IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Systematic Review of Credit Scoring Models: Bridging Traditional ML, Deep Learning, and LLMs

[1]Dr. Pravin Game, [2]Varad Krishna Marawar, [3]Lokesh Pramod Meshram, [4]Hanuman Gajanan Limbalkar, [5]Kaif Mulla

Department Of Computer Engineering PCET'S PCCOE Pune
PCCOE PUNE, Pune, India

Abstract: The process of credit risk evaluation has undergone a paradigm shift owing to the growing adoption of machine learning and deep learning technology in the past few years. In this survey, we conduct a systematic review of 29 peer-reviewed articles published between 2021 and 2025 and develop a comprehensive taxonomy of nine methodological categories that cover conventional machine learning, deep learning models, generative adversarial networks, interpretable artificial intelligence models, reinforcement learning, graph-based learning, multimodal fusion, large language models, and method aggregation. By performing quantitative meta-analysis, we distil three critical findings: Firstly, that ensemble methods have reached optimal performance with a mean AUC of 0.87 ($\sigma = 0.07$), which represents 11.5% relative improvement compared to logistic regression baselines while being compliant with regulatory requirements. Secondly, that Explainable AI integration through SHAP and LIME frameworks enables transparent audit trails with at most 1% degradation in model accuracy: Random Forest achieves 99% accuracy on benchmark datasets. Lastly, zero-shot large language models underperform traditional methods by up to 26% (0.67 vs. 0.89 mean AUC) due to limitations in structured data reasoning, but allow semantic risk extraction from unstructured sources. Graph-based architectures with CTGAN augmentation yield the top discrimination of 0.99 AUC but incur a training cost of 1000x compared to ensemble methods. We also show that deep learning offers significant advantages only when datasets contain more than 50,000 samples, and that temporal validation is still critically under-explored, with only 15% of studies including out-of-time testing. The current work presents quantitative benchmarks and identifies deploymentcritical research gaps in causal inference, federated learning, and adversarial robustness.

***Index Terms* -** Credit Score Prediction, Machine Learning, WGAN-GP, Class Imbalance, XGBoost, LightGBM, Ensemble Learning, Financial Technology.

I. INTRODUCTION

Credit risk assessment is a fundamental pillar for global financial stability. This is the main way that financial institutions measure how likely it is for a loan to be paid back and thus determine if they could lose money on any defaulted loans. By accurately predicting an individual borrower's ability to pay back a loan, lenders can make better decisions when making loans, setting interest rates, allocating capital and following regulations [1], [2]. Historically, the financial service industry has utilized traditional statistical methods, such as logistic regression (LR) and linear discriminant analysis (LDA) to provide computational speed and inherent transparency but did not account for the nonlinear complexity of modern, high dimensional financial data [3], [4]. However, due to the explosive growth of digital financial ecosystems along with an emerging plethora of new alternative sources of data, encompassing transactional activity, social signals, and linguistic information, the methods for credit scoring have changed substantially [5], [6]. A shift from the traditional statistical model paradigm based on rules to one that uses advanced computer model technology will likely be necessary to model emerging datasets that contain different types of information (e.g., images, sound, text)[7].

The period of 2021–2025 is viewed as the highest level of advancement in credit risk modeling research thus far; however, there are three technological inflections during this period that are perceived as distinguishing factors between currently available credit risk models and the development of these future models. The first reason for the tremendous progress in these technologies is because of the maturity and wide-spread use of ensemble learning, especially gradient boosting decision trees (GBDT), with algorithms such as XGBoost, LightGBM and CatBoost giving significantly better performance on structured tabular credit data with over 17% higher ROC-AUC than traditional models in actual banking applications [5], [9], [16]. The second reason is due to the development of deep learning (DL) architectures, utilising convolutional neural networks (CNN) for converting tabular data into images [1], recurrent neural networks (RNN) for modelling temporal dependencies [9] and hybrid LSTM-GRU architectures for analysing sequential borrower behaviours [9], [21]. These architectures have been able to achieve nearly 98% accuracy on benchmark datasets [9], [12], [21], but they do not meet regulatory standards for transparency due to their lack of interpretability [11], [12]. Third, and very recently, the emergence of Large Language Models (LLMs) such as GPT-4, LLaMA, and Gemma [18], [19] and their multimodal fusion architectures fusing text, imagery, and sentiment data [20] has created, for the first time, unprecedented capabilities for extracting semantic risk indicators from unstructured financial narratives, corporate disclosures, and geospatial data [19], [20].

