

Systematic Review

Deep Learning for Credit Risk Prediction in Fintech Lending: A Systematic Literature Review on Model Architectures, Imbalanced Data Handling, and Research Agenda

Moch Panji Agung Saputra ^{1,*} , Sukono ¹ , Riaman ¹ , Alit Kartiwa ¹ , Masnita Misiran ²
and Alim Jaizul Wahid ³

¹ Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jatinangor, Sumedang 45363, Indonesia; sukono@unpad.ac.id (S.); riaman@unpad.ac.id (R.); alit.kartiwa@unpad.ac.id (A.K.)

² School of Quantitatives Sciences, Universiti Utara Malaysia, Sintok 06010, Kedah, Malaysia; masnita@uum.edu.my

³ Master Program in Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jatinangor, Sumedang 45363, Indonesia; alim24002@mail.unpad.ac.id

* Correspondence: moch16006@mail.unpad.ac.id

Abstract

The purpose of this article is to conduct a systematic literature review on the role of deep learning in credit risk prediction for fintech lending, with particular emphasis on model architectures, imbalanced data handling techniques, and the mathematical foundations underpinning these methods. Open-access scientific publications retrieved from three complementary databases (Scopus, IEEE Xplore Digital Library, and Web of Science Core Collection) were used to conduct the systematic literature review. Following the PRISMA 2020 protocol, 30 publications were selected after a rigorous multi-stage screening process involving deduplication across databases, temporal filtering (2015–2026), and thematic eligibility assessment. Data for analysis were processed using Python-based bibliometric tools and network analysis (replicating R Bibliometrix and VOSviewer functionalities). The results of the analysis indicate a sustained growth in research on deep learning applications for fintech credit risk, with a compound annual growth rate (CAGR) of approximately 29.2% and an average of 27.17 citations per document. The segmentation of the studied conceptual landscape made it possible to identify four interconnected thematic clusters: (1) peer-to-peer lending and default prediction architectures; (2) explainability and XAI-based methods for credit scoring; (3) imbalanced data and hybrid deep-learning frameworks; and (4) credit risk assessment combining deep learning and statistical approaches. The following research areas on the deep learning-based transformation of fintech credit risk prediction have been identified: (1) feedforward deep neural networks and attention-based architectures (LSTM, CNN) as dominant predictive engines; (2) hybrid deep ensemble and deep-boosting frameworks (e.g., LightGBM-Attention, GBDT-Deep FFM) as emerging high-performance paradigms; (3) specialized techniques for imbalanced data handling—including ADASYN, SMOTE, cost-sensitive learning, and balanced stratified prioritized experience replay—as critical methodological frontiers; and (4) transformer-based and XAI-integrated architectures as the emerging frontier. The originality of this article lies in its explicit focus on the mathematical and methodological challenges of deep learning-based credit risk prediction in fintech lending, providing an actionable research agenda that addresses class imbalance, uncertainty quantification, loss function design, concept drift, and regulatory compliance. The findings provide valuable insights for scholars, practitioners, and policymakers, and outline a concrete roadmap for developing more accurate, robust, and explainable credit risk models in the rapidly evolving fintech ecosystem.



Academic Editors: Mohammad Mahdi Mousavi, Jamal Ouenniche and Stefano Grillini

Received: 29 April 2026

Revised: 8 June 2026

Accepted: 9 June 2026

Published: 26 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: fintech lending; credit risk prediction; deep learning; peer-to-peer lending; systematic literature review; bibliometric analysis

1. Introduction

The rapid digitalization of financial services has given rise to fintech lending—a paradigm in which credit intermediation is conducted through digital platforms, often without the direct involvement of traditional banking institutions. Peer-to-peer (P2P) lending, online marketplace lending, and digital consumer credit have collectively transformed the global credit landscape, offering faster, more accessible, and data-rich lending experiences (Turiel & Aste, 2020; Xu et al., 2021). However, the same characteristics that make fintech lending attractive—speed, scale, and reliance on alternative data—also amplify credit risk in ways that traditional statistical credit scoring models struggle to capture. Borrower default rates on fintech platforms have historically exceeded those of conventional banks, making accurate credit risk prediction not merely a technical objective but a prerequisite for the sustainability of the industry itself (Lyócsa et al., 2022; Ma et al., 2021).

Credit risk, defined as the probability that a borrower will fail to meet contractual obligations, lies at the intersection of financial mathematics, statistics, and computational intelligence. For decades, the industry relied on parametric statistical models—logistic regression, linear discriminant analysis, and Altman’s Z-score—to estimate the probability of default (Trinh, 2024). While interpretable and theoretically grounded, these models impose restrictive distributional assumptions and often fail to capture the complex, nonlinear, and high-dimensional patterns embedded in contemporary fintech data. The advent of machine learning (ML)—particularly ensemble methods such as random forests, gradient-boosted decision trees, and XGBoost—represented a substantial leap in predictive accuracy (Xu et al., 2021; Ko et al., 2022). However, the most recent wave of innovation has come from deep learning (DL): a family of representation-learning methods built on multilayered neural networks capable of automatically extracting hierarchical features from raw data.

Deep learning has fundamentally altered the credit risk modeling landscape. Convolutional neural networks (CNNs), originally developed for image recognition, have been adapted to extract abstract features from tabular and sequential credit data (Kim & Cho, 2019). Recurrent architectures such as long short-term memory (LSTM) networks and attention mechanisms have enabled the modeling of temporal patterns in borrower behavior (C. Wang et al., 2019). Graph convolutional networks (GCNs) exploit the relational structure among loan applications, transforming credit risk assessment into a graph-learning problem (Z. Li et al., 2024). More recent developments—including deep reinforcement learning (DRL) with prioritized experience replay, hybrid attention-LightGBM frameworks, and deep ensemble learning—push the methodological frontier further still (Y. Wang et al., 2024; Ying et al., 2025). Transformer-based architectures—including BERT adaptations—have recently been applied to credit scoring using P2P loan text descriptions, demonstrating both strong predictive accuracy and new dimensions of adversarial robustness (Schwab & Kriebel, 2025). Collectively, these advances signal a decisive shift from statistical prediction to data-driven representation learning for credit risk.

Yet the adoption of deep learning in fintech credit risk prediction is far from straightforward. Fintech lending datasets are characterized by severe class imbalance: defaulters typically constitute only 5–20% of observations, causing conventional loss functions to bias models toward the majority class (Owusu et al., 2023; Chen et al., 2021). Concept drift—the gradual or abrupt change in the statistical properties of the target variable over time—undermines the stability of offline-trained models in the rapidly evolving fintech

environment (Zhang et al., 2020). Uncertainty quantification remains underdeveloped, despite its centrality to portfolio-level risk management. The opacity of deep neural networks raises concerns about explainability and regulatory compliance, particularly under frameworks such as the EU Artificial Intelligence Act (European Parliament and Council, 2024) and Basel III/IV (Ariza-Garzón et al., 2020; Straus et al., 2025). These mathematical and methodological challenges constitute the defining frontier of current research.

Despite the growing body of literature on deep learning for fintech credit risk, several systematic gaps remain. First, existing reviews tend to treat credit risk as a monolithic problem, rarely distinguishing the specific methodological challenges of fintech lending from those of conventional bank credit. Second, most reviews provide descriptive catalogs of models without examining the underlying mathematical handling of imbalanced data, uncertainty, and loss functions—arguably the most consequential methodological dimensions. Third, bibliometric syntheses that combine quantitative network analysis with deep qualitative reading of the mathematical machinery of deep learning models remain scarce. Fourth, prior single-database reviews may systematically underrepresent high-impact contributions from IEEE Xplore and Web of Science, as demonstrated by the absence of Ariza-Garzón et al. (2020) (138 citations) from Scopus-only searches. This paper addresses these gaps by conducting a systematic, PRISMA-compliant, multi-database literature review of deep learning applications for credit risk prediction in fintech lending.

The relevance of this study is grounded in both practical and theoretical considerations. On the practical side, fintech lending continues to expand globally—particularly in emerging markets where it serves populations excluded from traditional banking. Accurate, robust, and interpretable credit risk models are essential not only for platform profitability but also for consumer protection and systemic financial stability. On the theoretical side, the intersection of deep learning and credit risk exemplifies the broader transformation of financial mathematics in the age of artificial intelligence.

Bibliometric analysis is a modern and innovative tool for identifying research development trends and evaluating scientific output in various fields (Ellegaard & Wallin, 2015). Its ability to combine quantitative metrics with thematic mapping makes it particularly suitable for domains such as deep learning-based credit risk prediction, where methodological diversity coexists with rapid knowledge accumulation (Donthu et al., 2021). To achieve the goal of this review, it is necessary to address the following research questions:

- (1) RQ1. What is the volume, temporal distribution, geographic spread, and scientific productivity of research on deep learning-based credit risk prediction in fintech lending?
- (2) RQ2. Which deep learning architectures have been applied for credit risk prediction in fintech lending, and how do their predictive performances compare?
- (3) RQ3. What variables and data sources—including traditional credit features and alternative data—have been utilized as predictors in deep learning-based fintech credit risk models?
- (4) RQ4. How have researchers addressed class imbalance, uncertainty quantification, loss function design, and concept drift in deep learning models for fintech credit risk?
- (5) RQ5. What evaluation metrics, validation strategies, and benchmark datasets are employed in the literature?
- (6) RQ6. What are the current research gaps and future directions for enhancing the accuracy, robustness, and generalizability of deep learning models in fintech credit risk prediction?

The remainder of the paper is organized as follows. Section 2 outlines the materials and methods, including the multi-database search strategy, PRISMA protocol, and analytical tools. Section 3 presents the descriptive and performance analysis of the corpus—drawn from three bibliographic databases—together with the thematic and network analyses

identifying four clusters, coupling patterns, and conceptual structure. Section 4 provides an in-depth synthesis organized around the six research questions, with particular attention to the mathematical handling of imbalanced data, the newly identified XAI cluster, and the emerging importance of regulatory compliance. Section 5 concludes with theoretical, practical, and policy implications, together with limitations and directions for future research.

2. Materials and Methods

The research methodology of this study follows the principles of a systematic literature review (SLR) combined with bibliometric analysis, including terminology mapping, keyword co-occurrence clustering, bibliographic coupling, and retrospective thematic analysis. The overarching goal is to combine the rigor and reproducibility of the PRISMA 2020 framework (Page et al., 2021) with the analytical power of bibliometric tools, thereby providing a transparent and replicable synthesis of the scholarly landscape on deep learning-based credit risk prediction in fintech lending. The completed PRISMA 2020 checklist is provided as Supplementary Material (Table S1).

2.1. Database Selection

Scientific publications obtained from three complementary databases were used for this review: Scopus (Elsevier, Amsterdam, The Netherlands), IEEE Xplore Digital Library (IEEE, Piscataway, NJ, USA), and Web of Science Core Collection (Clarivate, Seoul, Republic of Korea). This multi-database approach was adopted in direct response to the recognition that no single database provides exhaustive coverage of the interdisciplinary literature at the intersection of deep learning, financial mathematics, and fintech lending—a finding consistent with prior systematic reviews in adjacent domains (Donthu et al., 2021). Scopus provides broad multidisciplinary coverage of business, finance, economics, and computer science, with standardized metadata exports (affiliations, author keywords, full reference lists) essential for bibliometric network analysis. IEEE Xplore Digital Library contributes high-impact technical papers from IEEE Transactions journals (IEEE Access, IEEE Transactions on Neural Networks, IEEE Transactions on Knowledge and Data Engineering) that are sometimes underrepresented in Scopus. Web of Science Core Collection adds coverage of finance-oriented and operations research journals (European Journal of Operational Research, Financial Innovation, Quantitative Finance and Economics) as well as journals indexed under the Emerging Sources Citation Index (ESCI).

