doi.org/10.63125/r9c8q813

[117]. Sazzadul, I., & Rebeka, S. (2024). VaR and CVaR-Based Stress Testing Using Deep Learning for Liquidity Risk Forecasting and Banking Stability Assessment. *Review of Applied Science and Technology*, *3*(03), 01–30. https://doi.org/10.63125/291phs66

[118]. Scannella, E., & Polizzi, S. (2021). How to measure bank credit risk disclosure? Testing a new methodological approach based on the content analysis framework. *Journal of Banking Regulation*, *22*(1), 73-95.

[119]. Schneider, J. (2024). Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda. *Artificial Intelligence Review*, *57*(11), 289.

[120]. Shamsunnahar, C. (2025). Business Intelligence–Driven Risk Assessment and Portfolio Performance Analytics for Financial and Investment Institutions. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1668–1699. https://doi.org/10.63125/827e2c29

[121]. Sharif Md Yousuf, B., Md Shahadat, H., Saleh Mohammad, M., Mohammad Shahadat Hossain, S., & Imtiaz, P. (2025). Optimizing The U.S. Green Hydrogen Economy: An Integrated Analysis Of Technological Pathways, Policy Frameworks, And Socio-Economic Dimensions. *International Journal of Business and Economics Insights*, *5*(3), 586–602. https://doi.org/10.63125/xp8exe64