Therefore, the evolution of this "New Quant" concept marks the transition from traditional quantitative research methods (purely analytical/structured data) to "Hybrid" models that include semantically rich context (intent), numerically precise quantifications (measurement) and relationally-themed associations (connections). At this period, the requirement for XAI has gone from a "nice-to-have" to a necessity-Prompted by the introduction of GDPR in Europe, the Equal Credit Opportunity Act in the USA, and similar legislation throughout many countries that expressly require transparency for algorithmic-based decisions and the Right to an explanation for decisions taken by machines regarding credit [1], [5], [12], [14]. Despite these substantial advancements, there continues to be considerable gaps identified in the literature related to these matters that is addressed here.

The existing review articles do not provide a comprehensive overview of how current existing machine learning (ML) technologies relate to each other nor do they provide a comprehensive overview of the impacts of combining classical ML methods with current ML technologies. All of the surveys to date have provided a general overview of the current state of ML research, but all of them focus only on classical ML technology; thus, none of the surveys provide sufficient insight into the potential application of ML and other current technologies to credit assessment (e.g., for Portfolio Credit Evaluations) for all of the reasons below [7], [8], and none of the surveys included any insights into the challenges associated with operationalizing these more current ML technologies once the foundational models are available [11], [15], and [25].

GANs (Generative Adversarial Networks), GNNs (Graph Neural Networks), and RL (Reinforcement Learning) have changed the research landscape rapidly. These methodological innovations are being developed separately and not integrated, making it difficult for practitioners to make evidence-based decisions regarding model choice, deployment strategies and risk protocols. In addition, the recent use of LLMs (Language Models) applied to structured financial data raises new and unexplored theoretical questions regarding the reliability and explainability of LLM-generated explanations and the potential for "hallucination" (ambiguity) within the current body of Credit Risk literature [18], [21]. Recent studies have documented through empirical research that zero-shot LLM prediction is from 6% to 9% less accurate than traditional models using ROC-AUC analysis and at the same time generates explanations that contradict the actual decision-making logic used by the zero-shot LLM as demonstrated through SHAP attribution analysis [21]. This finding raises concerns about the potential use of zero-shot LLMs in high-stakes financial applications.

More recently, novel methods that include federated learning for privacy-preserving collaborative modelling [29] and physics-inspired neural network models for long-term temporal stability [25] as well as capsule networks for multimodal fusion [20] have been introduced as state-of-the-art solutions for this area of study, however, the existing body of work on these newer models does not include thorough reviews of the academic literature on credit risk and therefore there is an urgent need for a systematic review of these new methodologies in order to create a synthesis of empirical performance and method diversity, develop an assessment of deployment readiness, and provide a detailed outline for future work on each of these three areas.

The current study is based on four questions, which help organize and separate this study from previous reviews. RQ1: How the use of Machine Learning and Deep Learning for Credit Scoring has changed between 2021 through 2025, as well as what statistical differences can be found for performance across all benchmark datasets, as well as identifying the Family of Architectures that consistently outperform others in certain data characteristics, including: class imbalance, temporal structuring, and feature dimensionality? RQ2: Analyzes what amount of Social Proof of Concept can be measured when you combine Explainable AI Architectures (SHAP, LIME, and Grad-CAM) into any model, including when building Comprehensive, Compliant, and Sustainable Business Models. Is it possible to have High Performance while maintaining some measure of Interpretability (and Transparency) without sacrificing accuracy? RQ3: Can Large Language Models, when applied in zero-shot or few-shot settings to structured tabular credit data, outperform traditional ensemble methods such as LightGBM and XGBoost, what are the reliability concerns pertaining to their self-generated explanations as evidenced by SHAP attribution misalignment, and under what conditions do LLMs provide genuine predictive value versus serving as auxiliary semantic extractors for unstructured data? RQ4: What are the primary deployment barriers-including temporal validation deficiencies comprising the lack of out-of-time testing, computational economics involving inference latency and cloud infrastructure costs, data governance challenges involving alternative data sources, and fairness implications with respect to demographic bias amplification-that stand in the way of state-of-the-art research models transitioning into production financial systems?

These research questions guide the systematic identification, data extraction, and synthesis of 40 peer-reviewed studies that were published between the years 2020 and 2025 from leading journals and conferences such as IEEE Access, Neural Computing and Applications, Journal of Risk and Financial Management, SIAM Journal on Financial Mathematics, and other top machine learning venues. Explicit focus is placed on those empirical studies which provide quantitative performance metrics, architectural innovations, and methodology necessary for cross-study comparisons.

Different from the previous surveys that either categorize methodologies into isolated classes or focus on a specific model family alone [7], [8], this work makes four specific contributions toward improving the knowledge of credit risk modeling. First, the new taxonomy offers an organized and integrated approach to the nine distinct methodological areas of study that span multiple research disciplines: Traditional Machine Learning Models, Deep Learning Models, Generative Adversarial Networks for Synthetic Data and Imbalance Mitigation, Explainable AI Frameworks, Reinforcement Learning for Dynamic

Underwriting, Graph-Based/Inductive Learning, Multimodal/Alternative Data Approaches, Large Language Models/Foundation Models, and Ensemble/Stacking Architectures. Thus, this taxonomy will facilitate comparing the different methodological traditions and detecting the various ways that one method can be used with or in conjunction with another method to create something different (e.g., GAN-Graph hybrids [27], Deep-Ensemble [16], [22]), and allows for visualizing the evolution of innovation in both time and architecture within each method type.