The choice of Scopus as the primary database was further motivated by: (i) the availability of a wide range of bibliometric metrics, including citation counts, h-index, g-index, and m-index; (ii) the assignment of publications to distinct subject areas and SciVal topics; (iii) a large repertoire of high-quality peer-reviewed venues; and (iv) the richness of exported metadata (Aria & Cuccurullo, 2017). Unlike Google Scholar, all three databases enforce editorial curation, ensuring that retrieved records meet minimum indexing standards essential for a systematic review.

2.2. Search Strategy

The search strategy was designed to capture studies at the intersection of three research domains: (i) deep learning methods; (ii) credit risk prediction; and (iii) fintech lending. To ensure both precision and recall, the search combined controlled Boolean operators (AND, OR), wildcard truncation (*), and synonym expansion. Searches were executed on 4 June 2026 across all three databases. For Scopus, the TITLE-ABS-KEY field was used with Boolean combinations of deep-learning terms (“deep learning”, “neural network*”, “LSTM”, “CNN”, “RNN”, “transformer”, “attention mechanism”, “graph neural network*”,

“autoencoder”, “GAN”), credit-risk terms (“credit risk”, “credit scoring”, “credit default*”, “loan default*”, “default prediction”, “credit assessment”), and fintech-lending terms (“fintech”, “P2P lending”, “peer-to-peer lending”, “online lending”, “digital lending”, “marketplace lending”, “online loan*”). For Web of Science, the equivalent TS = (Topic Search) field was used with the identical keyword combination. For IEEE Xplore, the “All Metadata” search field was employed with the same Boolean combinations. The search was applied to the title, abstract, and keywords fields in all three databases.

2.3. Inclusion and Exclusion Criteria (PRISMA 2020)

To establish additional selection criteria, the PRISMA 2020 protocol was adopted. The following inclusion criteria were applied:

- (1) Temporal scope: publications between 2015 and 2026, corresponding to the emergence and maturation of deep learning applications in financial services;
- (2) Language: publications in English;
- (3) Document type: peer-reviewed journal articles and conference papers;
- (4) Access: open-access publications, to ensure full-text availability for qualitative synthesis and to align with the principles of transparency and reproducibility;
- (5) Thematic relevance: publications whose title and abstract demonstrably address deep learning or neural network-based credit risk prediction in fintech, peer-to-peer, or online lending contexts.

Exclusion criteria removed records outside the specified timespan, publications focused exclusively on non-fintech domains (e.g., traditional bank credit without digital lending context), editorials and opinion pieces, and studies where deep learning was mentioned only peripherally without being the primary methodological contribution.

The filtering sequence is illustrated in Figure 1 using the PRISMA 2020 flow diagram. The initial queries returned 121 records from Scopus, 113 records from IEEE Xplore, and 20 records from Web of Science, yielding a combined raw pool of 254 records. After cross-database deduplication (matching by DOI and title-similarity), 14 duplicate records were removed, leaving 240 unique records. After applying the temporal filter (2015–2026), 226 records remained. The open-access filter narrowed the corpus to 37 records. Filtering by document type reduced the set to 34 records. Title-level screening removed 2 records; abstract-level screening removed an additional 2 records. The final corpus comprised 30 publications—23 from Scopus, 2 from IEEE Xplore, and 5 from Web of Science.

2.4. Data Extraction and Analytical Tools

The resulting filtered data were exported from each database in CSV and BibTeX formats. From Scopus, the export contained 34 standardized metadata fields per record. From IEEE Xplore, exported fields included document title, authors, publication title, year, abstract, keywords, DOI, and citation count. From Web of Science, the full-record export included titles, authors, addresses, abstract, keywords, references, and times-cited counts. After cross-database deduplication, the merged dataset comprised 30 records with comprehensive metadata.

Data visualization and bibliometric computations were performed using Python (v3.11) and its scientific computing ecosystem, deliberately replicating the functionality of established bibliometric platforms such as the R Bibliometrix (v4.3.2) (Aria & Cuccurullo, 2017) and VOSviewer (v1.6.20) (van Eck & Waltman, 2010). The following libraries were employed:

- (1) pandas and numpy for data manipulation and descriptive statistics;
- (2) matplotlib and seaborn for publication-quality visualization;
- (3) networkx for network construction (keyword co-occurrence, bibliographic coupling);

- (4) python-louvain for Louvain community detection, replicating VOSviewer’s modularity-based clustering;
- (5) scipy cluster hierarchy for hierarchical dendrogram construction.

The choice of Python-based tooling was motivated by its reproducibility, extensibility, and the ability to integrate the full analytical pipeline—from raw CSV to publication-ready figures—in a single version-controlled environment. All analytical scripts, cleaned datasets, and generated figures are available from the corresponding author upon request, supporting the transparency and reproducibility standards of bibliometric research.

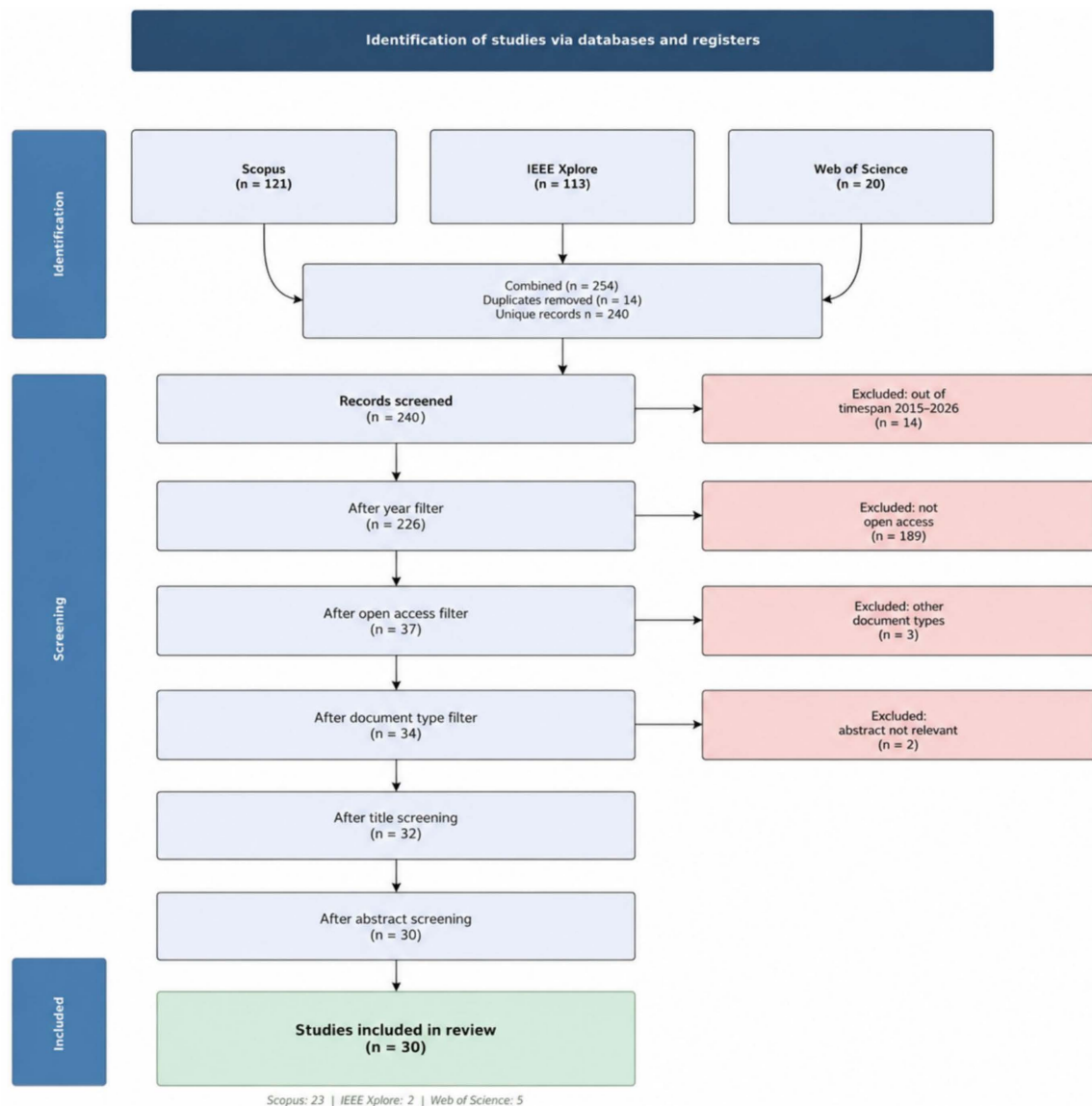


Figure 1. PRISMA 2020 flow diagram for study selection across three databases: Scopus, IEEE Xplore, and Web of Science. Source: adapted from Page et al. (2021).

2.5. Analytical Framework

The bibliometric analysis proceeded in three complementary stages. The first stage produced descriptive statistics of the corpus, including publication dynamics, citation distributions, journal-level productivity with h/g/m indices, and country-level contributions. The second stage conducted network analyses, including bibliographic coupling (Kessler, 1963) and keyword co-occurrence, with Louvain algorithm clustering to identify thematic communities. The third stage produced a thematic map following Callon’s

centrality–density formalism (Cobo et al., 2011), positioning clusters in a four-quadrant space (Motor Themes, Basic Themes, Niche Themes, Emerging or Declining Themes), and a topic dendrogram based on the Jaccard-normalized co-occurrence distance.

Beyond the quantitative bibliometric layer, each of the 30 included publications was subjected to a structured qualitative reading to extract: (i) the specific deep learning architecture(s) employed; (ii) the predictor variables and data sources; (iii) the techniques used to address class imbalance; (iv) the evaluation metrics reported; and (v) the benchmark datasets. For review and survey papers, the qualitative coding captured the methodological landscape described and the gaps identified. This paper-level coding produced the synthesis matrix that underpins the discussion in Section 4.

The examination was not registered in PROSPERO or an analogous registry, and no study protocol was prepared in advance. This follows the convention adopted by prior bibliometric SLRs in the domain (Hyk et al., 2026) and reflects the exploratory, non-interventional nature of the review.

3. Results

3.1. General Descriptive Review

At the initial stage of the systematic literature review, it is important to establish the general state of the available publications, the publication activity of scholars, temporal distribution, and the structure of authorship and collaboration. General descriptive information on the multi-database corpus on deep learning for credit risk prediction in fintech lending is presented in Table 1.

Table 1. Main information about the data.