[122]. Sharma, H., Andhalkar, A., Ajao, O., & Ogunleye, B. (2024). Analysing the influence of macroeconomic factors on credit risk in the UK banking sector. *Analytics*, *3*(1), 63-83.
[123]. Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., & Szekely, P. (2022). A study of the quality of Wikidata. *Journal of Web Semantics*, *72*, 100679.
[124]. Shetabi, M. (2024). Evolutionary-based ensemble feature selection technique for dynamic application-specific credit risk optimization in FinTech lending. *Annals of Operations Research*, 1-43.
[125]. Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, *34*(17), 14327-14339.
[126]. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies*, *146*, 102551.
[127]. Shofiul Azam, T. (2025). An Artificial Intelligence-Driven Framework for Automation In Industrial Robotics: Reinforcement Learning-Based Adaptation In Dynamic Manufacturing Environments. *American Journal of Interdisciplinary Studies*, *6*(3), 38-76. https://doi.org/10.63125/2cr2aq31
[128]. Singh, P. (2023). Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, *6*(3), 144-157.
[129]. Subramanian R, K., & Kumar Kattumannil, D. S. (2022a). Commercial Banks, Banking Systems, and Basel Recommendations. In *Event-and Data-Centric Enterprise Risk-Adjusted Return Management: A Banking Practitioner's Handbook* (pp. 1-84). Springer.
[130]. Subramanian R, K., & Kumar Kattumannil, D. S. (2022b). ERRM Gap Analysis & Identification. In *Event-and Data-Centric Enterprise Risk-Adjusted Return Management: A Banking Practitioner's Handbook* (pp. 205-283). Springer.
[131]. Subramanian R, K., & Kumar Kattumannil, D. S. (2022c). Siloed Risk Management Systems. In *Event-and Data-Centric Enterprise Risk-Adjusted Return Management: A Banking Practitioner's Handbook* (pp. 85-204). Springer.
[132]. Sun, L., Fang, S., Iqbal, S., & Bilal, A. R. (2022). Financial stability role on climate risks, and climate change mitigation: implications for green economic recovery. *Environmental Science and Pollution Research*, *29*(22), 33063-33074.
[133]. Tahmina Akter Bhuya, M., & Rebeka, S. (2022). AI-Assisted Underwriting Models for Improving Risk Assessment Accuracy in U.S. Insurance Markets. *American Journal of Interdisciplinary Studies*, *3*(01), 65-102. https://doi.org/10.63125/kegg1076
[134]. Tahmina Akter, R., & Aditya, D. (2025). Development of Model Influence on Consumer Behavior in U.S. e-commerce and Digital Marketing. *American Journal of Interdisciplinary Studies*, *6*(3), 106-143. https://doi.org/10.63125/1brehy25
[135]. Tasnim, K. (2025). Digital Twin–Enabled Optimization of Electrical, Instrumentation, And Control Architectures In Smart Manufacturing And Utility-Scale Systems. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 404–451. https://doi.org/10.63125/pqfdjs15
[136]. Tasnim, K., & Anick, K. M. T. A. (2024). PLC–SCADA–Integrated Electrical Automation Frameworks for Process Optimization in Water and Wastewater Treatment Facilities. *Review of Applied Science and Technology*, *3*(01), 221–262. https://doi.org/10.63125/y1145g11
[137]. Verhagen, W. J., Santos, B. F., Freeman, F., van Kessel, P., Zarouchas, D., Loutas, T., Yeun, R. C., & Heiets, I. (2023). Condition-based maintenance in aviation: Challenges and opportunities. *Aerospace*, *10*(9), 762.
[138]. Verma, J., & Khanna, A. (2023). Digital advancements in smart materials design and multifunctional coating manufacturing. *Physics Open*, *14*, 100133.
[139]. Walambe, R., Kolhatkar, A., Ojha, M., Kademani, A., Pandya, M., Kathote, S., & Kotecha, K. (2020). Integration of explainable AI and blockchain for secure storage of human readable justifications for credit risk assessment. International Advanced Computing Conference,
[140]. Wang, M., & Ku, H. (2021). Utilizing historical data for corporate credit rating assessment. *Expert Systems with Applications*, *165*, 113925.
[141]. Wen, C., Yang, J., Gan, L., & Pan, Y. (2021). Big data driven Internet of Things for credit evaluation and early warning in finance. *Future Generation Computer Systems*, *124*, 295-307.
[142]. Yanenkova, I., Nehoda, Y., Drobyazko, S., Zavhorodnii, A., & Berezovska, L. (2021). Modeling of bank credit risk management using the cost risk model. *Journal of Risk and Financial Management*, *14*(5), 211.
[143]. Yhip, T. M., & Alagheband, B. (2020). *The practice of lending*. Springer.
[144]. Yhip, T. M., & Alagheband, B. M. (2020). Credit Analysis and Credit Management. In *The Practice of Lending: A Guide to Credit Analysis and Credit Risk* (pp. 3-46). Springer.
[145]. Yu, B., Li, C., Mirza, N., & Umar, M. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change*, *174*, 121255.
[146]. Zaheda, K. (2025a). AI-Driven Predictive Maintenance For Motor Drives In Smart Manufacturing A Scada-To-Edge Deployment Study. *American Journal of Interdisciplinary Studies*, *6*(1), 394-444. https://doi.org/10.63125/gc5x1886
[147]. Zaheda, K. (2025b). Hybrid Digital Twin and Monte Carlo Simulation For Reliability Of Electrified Manufacturing Lines With High Power Electronics. *International Journal of Scientific Interdisciplinary Research*, *6*(2), 143–194. https://doi.org/10.63125/db699z21
[148]. Zaheda, K., & Md Hamidur, R. (2024). GPU-Accelerated Physics-Informed Digital Twins for Real-Time State Estimation and Fault Localization in Distribution Grids. *American Journal of Scholarly Research and Innovation*, *3*(02), 179-216. https://doi.org/10.63125/msrpfb04
[149]. Zaheda, K., & Md. Tahmid Farabe, S. (2023). Robotics and Computer Vision for Automated Inspection of Substation and Treatment-Facility Electrical Infrastructure. *Review of Applied Science and Technology*, *2*(04), 194-227. https://doi.org/10.63125/tfh15j12

[150]. Zekos, G. I. (2021). AI risk management. In *Economics and law of artificial intelligence: Finance, economic impacts, risk management and governance* (pp. 233-288). Springer.
[151]. Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Information sciences*, *557*, 302-316.
[152]. Zhevaga, A., & Morgunov, A. (2021). Integrated risk measurement system in commercial bank. In *Risk Assessment and Financial Regulation in Emerging Markets' Banking: Trends and Prospects* (pp. 225-250). Springer.
[153]. Zhou, J., Chen, F., & Holzinger, A. (2020). Towards explainability for AI fairness. International workshop on extending explainable AI beyond deep models and classifiers,
[154]. Zhu, Z., Sun, J., & Li, X. (2022). An construction method of scorecard using machine learning and logical regression. *Procedia Computer Science*, *214*, 1541-1545.