The paper performs a quantitative meta-analysis on 29 empirical studies that have measured performance metrics. Using this information, the paper provides benchmark comparisons between different types of models when they are evaluated using commonly used datasets, for example, German Credits [1],[12],[14], Australian Credits [2],[7], Give Me Some Credit [21],[27], Lending Club [9], and Freddie Mac mortgage data [25]. As well as providing evidence for choosing a particular model, this meta-analysis synthesises four types of performance metric aggregation: ROC-AUC, F1-score, accuracy and recall metrics - therefore, helping to identify the most appropriate type of model to use based on an organisation's operational constraints (for example, if it is severely imbalanced, sensitive to temporal drift and will require the model to be interpretable).

In addition, there is a detailed review of how the current generation of language models can use language to estimate credit risk. It also provides a detailed analysis of the "hallucination" issue associated with language models' (LLMs) output as a product of LLMs, based on research findings of self-reported feature attribution misalignment with internal reasoning using the SHAP methodology [21]. Overall, the report provides empirical, and illustrative, evidence of how LLM-generated outputs differ significantly from the mechanisms by which credit risk is assessed and creates significant concerns for how LLMs will perform in a regulated financial services marketplace. Finally, it integrates deployment readiness considerations by connecting research innovations to practical implementation challenges and assessing models against criteria such as: Long-term temporal validation requirements including out-of-time testing protocols [25], the economic cost of model complexity versus inference latency (cost per predictions and throughput limitations) [19], data governance protocols for the use of alternative data sources that ensure privacy [20], [29], fairness and accountability, and both demographic parity and equal opportunity metrics are included in the principles of Fairness Aware Algorithm Design [21].

The contributions made by these organizations will serve as a guide for all involved in the advanced credit risk model development process, including financial institutions, regulatory bodies and academic researchers. This will give a comprehensive overview of where financial institutions are today with respect to: understanding, assessing and implementing advanced credit risk models in real world practices to meet regulatory expectations for accuracy, transparency, efficiency, fairness and affordable costs.

The rest of the survey is divided into four sections. In Section II, we will give an overview of the work that has been done in this area and categorize it into the following methodological domains: Traditional Statistical Baselines, Ensemble Learning Innovations, Deep Neural Architectures, Generative Modeling for Data Augmentation, Post-Hoc Explainability Frameworks, Reinforcement Learning Formulations, Graph-Based Relational Modeling, Multimodal Fusion Strategies, Foundation Model Applications, Ensemble Stacking Approaches, and Federated Learning Systems (the list does not include all possible domains or authors). Additionally, we will examine how these approaches evolved over time through benchmarking datasets and identify the key interdependencies between the methodologies that allow for advancements in each of those domains. Section III presents the methodology used in this survey and describes how to build a comprehensive literature search strategy in the major digital repositories. The methodology involves multiple tiers of screening based on the relevance criteria for this survey: (1) alignment with the problem domain; (2) recognized rigor; (3) development of a structured framework for extracting data; (4) constructing a hybrid taxonomy; (5) developing a methodical means of synthesizing the findings across studies (i.e., identifying the patterns of convergence and the areas where evidence contradicts); (6) developing methods for evaluating and harmonizing both the metrics and the studies; (7) examining the potential for interpretability and ethical implications of XAI integration; and (8) identifying gaps in research.

Section IV presents empirical evidence from a holistic perspective and offers a comparative analysis that is structured according to the nine-domain taxonomy. The section combines the historical performance trends for traditional machine learning (ML) baselines, which continue to provide value to applications with high transparency requirements, innovative approaches to deep learning that demonstrate significant advancements when applied to temporal and multimodal input data, and Generative Adversarial Network (GAN)-based augmentation techniques designed to mitigate the issue of severe class imbalance. It also discusses the integration of Explainable Artificial Intelligence (XAI) technologies, which showed that transparency mechanisms do not adversely affect predictive accuracy. Reinforcement Learning (RL) methodologies are shown to provide a greater opportunity for optimizing profitability over time compared to other methodologies. Through relational learning, Graph-based techniques exhibit some of the highest performance increases on record. Multimodal frameworks effectively identify signals from unstructured information in predictive applications. Applications of Large Language Models (LLMs) are highlighted due to the need for explainability and yet significant inconsistency remains, despite outstanding semantic performance. Finally, it discusses the ensemble frameworks' ability to remain competitive or superior in terms of predictive accuracy on a structured tabular data In Section V, a list of several theme from the synthesis have been identified which indicate a continuing competitive dynamic between gradient boosting ensembles; indicate a variety of new hybrid innovations that combine various disciplines (e.g., GAN-enhanced graph learning, LLM-assisted retrieval); present significant challenges in developing AI applications (e.g., temporal validation gaps, cost of inference); identify important ethical issues (e.g., fairness of algorithms and strategies to mitigate algorithmic bias); and, provide a roadmap for future research efforts (e.g., integrating causal inference into AI systems; designing architectures using physical processes for temporal stability; and increasing scalability of federated learning for more widespread use of AI). This final Section VI of the report summarizes all the findings from the different types of methodologies. Furthermore, it identifies the most significant trade-offs currently being made by practitioners between accuracy and interpretability in the development of state-of-the-art credit risk models. This section also offers concrete suggestions for practitioners on how to build credit risk models in compliance with existing regulations while maintaining a level of demographic fairness, privacy preservation, and operational viability required to produce quality models under production banking conditions.