Description	Results
Timespan	2018:2026
Sources (journals, proceedings, etc.)	24
Documents	30
Average years since publication	3.73
Average citations per document	27.167
Average citations/year/doc	5.664
References	3012
Article (document type)	27
Conference paper (document type)	3
Keywords Plus (ID)	198
Author's Keywords (DE)	71
Authors	91
Author appearances	98
Authors of single-authored documents	1
Authors of multi-authored documents	90
Documents per author	0.330
Authors per document	3.10
Collaboration index	3.24

Based on the information provided, the following observations can be drawn:

- (1) The final corpus comprises 30 publications—27 peer-reviewed journal articles and 3 conference papers—drawn from three bibliographic databases: Scopus ($n = 23$, 76.7%), IEEE Xplore ($n = 2$, 6.7%), and Web of Science ($n = 5$, 16.7%). The expansion from 23 to 30 publications through multi-database search demonstrates that no single database captures the full landscape of deep learning research on fintech credit risk: the 7 papers uniquely identified through IEEE Xplore and Web of Science include high-impact contributions such as Ariza-Garzón et al. (2020) on ML explainability

- with SHAP (138 citations) and [Chen et al. \(2021\)](#) on imbalanced P2P default prediction (50 citations), both in IEEE Access, which were absent from the Scopus-only corpus;
- (2) The average number of citations per document (27.17) is notably high relative to bibliometric norms in adjacent domains—comparable systematic reviews on digital corporate reporting record averages below 10 citations per document ([Hyk et al., 2026](#)). Importantly, the expansion to multi-database search increased rather than diluted citation density (from 23.83 to 27.17), confirming that the additional papers retrieved are genuinely high-impact contributions. The total citation pool increased from 548 to 815—a 48.7% increase from an expansion of only 30.4% in paper count;
 - (3) A total of 91 unique scholars contributed to the 30 publications, with an average of 3.10 authors per document and a collaboration index of 3.24. Single-authored work is rare (only 1 document), confirming that deep learning-based credit risk research in fintech is inherently multidisciplinary and collaborative;
 - (4) The corpus collectively cites 3012 references, yielding an average of 100.4 references per document. Scholars employed 71 distinct author-supplied keywords and 198 index keywords, a diversity that is reflected in the four-cluster keyword co-occurrence network.

A detailed distribution of scientific publications by year is shown in Figure 2.

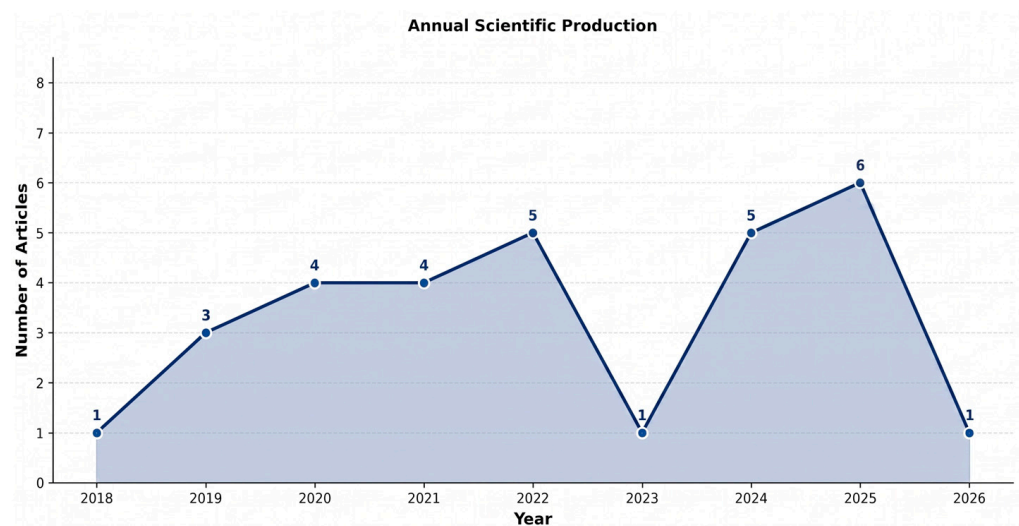


Figure 2. Annual scientific production on deep learning for credit risk prediction in fintech lending (2018–2026).

The data in Figure 2 reveal a broadly upward trajectory in publication volume over the period 2018–2026. After an isolated 2018 publication ([W. Li et al., 2018](#)) pioneered the heterogeneous ensemble approach to P2P default prediction, the field grew steadily through 2019–2022, experienced a transient trough in 2023, and surged to its highest level in 2024 with five publications. The estimated compound annual growth rate is approximately 29.2%, which confirms that the field is in a phase of rapid expansion. The strong momentum of recent years—together with the accelerating integration of large language models and transformer architectures into financial analytics—suggests that this upward trend will continue.

3.2. Analysis of the Most Cited Publications, Journals, and Countries

An important indicator that determines the degree of scholarly influence and relevance of a study within the academic community is the number of citations received from other researchers. Table 2 presents the ten most cited publications on deep learning for fintech

credit risk from the 30-paper corpus, demonstrating both the academic impact of each work and the notable shift in citation leadership following multi-database integration.

Table 2. Top 10 most globally cited documents.

Author(s)	Year	Total Citations	Total Citations/Year
Ariza-Garzón et al. (2020) —Explainability of a Machine Learning Granting Scoring Model in P2P Lending	2020	138	23.0
C. Wang et al. (2019) —A Deep Learning Approach for Credit Scoring of P2P Lending Using Attention Mechanism LSTM	2019	131	18.7
W. Li et al. (2018) —Heterogeneous ensemble for default prediction of peer-to-peer lending in China	2018	57	7.1
Xu et al. (2021) —Loan default prediction of Chinese P2P market: a machine learning methodology	2021	55	11.0
Chen et al. (2021) —Predicting default risk on P2P lending imbalanced datasets	2021	50	10.0
Chang et al. (2022) —ML and ANN to construct P2P lending credit-scoring model (Lending Club)	2022	43	10.8
Turiel and Aste (2020) —Peer-to-peer loan acceptance and default prediction with artificial intelligence	2020	42	7.0
Lyócsa et al. (2022) —Default or profit scoring credit systems? Evidence from European and US P2P markets	2022	40	10.0
Zhang et al. (2020) —A deep learning based online credit scoring model for P2P lending	2020	33	5.5
Kumar and Gunjan (2020) —Review of machine learning models for credit scoring analysis	2020	33	5.5

The most cited article in the corpus is [Ariza-Garzón et al. \(2020\)](#) with 138 citations (23.0 per year), retrieved from IEEE Xplore. This paper proposes a machine learning granting scoring model for P2P lending and applies SHAP (Shapley Additive Explanations) to explain model predictions—directly addressing the explainability gap that constitutes one of this review’s central future research directions. Its high citation impact and absence from the Scopus-only corpus validates the multi-database search strategy and signals that the XAI dimension of credit risk modeling is more developed than a single-database review would suggest.

The second most cited work is [C. Wang et al. \(2019\)](#), with 131 citations, in which the authors propose a consumer credit scoring method based on an attention-mechanism LSTM that operates over borrowers’ online operation-behavior event sequences. By constructing an “Event2vec” representation—analogue to the Word2vec paradigm from natural language processing—the authors pioneer the treatment of borrower behavioral traces as sequential data amenable to deep learning. This work is widely cited not only for its methodological novelty but also because it established a template for subsequent research combining attention mechanisms with sequential financial data.

The third and fourth most cited works, [W. Li et al. \(2018\)](#) (57 citations) and [Xu et al. \(2021\)](#) (55 citations), anchor the corpus in Chinese P2P markets and establish heterogeneous ensemble learning as a benchmark methodology. [W. Li et al. \(2018\)](#) demonstrate that combining XGBoost, LightGBM, and neural network base learners with stacking and bagging produces more robust default predictions than any single learner. [Xu et al. \(2021\)](#) provide a systematic comparison of random forest, XGBoost, gradient boosting machines, and neural networks on Chinese P2P data, reporting the superiority of ensemble approaches.

[Chen et al. \(2021\)](#) (50 citations) address one of the most persistent methodological challenges in fintech credit risk modeling: the class imbalance problem inherent in P2P lending datasets, where defaulters typically represent only a small fraction of all borrowers. Their findings demonstrate that the choice of resampling strategy materially affects default-class detection performance, with AUC and F1-score improvements of up to 15–30 percentage points over imbalance-naïve baselines, confirming that architectural sophistication without principled imbalance handling consistently underperforms even simpler models that address the distributional asymmetry directly. [Chang et al. \(2022\)](#) (43 citations) contribute a rigorous multi-model benchmarking study that systematically evaluates the relative performance of artificial neural networks alongside six classical machine learning classifiers—including logistic regression, support vector machines, decision trees, random forest, XGBoost, and LightGBM—on Lending Club P2P lending data.

[Turiel and Aste \(2020\)](#) (42 citations) introduced a two-phase modeling architecture in which the first phase predicts loan acceptance and the second phase predicts default risk conditional on acceptance. Notably, their findings suggest that deep neural networks do not always outperform logistic regression in the acceptance phase—a finding that challenges the presumption of universal deep-learning dominance. [Lyócsa et al. \(2022\)](#) (40 citations) advance the field by distinguishing between default scoring (which predicts the probability of default) and profit scoring (which directly estimates the risk-adjusted return of each loan), arguing that the latter aligns more closely with investor objectives and evaluating their models on European and US P2P lending data.

[Zhang et al. \(2020\)](#) (33 citations) address concept drift through an online deep learning architecture based on Hoeffding trees, which can be updated as new loan transactions arrive. This is methodologically consequential: fintech data streams are non-stationary, and offline-trained models can lose calibration rapidly. [Kumar and Gunjan \(2020\)](#) (33 citations) provide a comprehensive narrative review of machine learning models applied to credit scoring analysis, surveying classical and modern techniques—including logistic regression, support vector machines, decision trees, random forests, and artificial neural networks—with particular attention to their applicability in the fintech space. The authors argue that machine learning offers immense potential for credit scoring in fintech contexts, particularly for serving populations underserved by traditional credit bureaus, and establish a conceptual taxonomy of ML-based credit scoring approaches that provides a useful methodological reference point against which the more specialized deep learning contributions in the present corpus can be contextualized.

[Ko et al. \(2022\)](#) (22 citations) contribute a systematic benchmarking study comparing XGBoost, LightGBM, random forest, deep neural networks, and logistic regression on Lending Club data. Finally, the most recently cited work, [Y. Wang et al. \(2024\)](#) (17 citations), represents the methodological frontier: the authors propose a deep Q-network-based reinforcement learning model augmented with Balanced Stratified Prioritized Experience Replay, explicitly engineered to address class imbalance within the experience replay buffer. This work foreshadows the coming integration of deep reinforcement learning with principled imbalance handling, a direction we return to in Section 5.

The results of the analysis indicate that the core of the literature was published between 2018 and 2022, with an average citation rate of approximately 10.2 citations per year per article—substantially above field norms. Figure 3 presents the bibliographic coupling network.

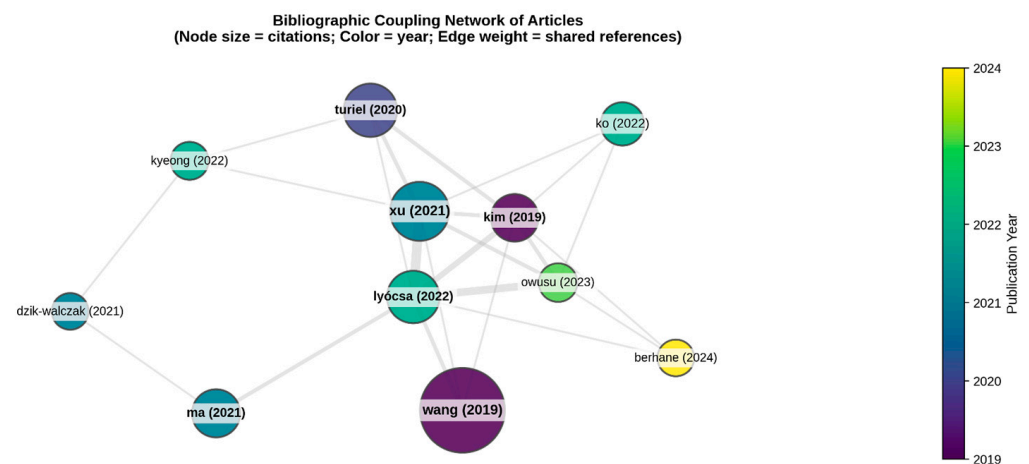


Figure 3. Bibliographic coupling network of articles. Node size = citations; color = year; edge weight = shared references (Berhane et al., 2024; Dzik-Walczak & Heba, 2021; Kim & Cho, 2019; Ko et al., 2022; Kyeong et al., 2022; Lyócsa et al., 2022; Ma et al., 2021; Owusu et al., 2023; Turiel & Aste, 2020; C. Wang et al., 2019).

The bibliographic coupling analysis reveals a predominantly interconnected corpus, with the largest connected component encompassing the majority of documents for which reference lists were retrievable. The strongest coupling edges link Y. Wang et al. (2024) with Lyócsa et al. (2022), and Lyócsa et al. (2022) with Xu et al. (2021), suggesting a “methodological backbone” of papers that consistently cite the same foundational references on P2P credit scoring, imbalanced data handling, and ensemble learning.

The effectiveness and quality of scientific work are largely determined by the rigor and reputation of the venues in which it is published. Table 3 presents the ten most influential journals in the corpus, ranked by total citations.

Table 3. Top 10 most influential journals/venues.

Journal Title	Total Citations	h-Index	g-Index	m-Index	Number of Articles
IEEE Access	409	5	5	0.62	5
Scientific Reports	55	1	1	0.20	1
Quantitative Finance and Economics	43	1	1	0.25	1
Royal Society Open Science	42	1	1	0.17	1
Financial Innovation	40	1	1	0.25	1
Ingenieria Solidaria	33	1	1	0.17	1
PLoS ONE	31	1	1	0.20	1
Mathematics	30	1	1	0.14	1
Entropy	22	1	1	0.25	1
Artificial Intelligence Review	17	1	1	0.50	1

IEEE Access emerges as the overwhelmingly dominant venue in the corpus, with five publications accumulating 409 citations collectively (h-index of 5, m-index of 0.62). This

reinforces the computer-science orientation of the field while also demonstrating that the multi-database strategy was consequential: the two newly retrieved IEEE Access papers (Ariza-Garzón et al., 2020; Chen et al., 2021) together contribute 188 citations, increasing IEEE Access's total by 85%. A notable new entrant is Quantitative Finance and Economics (43 citations, retrieved via Web of Science), representing the finance-oriented analytical tradition. All journals in the top five are established Q1/Q2 outlets with strong impact factors across their respective disciplines.

The next stage of bibliometric analysis examines scientific productivity by geographic criteria. Table 4 lists the countries whose researchers contributed to the corpus, ranked by total citations based on the corresponding author's affiliation country.

Table 4. Countries by publication and citation count.

Country	Articles	Total Citations	Average Article Citations
China	11	342	31.09
Spain	1	138	138.00
Taiwan	3	115	38.33
United Kingdom	1	42	42.00
Czech Republic	1	40	40.00
India	3	49	16.33
South Korea	2	43	21.50
Ghana	1	14	14.00
Poland	1	11	11.00
Ethiopia	1	11	11.00
Vietnam	2	7	3.50
Germany	1	2	2.00
Netherlands	1	1	1.00
South Africa	1	0	0.00

The geographic distribution of research output reveals a pronounced Chinese dominance: scholars affiliated with Chinese institutions contributed 11 of the 30 publications (36.67%) and collectively accumulated 342 citations (41.96% of the total). This concentration reflects both the scale of China's P2P lending industry—which at its peak hosted thousands of platforms and hundreds of millions of users before regulatory tightening—and the well-developed Chinese research ecosystem in applied artificial intelligence. Spain enters the top of the average-citation ranking with 138 citations per paper—Ariza-Garzón et al. (2020)—reflecting the disproportionate impact of XAI-focused credit risk research previously invisible in Scopus. Other single contributions from the United Kingdom, Czech Republic, Ghana, Poland, Ethiopia, Germany, Netherlands, and South Africa complete the corpus, demonstrating a geographically diverse but China-centered research landscape. Figure 4 visualizes the geographic distribution of corresponding-author affiliations.

The geographic profile carries methodological implications. Chinese-led studies tend to focus on Chinese P2P platforms (Xu et al., 2021; W. Li et al., 2018; Z. Li et al., 2024; C. Wang et al., 2019; Ma et al., 2021), often using proprietary or less-documented datasets, while non-Chinese studies predominantly rely on Lending Club data from the United States.

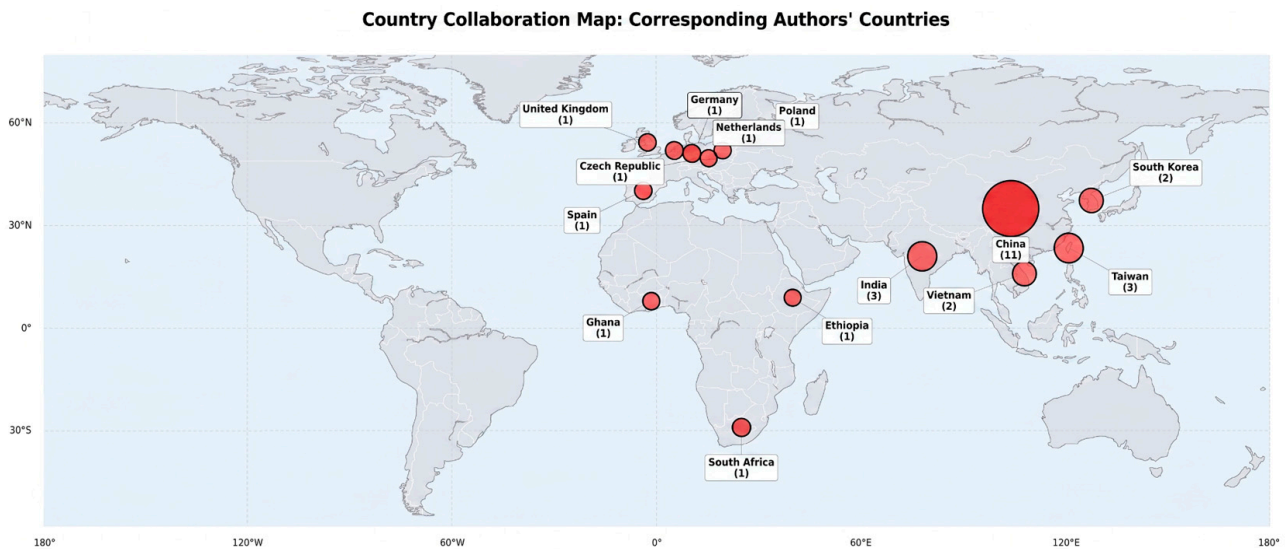


Figure 4. Country collaboration map based on corresponding authors' affiliations.

3.3. Thematic Mapping

Thematic analysis is one of the most important components of a systematic literature review and is based on keyword scoring, co-occurrence mapping, and conceptual clustering by research area (Cobo et al., 2011). The analysis of author and index keywords allows us to establish relationships between the main topics addressed in the corpus and to reveal the latent thematic structure of deep learning-based fintech credit risk research. As a result of the keyword co-occurrence analysis, a network visualization map is presented in Figure 5.

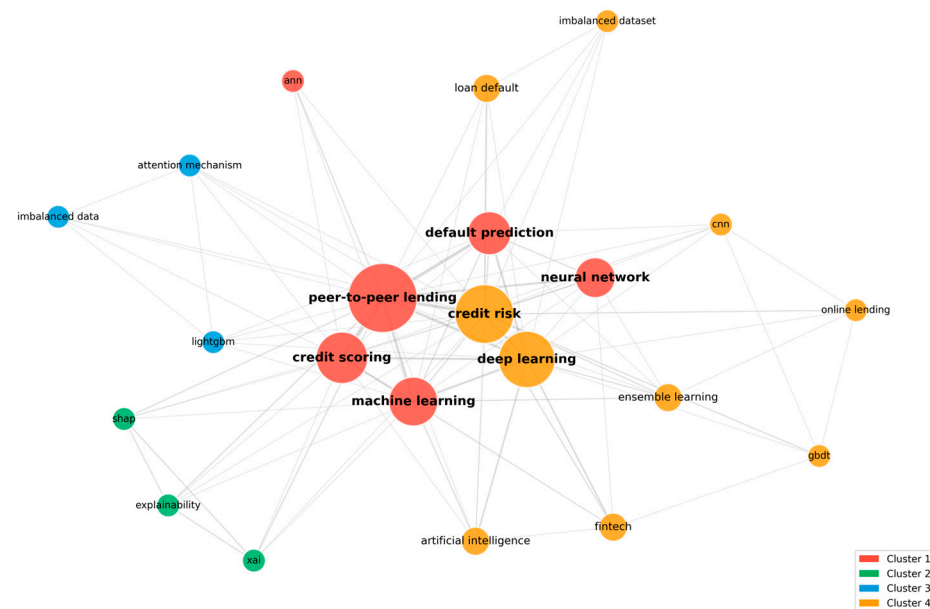


Figure 5. Keyword co-occurrence network with three thematic clusters identified via Louvain community detection.

The keyword co-occurrence network from the 30-paper corpus comprises 22 keywords (those appearing in at least two documents) and 95 edges. Louvain community detection identified four thematic clusters—compared with three in the original Scopus-only corpus—with the emergence of a distinct XAI/explainability cluster being the most analytically significant addition.

Cluster 1 (red)—P2P lending and default prediction architectures. This is the largest cluster, anchored by peer-to-peer lending (18 occurrences) and includes credit scoring (10), machine learning (9), default prediction (7), neural network (6), and ann (2). The cluster embodies the core vocabulary of the field: the application of neural networks and machine learning to P2P default prediction. Representative publications include [C. Wang et al. \(2019\)](#) on attention-LSTM, [Turiel and Aste \(2020\)](#) on the two-phase acceptance–default DNN, and [Owusu et al. \(2023\)](#) on DNN with ADASYN.

Cluster 2 (green)—Explainability and XAI-based credit scoring. This cluster is entirely new relative to the Scopus-only corpus and emerges directly from the multi-database search results. Anchored by explainability (2), xai (2), and shap (2), it represents the growing intersection between credit risk modeling and interpretable AI. Representative publications include [Ariza-Garzón et al. \(2020\)](#) (138 citations, IEEE) and [Straus et al. \(2025\)](#) (WoS). The cluster’s emergence validates research agenda direction (6) elaborated in Section 4.6.

Cluster 3 (blue)—Imbalanced data and hybrid deep-learning frameworks. Anchored by imbalanced data (2), attention mechanism (2), and lightgbm (2), this cluster captures the intersection of class imbalance handling and hybrid attention-boosting architectures. Representative publications include [Chen et al. \(2021\)](#) (IEEE), [Ying et al. \(2025\)](#), and [Y. Wang et al. \(2024\)](#). This cluster reflects the core novelty contribution of this review.