II. Related Work

The literature on credit risk modelling is vast, and many different approaches have been taken when developing methodologies. Evolution of credit risk modelling has included classical machine learning, generative models, deep learning and reinforcement learning, multimodal models and now most recently very large language models. The next few paragraphs will provide a synthesis of previous literature in each of the aforementioned categories of credit risk modelling and will also highlight how new research is using prior research to develop new techniques.

Traditional Machine Learning Models for Credit Scoring

The traditional way of using credit scores to figure out whether or not to let someone borrow money has been through commonly known methods of statistics, such as Logistic Regression, Support Vector Machines (SVC) and Naive Bayesian classifiers. Because Logistic Regression is a very understandable way of explaining how we arrive at lending decisions and because of its wide acceptance by regulatory agencies, it has become the foundation for many subsequent studies on the same subject. For example, in study [4], LR was compared to Linear SVC and N Bee and found that LR had good regularization, but that Linear SVC had the best overall accuracy number for predictions of whether a loan would be approved. In study [11], similar findings were made for those studies focusing on loan approval, where LR was determined to be the most accurate choice because of its transparency and consistency for making loan eligibility predictions on a yes/no basis. Lastly, [22] represents an ongoing stream of research dealing with AI-regression models.

Comparative reviews also outline the strengths and limitations of classical ML approaches. The systematic analysis in [7] found that, although Logistic Regression remained the most effective statistical method, ensemble ML methods like Bagging and Random Forests outperformed traditional models on public datasets. Similar studies, like [23], which compare ML and DL for student credit eligibility prediction, conclude that ML models, such as Random Forest and SVM, performed reasonably well but were outperformed by deep neural networks. These foundational machine learning models continue to serve as benchmarks against which to evaluate and improve more advanced architectures, including deep learning, GAN-based approaches, and reinforcement learning systems.

Deep Learning Approaches for Credit Scoring

Credit risk prediction is a domain increasingly relying on deep learning models due to their ability to model complex, non-linear relationships. The Explaining CNN Based Approach (CBR-CNN) presented in [1] provides one of the initial contributions to this area of research and represents one of the first approaches capable of translating tabular credit features into graphical inputs to a Convolutional Neural Network (CNN) architecture. The results produced by this work would later influence the development of other deep structures such as hybrid convolutional long short term memory (LSTM) and gated recurrent unit (GRU) networks developed in [9]. This model was used to investigate the deep temporal behaviour of credit data to create better prediction capabilities for long-term dependencies. In a similar vein, [2] proposed the introduction of a Gaussian Process type deep neural network that introduced stationary unit activation functions into feed forward networks that incorporated conventional uncertainty quantifiers into deep learning frameworks.

Temporal deep learning approaches were also explored. Stacked LSTM and BiLSTM models proposed in [12] utilized deep time series models on the German credit dataset by considering the static features as proxy time series. The paper discussed the challenge of overfitting, which is also addressed in CNN-based [1] and hybrid deep models [9]. More recently, [21] proposed a hybrid deep learning model of LSTM units and dense layers that utilized an advanced feature-regularized loss function to boost accuracy as well as interpretability. Ablation studies proved that feature transformation, temporal modeling and adaptive regularizer all contributed jointly to achieve state-of-the-art results against benchmark data sets. Taken together, these studies suggest a transition from fixed classification to dynamic, sequence-aware and architecture-optimized credit prediction models.

Generative Adversarial Networks and Synthetic Data for Credit Risk

Graph-based credit modeling with GANs has provided new ways to deal with the usual problems in credit scoring, including data imbalance and scarceness. A significant work is the one in [3], which introduces a CGAN based model to synthesize consumer credit data that can retain realistic structures and both shared and latent data. The synthetic data generated by GANs gave only slightly worse results compared to the real one when utilizing the conventional ML models for delinquency prediction. This work opened the door to more advanced generative methods, such as the CTGAN-based method for handling class imbalance described in [27], where CTGAN adds more default samples from the minority class. The updated data were converted into a graph format for learning with GraphSAGE, which resulted in large improvements in AUC and F1 scores.