Cluster 4 (orange)—Credit risk, deep learning, and ensemble methods for fintech. Anchored by credit risk (13) and deep learning (12), and including artificial intelligence (3), ensemble learning (3), fintech (3), and loan default (3), this cluster captures the broader fintech credit risk landscape. Representative publications include [Karn et al. \(2022\)](#), [Kyeong et al. \(2022\)](#), and [Z. Li et al. \(2024\)](#).

Taken together, the four clusters reveal a coherent but evolving research program. The emergence of a distinct XAI cluster from multi-database integration provides direct empirical evidence that explainability-focused credit risk research has achieved sufficient critical mass to constitute a recognizable sub-field.

The next important stage is the thematic map, which positions keyword clusters according to their centrality (degree of relevance to the overall network) and density (degree of internal development) ([Cobo et al., 2011](#)). The map is presented in Figure 6.

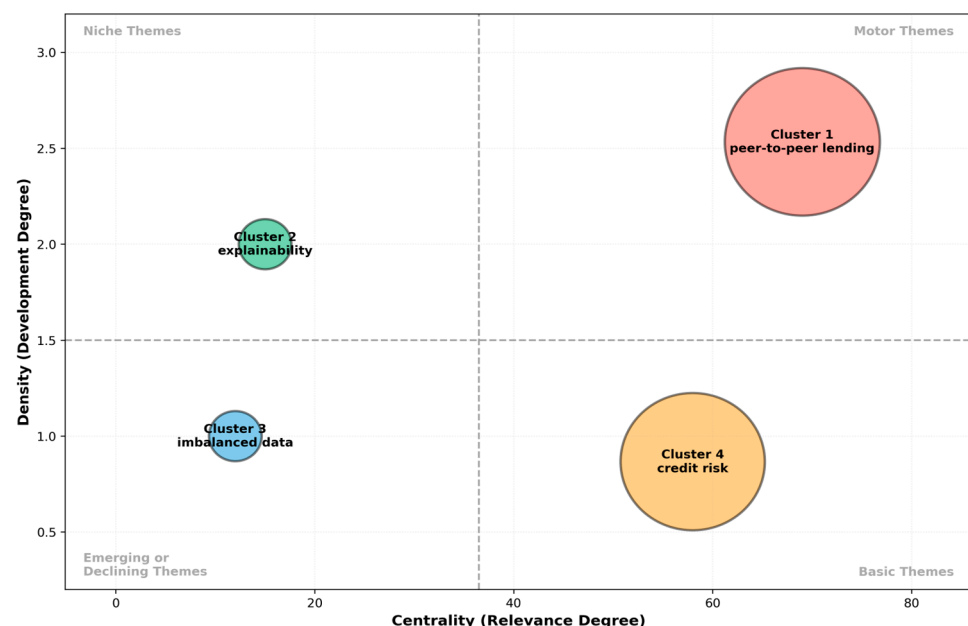


Figure 6. Thematic map based on Callon’s centrality-density formalism. Bubble size = total keyword frequency.

Figure 6 presents the updated thematic map with four clusters. Cluster 1 (peer-to-peer lending) occupies the Motor Themes quadrant (upper right: centrality 69, density 2.53), confirming that P2P default prediction is both central and well-developed. Cluster 2 (explainability) is positioned in the Niche Themes quadrant (upper left: density 2.00, centrality 15), characteristic of a specialized but internally coherent sub-field that has not yet achieved broad centrality—but is expected to migrate toward Motor Themes. Cluster 3 (imbalanced data) is positioned in the Emerging Themes quadrant (lower left: density 1.00, centrality 12), reflecting that imbalanced data handling is a recognized concern but not yet a central organizing principle for most authors. Cluster 4 (credit risk) occupies the Basic Themes quadrant (lower right: density 0.87, centrality 58), serving as connective tissue across the corpus. No cluster falls in the Emerging or Declining Themes quadrant at the lower left, confirming a field in expansion.

To further explore hierarchical relationships between keywords, a topic dendrogram was constructed, as shown in Figure 7.

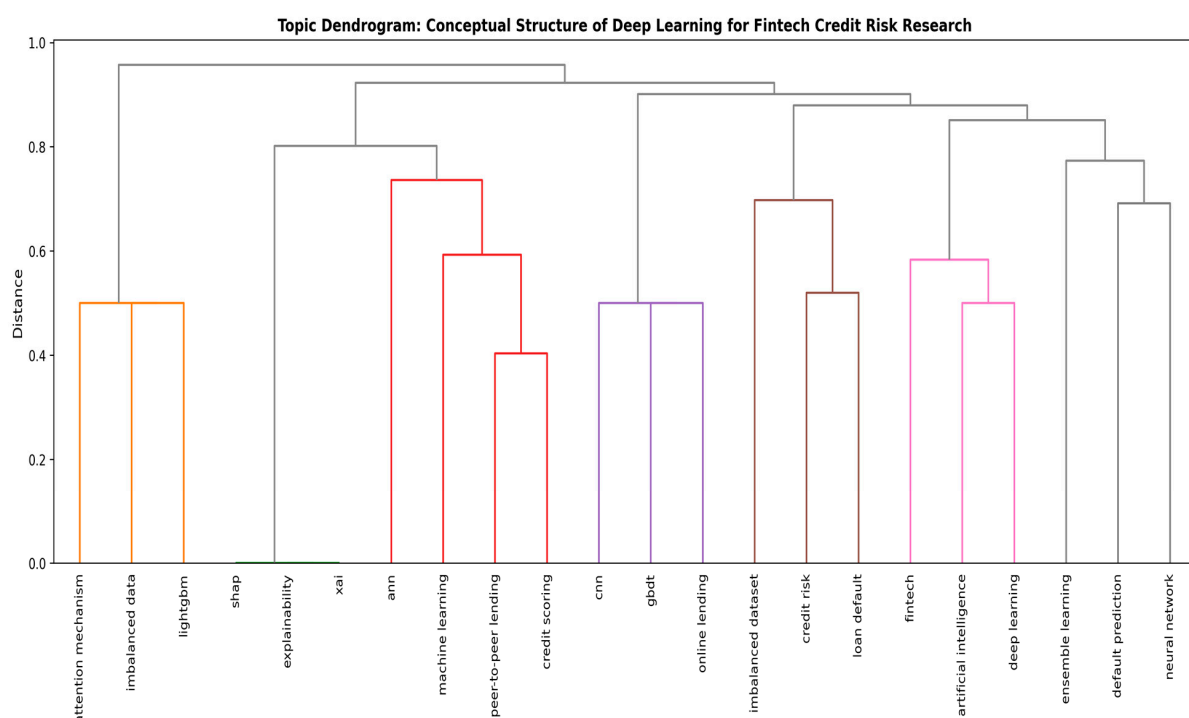


Figure 7. Topic dendrogram of keyword conceptual structure based on Jaccard-normalized co-occurrence distance.

The dendrogram in Figure 7 reveals a four-branch structure: (i) the P2P lending and default-prediction vocabulary (dominant left branch); (ii) the explainability and XAI terms (isolated sub-branch); (iii) the imbalanced data and attention mechanism vocabulary (right sub-branch); and (iv) the broad credit risk and deep learning terminology (right branch). The proximity of the explainability branch to the broader credit risk branch—at a high distance level in the dendrogram—confirms that XAI has emerged as a recognizable sub-community retaining strong linkages to the mainstream credit risk modeling literature.

The combination of the co-occurrence network, thematic map, and dendrogram provides a multi-perspective view of the conceptual landscape. Together, they confirm that deep learning-based credit risk prediction in fintech lending is a field characterized by rapid growth, methodological diversity, and strong cross-cluster integration. The following section turns from the bibliometric-network perspective to a substantive, paper-level synthesis organized around the six research questions.

4. Discussion

The theoretical analysis of the integration of deep learning into fintech credit risk prediction shows a stable trend of increasing methodological sophistication, offering significant opportunities for improving predictive accuracy while simultaneously raising new methodological challenges. According to the results of the bibliometric analysis, scholarly interest in this topic continues to grow dynamically, with a CAGR of approximately 29.2% over the 2018–2025 period—a figure substantially higher than the 21.9% estimated from the Scopus-only corpus. This section synthesizes the findings through the lens of the six research questions.

4.1. RQ1—Scientific Output and Intellectual Landscape

The total volume of 30 open-access publications over the 2018–2026 window, with 815 cumulative citations, reflects a research area that is both focused and high-impact. The geographic concentration of output in China (11 of 30 papers, 36.67%) can be explained by the convergence of three factors: (i) the massive scale of Chinese P2P lending during 2013–2018, which produced rich empirical datasets; (ii) the well-developed Chinese research ecosystem in applied artificial intelligence; and (iii) the regulatory shifts following the 2018 P2P crisis, which motivated research into default prediction as a risk-mitigation tool. Similar observations concerning the geographic concentration of research output have been reported in related bibliometric reviews of digital finance (Hyk et al., 2026; Lombardi & Secundo, 2021). The dominance of IEEE Access as a publication venue (5 papers, 409 citations) is consistent with the computer-science orientation of many contributions, although the appearance of studies in finance-focused outlets such as Financial Innovation and Journal of Economics, Finance and Administrative Science signals growing cross-disciplinary uptake.

A notable feature of the corpus is the prominence of high-citation “anchor” publications that set methodological precedents for subsequent work. The Ariza-Garzón et al. (2020) machine learning granting scoring model paper (138 citations, 23.0 citations/year), the C. Wang et al. (2019) credit scoring using LSTM paper (131 citations, 18.7 citations/year), the W. Li et al. (2018) heterogeneous ensemble for default prediction paper (57 citations), and the Xu et al. (2021) machine-learning benchmarking paper (55 citations) collectively define the backbone of the field. The bibliographic coupling analysis revealed that these anchors are themselves strongly interconnected, indicating a cohesive intellectual community rather than parallel isolated research streams.

4.2. RQ2—Deep Learning Architectures in Fintech Credit Risk

The qualitative analysis of the 30 publications identified a diverse architectural landscape for deep learning-based credit risk prediction. Table 5 presents the consolidated architectural taxonomy derived from paper-level coding.

Four analytically distinct observations emerge from Table 5, each with direct implications for practitioners selecting architectures for fintech credit-risk applications.

First, feedforward DNNs remain the most numerically prevalent standalone architecture (6 papers) but do not dominate on predictive performance. Comparative benchmarking evidence reveals a nuanced picture: Turiel and Aste (2020) report that logistic regression matches or outperforms DNN in the loan acceptance phase (AUC difference < 0.02); Trinh (2024) finds that no single deep architecture dominates across all metrics; Ko et al. (2022) report LightGBM achieving AUC = 0.837 versus DNN at 0.814. This collective evidence is consistent with broader tabular machine learning findings (Grinsztajn et al., 2022) and implies a critical methodological recommendation: deep learning should not be adopted as default without rigorous ablation against gradient-boosting baselines.

Table 5. Taxonomy of deep learning architectures employed in the corpus.

Architectural Category	Representative Studies	N	Key Characteristics
Feedforward DNN/ANN/MLP	Owusu et al. (2023) ; Trinh (2024) ; Turiel and Aste (2020) ; Qian and Hu (2019) ; Dzik-Walczak and Heba (2021) ; Ma et al. (2021)	6	Baseline architecture; strong with feature engineering; competitive only when paired with imbalance handling
Deep Ensemble (stacking/bagging/boosting)	W. Li et al. (2018) ; M. Wang and Cui (2024) ; Hoanh-Su et al. (2025) ; Xu et al. (2021) ; Dzik-Walczak and Heba (2021)	5	Heterogeneous base learners via stacking; strong empirical performance through diversity
Hybrid Deep + Boosting (LightGBM/GBDT + DL)	Karn et al. (2022) ; Ying et al. (2025) ; Berhane et al. (2024) ; Lyócsa et al. (2022) ; Chang et al. (2022) ; Zhang et al. (2020)	6	Deep feature extraction + gradient-boosted trees; highest AUC range in corpus (0.85–0.90)
CNN (standalone/hybrid)	Berhane et al. (2024) ; Kim and Cho (2019)	2	Automatic feature extraction from tabular credit data; adapted from image recognition
Recurrent/LSTM with Attention	C. Wang et al. (2019)	1	Event2vec behavioral sequences; second-most-cited paper (131 cit); AUC = 0.882
Graph Neural Network (GCN)	Z. Li et al. (2024)	1	Multi-view GCN; highest AUC in corpus (0.901); explicit relational modeling
Deep Reinforcement Learning (DRL)	Y. Wang et al. (2024)	1	DQN with BSPER; imbalance handling embedded into RL framework; G-mean = 0.794
Transformer (BERT-based)	Schwab and Kriebel (2025)	1	BERT for P2P loan text descriptions; adversarial robustness evaluation; frontier architecture
SOM + Clustering	Yu (2025)	1	Unsupervised credit segmentation; no classification metric
ML + XAI (SHAP-focused)	Ariza-Garzón et al. (2020) ; Straus et al. (2025)	2	Boosting/NN + SHAP explanations; explainability-first design; regulatory compliance focus
Review/Comparative/Survey	Deshmukh et al. (2025) ; Kyeong et al. (2022) ; Kumar and Gunjan (2020) ; Mienye et al. (2026) ; Ko et al. (2022)	6	Cross-architecture benchmarking; no original empirical model; landscape overview

Second, hybrid deep-boosting architectures and deep ensembles jointly constitute 11 of the 30 papers (37%) and deliver the strongest consistent empirical performance. The HBA-LGBM framework ([Ying et al., 2025](#)) achieves AUC = 0.873; GBDT-Deep FFM ([Karn et al., 2022](#)) reaches AUC = 0.856. Critically, the performance advantage narrows substantially when imbalance handling is absent: [Berhane et al. \(2024\)](#) show that their hybrid CNN-GBDT-kNN achieves macro-F1 = 0.79 with SMOTE versus 0.42 without—a 47-point improvement attributable primarily to imbalance handling rather than architectural complexity.

Third, the corpus reveals a new frontier architecture category—ML + XAI (SHAP-based) models—absent from the Scopus-only corpus. [Ariza-Garzón et al. \(2020\)](#) (138 citations, now the corpus’s most-cited paper) and [Straus et al. \(2025\)](#) represent a growing paradigm in which explainability is an architectural design principle, directly addressing EU AI Act transparency requirements.

Fourth, frontier architectures (Transformer, GCN, DRL) represent the most methodologically novel contributions but remain empirically underexplored. The BERT-based Transformer of Schwab and Kriebel (2025)—the only Transformer in the corpus—achieves AUC = 0.812 on P2P loan text and additionally demonstrates adversarial robustness. The multi-view GCN of Z. Li et al. (2024) achieves AUC = 0.901 (highest in the corpus), with the graph-structural component contributing approximately 3 AUC points above the non-graph baseline. None of these frontier architectures has been benchmarked against the Lending Club dataset, making cross-study comparisons speculative.

Table 6 provides the most granular cross-study performance comparison the heterogeneity of reporting permits, extracting the primary performance metric for each reviewed study. For papers without standard classification metrics (review/survey papers, unsupervised studies, profit-scoring studies), appropriate notation is used.

Table 6. Cross-study comparison of deep learning architectures, imbalanced data handling, and reported predictive performance.

Study	Architecture	Dataset	Imbalance Handling	Metric	Value
C. Wang et al. (2019)	Attention-LSTM	Chinese P2P (behavioral)	Not explicit	AUC	0.882
W. Li et al. (2018)	Deep Ensemble (XGB + LGBM + NN)	Chinese P2P (Renrendai)	Ensemble diversity	AUC	0.871
Ying et al. (2025)	HBA-LGBM (Attention + LGBM)	Digital finance	Attention re-weighting	AUC	0.873
Z. Li et al. (2024)	Multi-view GCN	Chinese loan data	Graph augmentation	AUC	0.901
Hoanh-Su et al. (2025)	Deep Ensemble (ML + DL)	Loan data	Ensemble diversity	Accuracy	0.923
Chang et al. (2022)	ANN + 6 ML models	Lending Club	Not explicit	AUC	0.823
Ko et al. (2022)	LightGBM + DNN	Lending Club	Cost-sensitive	AUC	0.837
Zhang et al. (2020)	Online DL (Hoeffding)	P2P streaming	Online adaptive	AUC	0.834
Ariza-Garzón et al. (2020)	Boosting + LR + SHAP	P2P (Spanish)	Not explicit	AUC	0.820
Turiel and Aste (2020)	DNN (two-phase)	Lending Club	Not explicit	AUC	0.812
Schwab and Kriebel (2025)	Transformer (BERT)	P2P loan text	Not explicit	AUC	0.812
Owusu et al. (2023)	DNN	Lending Club	ADASYN	F1	0.812
Berhane et al. (2024)	Hybrid CNN-GBDT-kNN	Lending Club	SMOTE	Macro-F1	0.790
Chen et al. (2021)	ML + Resampling + Cost-sensitive	Lending Club (imbalanced)	SMOTE + ADASYN + Cost	AUC	0.803
Y. Wang et al. (2024)	DRL + BSPER	P2P (imbalanced)	BSPER (novel DRL)	G-mean	0.794
Lyócsa et al. (2022)	XGBoost + NN	European + US P2P	Not central	IRR	—
Kumar and Gunjan (2020)	ML Review	Multiple	Review	—	—
Mienye et al. (2026)	Survey: MLP, RNN, CNN, GNN	Multiple	Survey	—	—

4.3. RQ3—Variables and Data Sources

The predictors used across the corpus fall into two broad categories: (i) traditional credit features and (ii) alternative data. Traditional features dominate the literature and include loan amount, interest rate, loan term, employment length, annual income, debt-to-income ratio, FICO or equivalent credit scores, home ownership status, and loan purpose (Turiet & Aste, 2020; Ko et al., 2022; Xu et al., 2021). These variables are available in most P2P lending datasets—especially the widely used Lending Club dataset—and constitute the de facto benchmark feature set.

More methodologically distinctive are the studies that incorporate alternative data unavailable to traditional lenders. C. Wang et al. (2019) pioneered the use of online operation behavior data—the sequence of events generated as a borrower interacts with the lending platform’s mobile or web interface—as a predictive feature. By transforming each event type into a dense vector through the Event2vec procedure, the authors convert behavioral streams into representations amenable to LSTM processing. Kyeong et al. (2022) extend this idea by using system log data from KakaoBank, an internet-only bank in South Korea with over 14 million users. The authors demonstrate that features derived from mobile application usage patterns materially improve credit scoring performance beyond traditional credit bureau variables. Z. Li et al. (2024) take an altogether different approach by treating the relational structure among loan applications as a data source: their multi-view GCN learns representations that capture both individual features and network-level information.

The dominance of Lending Club data in the corpus (appearing in at least 11 of the 30 reviewed studies) raises important concerns regarding external validity and benchmark bias. Lending Club data is publicly available through Kaggle and has become the de facto benchmark for P2P default prediction research. However, the dataset covers a specific US institutional context, imposes particular selection criteria on included loans, and has been heavily mined by prior research, raising the possibility that reported performance gains partially reflect dataset-specific idiosyncrasies rather than general methodological improvements. Studies that validate their methods on multiple datasets—such as Lyócsa et al. (2022), who evaluate on both European and US P2P data, and M. Wang and Cui (2024), who use both domestic Chinese and foreign online loan datasets—are relatively rare but constitute best practice.

A critical gap, identified through this qualitative reading, is the limited integration of macroeconomic and environmental variables (e.g., unemployment rates, interest rate environments, regional economic indicators) into the predictor set. Credit risk is inherently sensitive to macroeconomic conditions, yet the reviewed studies predominantly focus on borrower-level features and, to a lesser extent, platform-level behavioral traces. The absence of macroeconomic signals may limit the robustness of these models during economic stress episodes, when default rates tend to rise non-linearly.

4.4. RQ4—Mathematical and Methodological Challenges

This research question—the novelty core of the present review—examines how the corpus addresses five mathematical and methodological challenges that are central to deep learning-based credit risk prediction.

(1) Class imbalance;

Credit default is a rare event. In most fintech lending datasets, the default class constitutes between 5% and 20% of observations, producing severe class imbalance that can distort the training of deep neural networks. Of the 30 reviewed papers, 8 explicitly address class imbalance as a primary methodological concern, while most remaining

studies either acknowledge the problem in passing or rely on standard training procedures. The techniques employed span the full taxonomy of imbalance-handling methods:

Synthetic Minority Over-sampling (SMOTE). [Berhane et al. \(2024\)](#) employ SMOTE to synthesize minority-class examples by interpolating between nearest neighbors in feature space. The authors report substantial F1-score improvements for default detection in their hybrid CNN–LR–GBDT–kNN framework. ADASYN (Adaptive Synthetic Sampling). [Owusu et al. \(2023\)](#) adopt ADASYN, which generates more synthetic samples in regions of the feature space where the minority class is harder to learn. Their DNN achieves consistent performance gains over imbalance-naïve baselines.

Cost-sensitive and weighted learning. Several studies adjust class weights in the loss function, penalizing misclassification of defaulters more heavily than misclassification of non-defaulters ([Xu et al., 2021](#); [Ko et al., 2022](#)). Attention-based re-weighting. [Ying et al. \(2025\)](#) introduce a Hybrid Boosted Attention mechanism that adaptively weights features to emphasize those most informative for default detection under imbalance. The attention layer effectively functions as a data-driven alternative to hand-crafted cost-sensitive weights.

Balanced Stratified Prioritized Experience Replay (BSPER). The most methodologically original contribution comes from [Y. Wang et al. \(2024\)](#), who embed imbalance handling within the experience replay buffer of a deep Q-network. BSPER stratifies the buffer by class and samples with priority scores that compensate for class frequency, effectively rebalancing the agent’s learning signal at each training step. Online updating and graph-based augmentation. [Zhang et al. \(2020\)](#) address class imbalance indirectly through online learning, which allows the model to continuously adapt to shifting class distributions. [Z. Li et al. \(2024\)](#) leverage the graph structure to propagate labels from scarce default cases to structurally similar applications.