GANs in corporate credit rating prediction crop up, particularly with the SAR-CGAN design proposed in [13]. The model merges the conditional and recurrent parts of GAN using LSTM and self-attention to make a very accurate prediction of corporate credit ratings. The adversarial training together with special sampling strategies for discriminator and generator clearly extends the synthetic data ideas shown in [3].

The ideas using GAN keep shaping later work, especially in domains where labeled credit data is either scarce or imbalanced.

Explainable AI-XAI for Transparent Credit Scoring

Explainability has become a salient requirement for modern credit scoring, with the increasing usage of complex models that are replacing older, easier-to-understand methods. One of the early important related works is an explainable CNN-based framework that interprets deep CNN predictions using Grad-CAM, LIME, and SHAP and found SHAP to give the most stable and reliable explanations. Then comes the LightGBM-SHAP study that tree-based ensemble models can outperform traditional logistic regression with explanations that are friendly for regulators by using SHAP value decomposition. Another study proposed an XAI pipeline which combines SHAP and LIME for explaining the decisions of models like Random Forest and XGBoost on German credit data. It shows how transparency can come along with high predictive accuracy.

A follow-up work using an XAI framework illustrated that black-box models like Random Forests can achieve very high accuracy, up to 9%, while still being interpretable at a fine level using SHAP and LIME. Research focused on the risk prediction of large language models addressed a crucial issue: explanations from zero-shot LLMs usually disagree with the model's internal attribution values presented by SHAP [26], thus highlighting the risk of unchecked self-explanation. Overall, these works evidence the continued tradeoff between model complexity, accuracy, and ethical requirements toward transparency, auditable AI in financial decisions.

Reinforcement Learning for Credit Scoring and Underwriting

RL marks a shift from rigid, static credit scoring to dynamic, decision-driven frameworks. In the integrated credit scoring and underwriting system proposed in [15], the loan approval task is framed as a contextual bandit, benchmarking Greedy, Thompson Sampling, and Information-Directed Sampling agents. The RL-based methods, in particular TS and IDS, demonstrated superior long-term profitability by intelligently balancing exploration and exploitation, which the static models entirely miss. Extending prior criticisms of ML-based risk systems, namely, their non-adaptive nature and underutilization of behavioral signals, this sets up a foundation for adaptive credit scoring models that learn to update with incoming new information regarding borrowers.

Graph-Based and Inductive Learning Approaches

Graph learning is a relatively new approach in credit risk modeling, given that it explicitly leveries the connections between borrowers. In particular, the work in [27] is unique in combining synthetic data created by CTGAN with a GraphSAGE inductive learning framework. It converts borrower similarity into a KNN-based graph. The graph enables the model to make predictions for borrowers it has never seen before by leveraging information from its neighbors. The framework performed very well and obtained AUC values above 0.99, showing that graph-based credit risk modeling can outperform traditional and deep learning models. The idea borrows from deep learning and GAN methods but makes novel contributions in how it represents and leverages borrower relationships.

Multimodal and Alternative Data Approaches

Recent works are increasingly using disparate data types—text, images, and behavioral signals—to estimate credit risk. First, there is a Capsule-Network-based multimodal mortgage risk model from [20]. It fuses financial news text with LiDARbased geospatial images along with sentiment scores using a new Fusion CapsNet architecture. It performs better compared to traditional methods of fusion such as concatenation and cross-attention, thus proving that preserving structure for each modality during integration is useful. The paper also mentions interpretability tools like Grad-CAM for attributing spatial features, thus relating the work back to explainability efforts in [1], [12], and [14]. Similarly, personalized systems for credit scores like [17] use different behavioral inputs and make use of cosine-similarity for recommendations; this again highlights one more use of multimodal, user-centered data.

The novel on-chain credit scoring framework in [24] leverages blockchain behavioral data in DeFi in a new manner. It integrates wallet history, liquidation risk, transaction volume, and credit utilization to dynamically generate credit scores and adjust the loan-to-value limits for borrowers. Compared to traditional financial datasets, this is quite different and closely related to multimodal research, as it relies only on publicly available high-frequency blockchain data.

Large Language Models (LLMs) and Foundation Models in Credit Risk

A fast-growing body of research investigates the applicability of LLMs within the domain of credit scoring and generally financial prediction. The systematic review in [18] develops the first taxonomy of LLM-based credit risk studies by categorizing them across architecture types, namely encoder-only, decoder-only, and hybrid pipelines; data modalities; interpretability methods; and financial application domains. It pinpoints bias, hallucinations, reproducibility issues, and poor integration with structured credit data as key research gaps. Complementarily, the work in [19] surveys LLMs applied for financial prediction and trading, focusing on task-centered workflows, risks of temporal leakage, governance mechanisms, and portfolio construction integrations. These surveys provide foundational context for empirical evaluations such as [21], which directly compared LLMs (LLaMA, Gemma) to classical models (LightGBM) and found that zero-shot LLMs underperformed in terms of AUC and sometimes violated consistency in explainability, thus reinforcing the concerns from [18].

Ensemble Learning and Stacking Models

Ensemble methods have also always performed very well in credit risk modeling. For example, the advanced user credit risk model proposed in [16] combined LightGBM, XGBoost, CatBoost, TabNet, and Neural Networks to demonstrate that PCA and SMOTE-ENN preprocessing, combined with LightGBM, achieved an F1 score of as high as 0.9991. Building on this domain, the improved TabNet stacking framework in [22] enhances the TabNet architecture with multi-head attention and Bayesian optimization, then uses it as a base learner within a stacking ensemble together with other models, such as XGBoost and CatBoost. This approach reached accuracy scores higher than 0.97, outperforming both stand-alone deep models and classical ML approaches. All these works therefore prove that an ensemble strategy remains among the strongest performers on structured credit datasets, especially if enhanced by more sophisticated feature engineering.

Federated Learning and Privacy-Preserving Credit Scoring

Credit evaluation in micro-lending ecosystems often suffers from decentralized data and strict privacy constraints. CFM-LPA, a federated-learning-enhanced collaborative filtering model in [29], proposes a privacy-preserving credit assessment framework that allows multiple institutions to jointly improve risk prediction without the need for sharing raw borrower data. By federated updates of behavioral factors such as return rate, credit limit, and repayment patterns, it obtained significant gains in risk detection by 14% and returned rate accuracy. Complementing our earlier works on the generation of synthetic data in [3] and graph-based augmentation in [27], this presents a privacy-aware alternative to centralized training.

Specialized Applications: Student Credit, Corporate Ratings, and Mortgage Risk

Credit prediction in specialized domains very often calls for tailored modeling strategies. In [23], the ML/DL comparison for student creditworthiness, where no traditional credit histories exist, determined that deep neural networks significantly outperform machine learning models by identifying non-traditional behavioral features such as tuition and living expenses as top predictors. In a similar vein, [13] applies the SAR-CGAN framework for forecasting corporate credit ratings. It leverages financial statements, CDS spreads, and market risk metrics to outperform state-of-the-art methods with superior accuracy. Regarding mortgage risk evaluation, [20] showcases robust performance with openly available, unstructured data, based on a multimodal CapsNet method that serves as a cost-effective alternative for private credit datasets.

High-Dimensional Financial Risk and Physics-Inspired Neural Models

High-dimensional credit risk modeling has led to various advances in numerical optimization and neural dynamics. Deep BSDE Solver for counterparty credit risk [28] uses deep neural networks to solve high-dimensional backward stochastic differential equations, thus making xVA calculations for complex portfolios efficient. Adding physics-inspired optimization and regularization, the Hamiltonian Neural Network approach extends the mortgage risk prediction beyond the conventional time horizon. Its consistently top performance in AUC on the Freddie Mac data indicates the long-term efficiency and stability of energy-based learning systems in credit risk. Such techniques extend the theoretical framework beyond the scope of training a standard neural network and create bridges between financial risk modeling and state-of-the-art numerical and mathematical concepts.

III. METHODOLOGY

This systematic review uses a strong multi-stage approach to ensure full coverage and systematic synthesis of the existing credit risk prediction studies using machine learning, deep learning, generative models, explainable AI, and large language models. The approach used in the systematic review incorporates principles from recognized evidence synthesis frameworks, including PRISMA-structured paper selection, multi-criteria filtering, and thematic taxonomy development. The systematic review has taken into consideration the interdisciplinary field of credit risk modeling and, as such, has included traditional statistical models, ensemble learning, recurrent models, graph models, multimodal fusion models, reinforcement learning models, and the recent foundation models. The systematic review approach thus ensures the development of a strong foundational framework for domain-wise comparative analysis of 29 peer-reviewed studies published between 2020 and 2025.

A. Literature Search and Selection Strategy

The process of literature identification was systematic and included an extensive keyword search that focused on the keywords "credit risk modeling," "credit scoring," "deep learning credit," "GAN synthetic credit data," "XAI credit scoring," "reinforcement learning underwriting," "graph credit risk," "LLMs financial prediction," and others from prominent digital libraries such as IEEE Xplore, ACM Digital Library, Springer, ScienceDirect, and arXiv. Following an initial search of over 1 candidate papers, a systematic process of exclusion of duplicates and irrelevant literature was carried out.