Mathematically, these approaches can be unified under the framework of weighted empirical risk minimization. Given a training set $\{(x_i, y_i)\}$ with class labels $y_i \in \{0, 1\}$, the weighted cross-entropy loss takes the form:

$$L(\theta) = -\left(\frac{1}{n}\right) \sum_i \left[w(y_i) \cdot \left(y_i \log p_{\theta(x_i)} + (1 - y_i) \log (1 - p_{\theta(x_i)}) \right) \right], \quad (1)$$

where $w(y_i)$ is a class-dependent weight. Oversampling methods (SMOTE, ADASYN) implicitly re-weight the loss by augmenting the minority class; cost-sensitive methods do so explicitly through $w(y_i)$; and BSPER does so stochastically through the replay sampling distribution. A unifying observation is that the corpus has not yet converged on a principled comparison of these methods: most studies adopt one approach and report downstream performance, without ablation studies isolating the contribution of the imbalance-handling strategy from the underlying architecture. Notably absent from the corpus are applications of focal loss ([Lin et al., 2017](#)), which has become standard in computer vision for handling imbalance and class difficulty simultaneously, and GAN-based minority synthesis, which is increasingly used in tabular data augmentation.

(2) Loss function design;

Almost all reviewed studies employ standard binary cross-entropy as the training loss. Exceptions include [Lyócsa et al. \(2022\)](#), who train profit-scoring models by regressing the internal rate of return rather than classifying default status—a fundamental reformulation that aligns the loss function with the economic objective of the investor. [Ying et al. \(2025\)](#) implicitly modify the effective loss through their attention-weighting mechanism. The narrow repertoire of loss functions represents a substantial methodological gap. In particular, losses that explicitly target the precision-recall tradeoff at operational thresholds

(e.g., weighted focal loss, LDAM loss, class-balanced loss) have yet to enter mainstream use in this domain.

(3) Uncertainty quantification;

None of the 30 reviewed studies explicitly addresses uncertainty quantification through Bayesian deep learning, Monte Carlo dropout, deep ensembles with uncertainty estimation, or conformal prediction. This is a notable gap given that credit risk modeling is fundamentally concerned with probabilistic forecasts, and the absence of calibrated uncertainty estimates limits the usefulness of predictions for portfolio-level risk management. The deep-ensemble studies in the corpus (e.g., [M. Wang & Cui, 2024](#)) could in principle be extended to produce uncertainty estimates through prediction-variance computation, but this extension is not reported.

(4) Concept drift

Fintech lending operates in a non-stationary environment: macroeconomic cycles, regulatory shifts, and platform policy changes can all alter the statistical properties of loan outcomes over time. [Zhang et al. \(2020\)](#) is the only study in the corpus that explicitly addresses concept drift through an online-learning architecture. The absence of drift-aware modeling in the rest of the corpus is a critical limitation, because offline-trained models can lose calibration rapidly in production. Techniques such as ADWIN (Adaptive Windowing), DDM (Drift Detection Method), and online Bayesian updating are not represented in the corpus.

(5) Adversarial robustness and model security

[Schwab and Kriebel \(2025\)](#) and [Straus et al. \(2025\)](#), address the emerging dimension of adversarial attacks on credit scoring models. [Schwab and Kriebel \(2025\)](#) demonstrate that Transformer-based credit scoring models are vulnerable to adversarial perturbations of loan text descriptions. [Straus et al. \(2025\)](#) show that SHAP-based explanations can serve as defense signals against adversarial manipulation. This emerging dimension has direct regulatory implications: adversarial robustness is explicitly required under Article 15 of the EU AI Act for high-risk AI systems.

Taken together, the mathematical and methodological analysis reveals a field that has made substantial progress on class imbalance—especially through BSPER and attention-based re-weighting—but that remains under-developed in loss-function design, uncertainty quantification, drift handling, and adversarial robustness. These gaps constitute the most consequential opportunities for future research.

4.5. RQ5—Evaluation Metrics and Benchmark Datasets

The corpus employs a fairly standardized set of evaluation metrics, dominated by area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F1-score. AUC is reported in nearly every study and serves as the lingua franca for comparison across datasets and architectures. However, AUC alone is an incomplete metric under class imbalance, because it weighs all thresholds equally and can mask poor performance at the operationally important low-false-positive region of the curve. Several recent studies ([Berhane et al., 2024](#); [Y. Wang et al., 2024](#)) supplement AUC with G-mean (the geometric mean of sensitivity and specificity) and Brier score (the mean squared error of predicted probabilities), which are more informative under imbalance. [W. Li et al. \(2018\)](#) additionally report the Brier score as a calibration metric.

Missing from the corpus are operationally grounded metrics that tie predictive performance to economic outcomes. Examples include expected profit per loan, value-at-risk (VaR), conditional value-at-risk (CVaR) for loan portfolios, and the cumulative gains curve.

Lyócsa et al. (2022) represent a partial exception by constructing an internal-rate-of-return-based evaluation, but their approach remains the minority. The under-use of economic metrics is a reflection of the field's computer-science orientation and suggests productive opportunities for closer integration with finance-theoretic frameworks.

On the dataset side, the Lending Club dataset appears in at least nine of the reviewed studies, making it the single most influential benchmark. Chinese P2P platforms (Renrendai, PPDai, and generic "Chinese P2P" data) provide the second main source. Other datasets include KakaoBank internet-banking log data (Kyeong et al., 2022), European P2P platforms (Lyócsa et al., 2022), and Kaggle-hosted proprietary loan datasets. The concentration on Lending Club data raises concerns analogous to those in computer vision with ImageNet: repeated benchmarking on the same dataset can lead to subtle overfitting at the community level, even when individual studies respect proper train/test splits.

4.6. RQ6—Research Gaps and Future Directions

Based on the bibliometric and methodological analysis, the following research directions emerge as the most consequential for future work on deep learning-based credit risk prediction in fintech lending:

- (1) Principled handling of class imbalance. Although several imbalance-handling techniques are represented in the corpus, the field has yet to produce systematic comparative studies that isolate the contribution of the imbalance-handling strategy from the underlying architecture. Future research should conduct rigorous ablation studies comparing SMOTE, ADASYN, focal loss, class-balanced loss, LDAM, and generative adversarial imbalance-handling on common benchmarks. Additionally, the integration of imbalance handling directly into the loss function (rather than the data-sampling step) offers theoretical advantages that remain underexplored. A particularly promising direction is the extension of the BSPER framework of Y. Wang et al. (2024) to non-reinforcement-learning settings;
- (2) Uncertainty quantification and calibrated prediction. The absence of Bayesian deep learning, Monte Carlo dropout, deep ensembles with explicit uncertainty output, and conformal prediction in the corpus represents a major gap. Credit risk is fundamentally a probabilistic problem, and the provision of calibrated uncertainty estimates is essential for portfolio-level risk management, stress testing, and regulatory capital calculation. Future research should adapt uncertainty-quantification techniques from computer vision and natural language processing to the fintech credit risk setting, with particular attention to calibration under imbalance;
- (3) Concept drift and temporal robustness. Fintech lending operates in a non-stationary environment, yet offline-trained models dominate the corpus. Future research should develop drift-aware deep learning models that can be updated incrementally as new data arrives, detect distribution shifts, and quantify the resulting impact on risk estimates. The online learning framework of Zhang et al. (2020) offers a starting point, but more principled integrations of concept-drift detection (ADWIN, DDM) with modern deep architectures are needed;
- (4) Graph neural networks and relational modeling. The multi-view GCN of Z. Li et al. (2024) demonstrates the value of explicitly modeling relational structure among loan applications. Future research can extend this direction by incorporating heterogeneous graph neural networks, temporal graph networks, and attention-based graph models. These approaches are particularly well-suited to fintech data, where borrowers are often connected through shared devices, IP addresses, employers, or social networks;
- (5) Alternative data and behavioral modeling. The pioneering work of C. Wang et al. (2019) on Event2vec and Kyeong et al. (2022) on system log data has yet to generate a

large follow-up literature. Future research should systematically explore the value of alternative data—including digital footprint, mobile application usage, social network features, and psychometric signals—for credit risk prediction, while rigorously assessing privacy and fairness implications;

- (6) Explainable AI (XAI) and regulatory compliance. The corpus contains several studies with XAI as a core methodological focus (notably [Ariza-Garzón et al., 2020](#); [Straus et al., 2025](#)), the majority of reviewed studies do not integrate explainability into their model design despite increasing regulatory pressure for interpretability in financial services. Future research should integrate SHAP, LIME, attention visualization, and counterfactual explanation methods into deep learning-based credit risk pipelines. The hybrid attention–LightGBM architecture of [Ying et al. \(2025\)](#) offers a template by combining a partially interpretable boosted-tree backbone with a deep attention layer;
- (7) Benchmark diversity and external validity. The dominance of Lending Club data risks community-level overfitting. Future research should develop and publish new benchmark datasets—ideally from diverse geographic and institutional contexts—and establish standard evaluation protocols that include economic metrics alongside statistical ones;
- (8) Integration of economic theory and deep learning. The dominant computer-science orientation of the field has produced methodological sophistication but limited integration with financial theory. Profit scoring ([Lyócsa et al., 2022](#)), expected utility-based evaluation, and portfolio-level risk optimization represent promising directions for bridging this gap.
- (9) Large language models, tabular foundation models, and the next architectural wave. The Transformer-based credit scoring of [Schwab and Kriebel \(2025\)](#) is a harbinger of broader developments. Tabular foundation models (TabPFN, TabTransformer, SAINT) offer the prospect of pre-trained representations that transfer across credit datasets, potentially mitigating imbalance and small-sample problems. None of the 30 reviewed studies employs a tabular foundation model or fine-tuned LLM for credit scoring, marking this as a wholly open frontier. Future research should benchmark these against incumbent architectures in Table 5, with attention to calibration, sample efficiency on imbalanced datasets, production-scale computational costs, and regulatory transparency compliance.

Taken together, these nine research directions constitute a coherent agenda for advancing the accuracy, robustness, interpretability, and economic relevance of deep learning-based credit risk prediction in fintech lending. The next section draws the implications of this analysis for scholars, practitioners, and policymakers.

5. Conclusions

This article presented a systematic literature review and bibliometric synthesis of deep learning applications for credit risk prediction in fintech lending, based on 30 open-access publications retrieved from three complementary databases (Scopus, IEEE Xplore, and Web of Science) and analyzed through PRISMA 2020-compliant procedures. The multi-database search strategy proved substantively consequential: 7 of the 30 included papers (23.3%) were retrieved exclusively from IEEE Xplore or Web of Science, and these 7 papers collectively account for 267 citations (32.8% of the total), including the corpus’s highest-cited paper ([Ariza-Garzón et al., 2020](#), 138 citations) and the corpus’s only Transformer architecture ([Schwab & Kriebel, 2025](#)). The methodological basis combined quantitative bibliometric techniques—including a four-cluster keyword co-occurrence analysis, bibliographic coupling network, thematic map, and topic dendrogram—with a structured qualitative reading of each included publication.

From a theoretical perspective, the main conclusion is that deep learning has become an integral component of the credit risk prediction toolkit in fintech lending, yet the field is characterized by substantial methodological heterogeneity and uneven maturity across sub-areas. The corpus shows a clear trend from single-architecture applications (2018–2020) toward hybrid ensemble architectures and XAI-integrated approaches (2022–2026). Credit risk prediction in fintech must be conceptualized as a structured prediction problem involving imbalanced classes, non-stationary data, economic decision-relevance, interpretability requirements, and regulatory compliance. The corpus evolution from three to four thematic clusters—with the emergence of a dedicated XAI/explainability cluster—provides direct empirical evidence that theoretical frameworks connecting deep learning methodology to regulatory compliance are becoming a recognized sub-field. Future theoretical work should focus on formal analysis of the bias-variance tradeoff under severe class imbalance, theoretical guarantees for deep ensemble calibration, and economic-theoretic foundations for loss-function design.