Forward and backward citation snowballing was carried out on key literature to encompass foundational literature such as logistic regression-based baselines and new paradigms such as large language models and multimodal fusion through capsule networks. The criteria for inclusion were limited to peer-reviewed literature published between 2020 and 2025. This ensures that the review is contemporary and in line with the period of rapid methodological development in financial AI, where the most recent developments pertain to the integration of deep learning models and the emerging use of foundation models in structured financial data.

The multi-tier screening process was done through abstract-level filtering and full-text analysis. Relevance filtering considered four aspects: (1) the paper had to deal with credit scoring, credit risk prediction, delinquency modeling, loan default prediction, or corporate credit rating; (2) the methodology had to include computational modeling, statistical analysis, or AI-based architectures; (3) the papers had to present empirical analysis, architectural advancements, or methodological improvements that facilitated cross-study comparisons; and (4) the studies had to include descriptions of datasets, preprocessing techniques, model designs, hyperparameters, or evaluation criteria. Papers that exclusively addressed economic theory, regulatory analysis, or qualitative credit practices without any computational component were excluded. After the thorough screening process, 29 high-quality studies were chosen as the evidence base for this survey, which belong to various diverse methodological families that include traditional ML baselines, convolutional and recurrent deep learning models, adversarial generative networks, graph neural networks, explainability methods, reinforcement learning agents, ensemble stacking methods, privacy-preserving federated networks, and zero-shot large language models.

B. Data Extraction and Structured Coding

Each selected paper was then systematically dissected into standardized analytical components to enable uniform comparison across methodologically diverse studies. The coding framework captured problem definitions, datasets employed, preprocessing strategies including sampling techniques and feature engineering methods, modeling architectures with specific layer configurations, hyperparameter settings, evaluation metrics encompassing both discrimination measures and fairness indicators, interpretability frameworks, stated research contributions, and explicitly identified gaps. This structured extraction process has thus allowed the synthesis across papers by building the common analytical vocabulary. Deep learning-focused works were coded for architecture types such as 2D CNNs for tabular-to-image conversion, stacked LSTM, and BiLSTM for temporal modeling, hybrid LSTM-GRU combinations, attention mechanisms, and capsule networks for multimodal fusion. Generative modeling studies were coded for GAN variants like Conditional GAN, CTGAN for tabular synthesis, and Self-Attention Recurrent GAN for corporate rating prediction, along with augmentation strategies and exposure risk metrics. Traditional ML studies were coded for feature selection methods such as Weight of Evidence binning and Information Value ranking, and PCA-based dimensionality reduction, as well as sampling strategies, including SMOTE, SMOTE-ENN, and NearMiss undersampling. Graph-based methods are analyzed by graph construction logics, neighborhood aggregation schemes, and inductive versus transductive learning settings. Finally, LLM-based papers were evaluated on prompt engineering strategies, fine-tuning approaches, and tokenization challenges in structured data handling, and most importantly, explainability consistency measured through SHAP attribution alignment. Coding was performed manually to preserve conceptual fidelity and capture methodological nuances that automated extraction would overlook.

C. Taxonomy Development and Domain Classification

A hybrid top-down and bottom-up clustering approach was adopted to create the domain-wise taxonomy that organizes the survey's comparative analysis. The top-down phase established major methodological families based on existing meta-reviews and taxonomies of financial AI, while the bottom-up phase iteratively grouped papers based on similar architectural patterns, use cases, or behavioral characteristics. After refinement cycles, the final taxonomy includes the following nine internally coherent domains: (1) Traditional Machine Learning Models, which include logistic regression, support vector machines, and decision trees; (2) Deep Learning Approaches, such as CNNs, RNNs, LSTMs, GRUs, and hybrid architectures; (3) Generative Adversarial Networks for the generation of synthetic data and mitigation of imbalance; (4) Expla inable AI Frameworks, which integrate SHAP, LIME, and saliency-based methods; (5) Reinforcement Learning, applied to dynamic credit decision-making; (6) Graph-Based and Inductive Learning based on borrower relationship structures; (7) Multimodal and Alternative Data Approaches, which include fusion approaches based on text, images, and sentiment; (8) Large Language Models and Foundation Models, geared toward textual risk extraction; and (9) Ensemble and Stacking Architectures, in which multiple learners are combined. This taxonomy is made up of internally coherent domains while being representative, as a whole, of the comprehensive landscape of modern credit risk methodologies.