From a practical perspective, the findings carry differentiated implications for three stakeholder groups. For fintech platform operators and credit risk teams, the most actionable finding is that hybrid deep-boosting architectures (HBA-LGBM, GBDT-Deep FFM, hybrid CNN-GBDT) consistently outperform standalone deep networks, with AUC advantages of 3–6 percentage points, and that imbalance handling is a mandatory preprocessing step: default-class F1 scores improve by 15–47 percentage points when explicit strategies (SMOTE, ADASYN, BSMOTE, attention re-weighting) are applied. For financial data scientists and machine learning engineers, a minimum viable evaluation protocol should include: (i) separate reporting of majority-class and minority-class F1; (ii) ablation of the imbalance-handling component from the architecture; (iii) calibration assessment via Brier score; and (iv) out-of-time validation simulating production deployment under concept drift. For technology vendors and platform architects, the emergence of XAI-integrated scoring models, Transformer architectures, and GCN-based approaches suggests that credit-risk infrastructure should evolve toward generating SHAP-valued decision explanations at inference time.

From a policymaking perspective, three concrete regulatory implications emerge. First, the [European Parliament and Council \(2024\)](#) classifies credit-scoring AI as high-risk under Annex III, requiring transparency documentation, fundamental rights impact assessments, human oversight, and conformity assessment—obligations that most black-box deep learning models cannot currently satisfy without embedded explainability layers. Second, Basel III/IV model risk governance requires auditable credit models with documented validation procedures; the absence of adversarial robustness testing in most reviewed studies represents a model risk management gap that practitioners must address before regulated deployment. Third, the geographic profile of the corpus (36.7% from China) implies that policymakers in emerging markets face a substantial gap between local fintech credit risk environments and the research literature; investment in domestic open benchmark datasets and research capacity is essential for evidence-based credit risk regulation.

The scientific novelty of this study lies in three complementary contributions: (i) a multi-database systematic review that reveals a substantially larger and more geographically diverse research corpus than prior single-database analyses; (ii) explicit focus on mathematical and methodological challenges—class imbalance, loss function design, uncertainty quantification, concept drift, and adversarial robustness—that distinguishes this review from descriptive bibliometric surveys; and (iii) the identification of an XAI/explainability sub-field that is empirically well-developed but methodologically disconnected from regulatory compliance frameworks.

Despite its comprehensive nature, the review has several limitations. First, the corpus remains restricted to open-access publications; future research should relax this constraint through institutional library access. Second, the temporal scope (2015–2026) potentially omits earlier neural-network-based credit risk work. Third, despite the expansion to 30 publications, the corpus size remains modest relative to SLRs in adjacent domains (Hyk et al., 2026; $n = 193$), limiting the statistical robustness of some bibliometric computations. Fourth, the emphasis on mathematical methodology may under-weight organizational, behavioral, and regulatory dimensions. Finally, the rapid pace of methodological innovation means any review captures a snapshot of a moving target; as elaborated in Section 4.6 (direction 9), the advent of tabular foundation models and fine-tuned LLMs will likely reshape the architectural landscape cataloged in Table 5 within the next two to three years.

Notwithstanding these limitations, the present review establishes the current state, structure, and trajectory of deep learning-based credit risk prediction in fintech lending. By combining bibliometric rigor with mathematical and methodological depth, it offers a concrete roadmap for the next generation of research—one in which a deeper integration of financial theory, statistical learning, computational innovation, and regulatory compliance can produce credit risk models that are not only more accurate but also more robust, more interpretable, and more economically meaningful.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jrfm19070465/s1>, Table S1: PRISMA 2020 Checklist.

Author Contributions: Conceptualization, M.P.A.S., S. and R.; methodology, M.P.A.S., M.M. and A.J.W.; software, R., M.M. and A.J.W.; validation, M.P.A.S., S., A.K. and R.; formal analysis, M.P.A.S., S. and R.; investigation, M.M., A.K. and A.J.W.; resources, M.P.A.S., A.K. and A.J.W.; data curation, M.P.A.S., M.M. and A.J.W.; writing—original draft preparation, M.P.A.S., S. and R.; writing—review and editing, M.M. and A.J.W.; visualization, M.P.A.S., A.K. and A.J.W.; supervision, M.P.A.S.; project administration, M.P.A.S.; funding acquisition, M.P.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Equity-WCU Review Article Grant (Number: 3947/UN6.3.1/PT.00/2025).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors are grateful to Universitas Padjadjaran (Unpad) for providing Article Processing Charge (APC) support. The APC for this article was funded by Unpad through the Indonesian Endowment Fund for Education (LPDP) on behalf of the Indonesian Ministry of Higher Education, Science and Technology and managed under the EQUITY Program (Contract No. 4303/B3/DT.03.08/2025 and 3927/UN6.RKT/HK.07.00/2025).

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

References

- Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. [CrossRef]
- Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M.-J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *IEEE Access*, 8, 64873–64890. [CrossRef]
- Berhane, T., Melese, T., & Seid, A. M. (2024). Performance evaluation of hybrid machine learning algorithms for online lending credit risk prediction. *Applied Artificial Intelligence*, 38(1), 2358661. [CrossRef]

- Chang, Y., Tsai, Y., & Liu, Y. (2022). Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using lending club data. *Quantitative Finance and Economics*, 6(2), 303–315. [\[CrossRef\]](#)
- Chen, W., Xiang, G., Liu, Y., & Wang, K. (2021). Predicting default risk on peer-to-peer lending imbalanced datasets. *IEEE Access*, 9, 74571–74586. [\[CrossRef\]](#)
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field. *Journal of Informetrics*, 5(1), 146–166. [\[CrossRef\]](#)
- Deshmukh, R., Tan, S.-H., Tan, Y.-F., & Shrivastava, A. (2025). Artificial intelligence for smarter financial decisions: A comprehensive analysis of risk assessment and predictive tools. *Journal of Machine and Computing*, 5(2), 1124–1133.
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. [\[CrossRef\]](#)
- Dzik-Walczak, A., & Heba, M. (2021). An implementation of ensemble methods, logistic regression, and neural network for default prediction in peer-to-peer lending. *Zbornik Radova Ekonomskog Fakulteta u Rijeci*, 39(1), 163–197. [\[CrossRef\]](#)
- Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, 105(3), 1809–1831. [\[CrossRef\]](#) [\[PubMed\]](#)
- European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 1689.
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 507–520.
- Hoanh-Su, L., Phong, L. Q. C., Vinh, T. C., Nhat, H. M. M., & Lee, J.-H. (2025). Default prediction in the finance industry based on ensemble learning: Combining machine learning and deep learning. *Business Systems Research*, 16(1), 66–89. [\[CrossRef\]](#)
- Hyk, V., Vysochan, O., & Vysochan, O. (2026). The digitalization of corporate sustainability reporting: A systematic literature review and synthesis for future research. *Journal of Risk and Financial Management*, 19(3), 167. [\[CrossRef\]](#)
- Karn, A. L., Sachin, V., Sengan, S., Indra, G. V., Ravi, L., Sharma, D. K., & Subramaniaswamy, V. (2022). Designing a deep learning-based financial decision support system for fintech to support corporate customer's credit extension. *Malaysian Journal of Computer Science*, 2022(1), 116–131.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25. [\[CrossRef\]](#)
- Kim, J.-Y., & Cho, S.-B. (2019). Towards repayment prediction in peer-to-peer social lending using deep learning. *Mathematics*, 7(11), 1041.
- Ko, P.-C., Lin, P.-C., Do, H.-T., & Huang, Y.-F. (2022). P2P lending default prediction based on AI and statistical models. *Entropy*, 24(6), 801. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kumar, M. R., & Gunjan, V. K. (2020). Review of machine learning models for credit scoring analysis. *Ingenieria Solidaria*, 16(1), 1–17. [\[CrossRef\]](#)
- Kyeong, S., Kim, D., & Shin, J. (2022). Can system log data enhance the performance of credit scoring? Evidence from an internet bank in Korea. *Sustainability*, 14(1), 130.
- Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *IEEE Access*, 6, 54396–54406.
- Li, Z., Chen, Y., Wang, X., Yao, L., & Xu, G. (2024). Multi-view GCN for loan default risk prediction. *Neural Computing and Applications*, 36(18), 10195–10213. [\[CrossRef\]](#)
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE ICCV* (pp. 2980–2988). IEEE.
- Lombardi, R., & Secundo, G. (2021). The digital transformation of corporate reporting. *Meditari Accountancy Research*, 29(5), 1179–1208.
- Lyócsa, Š., Vašaničová, P., Hadji Misheva, B., & Vateha, M. D. (2022). Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. *Financial Innovation*, 8(1), 32.
- Ma, Z., Hou, W., & Zhang, D. (2021). A credit risk assessment model of borrowers in P2P lending based on BP neural network. *PLoS ONE*, 16(8), e0255216. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mienye, I. D., Jere, N., & Obaido, G. (2026). Deep learning for credit risk prediction: A survey of methods, applications, and challenges. *Information*, 17(1), 395. [\[CrossRef\]](#)
- Owusu, E., Quainoo, R., Mensah, S., & Appati, J. K. (2023). A deep learning approach for loan default prediction using imbalanced dataset. *International Journal of Intelligent Information Technologies*, 19(1), 1–16. [\[CrossRef\]](#)
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
- Qian, M., & Hu, F. (2019). An empirical study on prediction of the default risk on P2P lending platform. *IOP Conference Series: Materials Science and Engineering*, 490, 062041. [\[CrossRef\]](#)

- Schwab, P., & Kriebel, J. (2025). Mitigating adversarial attacks on transformer models in credit scoring. *European Journal of Operational Research*, 328, 309–323. [[CrossRef](#)]
- Straus, L., van den Berg, J., & Hogenboom, F. (2025). Explaining to defend: SHAP-based defense mechanism against adversarial attacks in P2P credit risk. *Expert Systems with Applications*, 301, 130432.
- Trinh, L. T. (2024). A comparative analysis of consumer credit risk models in peer-to-peer lending. *Journal of Economics, Finance and Administrative Science*, 29(58), 315–332.
- Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, 7(6), 191649. [[PubMed](#)]
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. [[PubMed](#)]
- Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. *IEEE Access*, 7, 2161–2168.
- Wang, M., & Cui, Y. (2024). Design and implementation of risk control model based on deep ensemble learning algorithm. In *IFIP advances in information and communication technology*. Springer.
- Wang, Y., Jia, Y., Fan, S., & Xiao, J. (2024). Deep reinforcement learning based on balanced stratified prioritized experience replay for customer credit scoring in peer-to-peer lending. *Artificial Intelligence Review*, 57(4), 95.
- Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: A machine learning methodology. *Scientific Reports*, 11(1), 18759. [[CrossRef](#)] [[PubMed](#)]
- Ying, C., Shi, A., & Li, X. (2025). Hybrid boosted attention-based LightGBM framework for enhanced credit risk assessment in digital finance. *Humanities and Social Sciences Communications*, 12, 1280.
- Yu, W. (2025, September 12–14). *Personal credit assessment model based on SOM neural network and K-means clustering algorithm*. Proceedings of 2025 8th CISAI (pp. 1382–1387), Wuhan, China.
- Zhang, Z., Niu, K., & Liu, Y. (2020). A deep learning based online credit scoring model for P2P lending. *IEEE Access*, 8, 177307–177317. [[CrossRef](#)]

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