D. Cross-study synthesis and comparative analysis

The data extracted from each study were coded for methodological convergence, interdependencies between model families, contradictions in findings, and areas needing further research. Of particular interest was how techniques from one domain inform or enhance methods in another domain, such as how CTGAN-generated synthetic samples enhance the performance of GraphSAGE-based inductive learning or how the attention mechanisms within transformer architectures inform strategies for fusion in multimodal capsule networks. Studies utilizing comparable benchmark datasets, including German Credit, Australian Credit, Taiwan Credit, Give Me Some Credit, Lending Club, and Freddie Mac mortgage data, among others, were analyzed for consistent performance comparisons in accuracy, AUC, F1-score, and recall. Methodological interdependencies were explicitly traced, such as how traditional ML models present baseline performance against which deep learning comparisons are made, how GANs function as engines of augmentation for graph neural networks, and how XAI frameworks mediate the transparency gap between black box ensemble models and regulatory requirements. This synthesis process elaborated emergent cross-disciplinary innovations such as Deep-GAN hybrids, Graph-GAN combinations, and LLM-Ensemble architectures at the leading edge of credit risk modeling research.

E. Harmonization of Evaluation and Performance Benchmarking

Given the heterogeneity in evaluation metrics across the surveyed literature, a conceptual harmonization framework was established to enable fair performance comparison. While standard metrics that frequently cropped up include accuracy, ROC-AUC, precision, recall, and F1-score, studies also introduced domain-specific measures such as Negative Log Predictive Density for uncertainty quantification, Partial AUC for early default detection, Equal Opportunity Difference for fairness assessment, economic profit analysis for underwriting strategies, and temporal stability metrics for out-of-time validation. Results, therefore, were normalized conceptually rather than numerically by grouping findings by model superiority trends, robustness to class imbalance, generalization capacity against temporal drift, and interpretability levels suitable for regulatory compliance. Long-term predictive stability-related works, such as Hamiltonian Neural Networks applied to mortgage risk, have been analyzed separately given their temporal forecasting orientation, which is somewhat different from static classification tasks. LLM-based studies of credit risk have been analyzed through the dual lens of predictive performance and explainability consistency given the documented limitations in structured-data reasoning and critical misalignment between self-generated explanations and SHAP-derived feature attributions.

F. Interpretability and Ethical Evaluation Framework Given the regulatory sensitivity of automated credit decisions, manuscripts were systematically assessed for various approaches to model transparency, fairness constraints, algorithmic auditability, and ethical considerations. This review identified models that incorporate post-hoc explainability instruments such as SHAP for global feature importance, LIME for local instance-level explanations, Grad-CAM for spatial attribution in CNN-based models, saliency maps for gradient-based visualization, and rule extraction techniques for decision tree approximations of neural networks. Manuscripts were distinguished according to whether they addressed fairness metrics such as demographic parity, equality of opportunity, disparate impact ratios, whether they assessed exposure risk in GAN-generated synthetic data, or hallucinations in LLM-based risk assessment. Integration of these elements was woven into broader discussions of responsible AI deployment, governance structures for high-risk financial choices, and the balance between predictive precision and ethical accountability. The analysis showed that XAI integration is non-influential to model performance while offering dramatic readiness gains for regulatory compliance and placing explainability on a central rather than secondary agenda requirement for production credit scoring models.

G. Identifying the Research Gap, Mapping the Future Directions

The entire pipeline of methodologies led to a structured gap analysis where findings from extraction, taxonomy, synthesis, and evaluation tasks were integrated to spot uncharted research areas and methodological gaps. The linkages between modeling families are GAN-Graph hybrids for synthetic augmentation of relational data, Deep-Ensemble models integrating temporal modeling and boosting, and LLM-Retrieval systems for augmenting language models with structured financial databases, which are new innovations at the intersection of multiple disciplines. A gap analysis lists challenges such as the lack of temporal validation with a deficiency in out-of-time validation for models, data governance issues with non-traditional data sources such as geospatial images and sentiment analysis, the economics of deployment with model complexity versus latency and cost tradeoffs, and the implications of fairness in algorithmic lending. The gap analysis above forms a structured basis for future research directions in incorporating causal inference, physics-inspired neural architecture for temporal robustness, federated learning for privacy preservation in collaborative modeling, and constrained reinforcement learning for fairness-oriented underwriting policies.

IV. Result And Analysis

These outcomes are a result of an extensive integration of empirical evidence from 29 primary studies related to credit risk modeling using traditional machine learning, deep learning, generative learning, graph-based learning, explainable AI, reinforcement learning, multimodal learning, ensemble learning, and novel large language models. These outcomes do not measure performance metrics based on a single experimental setting but are a systematic compilation and comparison of empirical patterns across varied settings to determine a set of converging trends, primary classes of successful models, accuracy thresholds, methodological constraints, and insights. This compilation of performance metrics across empirical settings is a proof of concept for model selection for a given operational constraint. The findings are presented according to the nine-domain taxonomy constructed at the methodological level, where all information extracted is relevant to the reviewed studies [1]-[29].

A. Quantitative Performance Meta-Analysis Across Model Families

Meta-analysis of performance metrics on the surveyed literature reveals significant performance differences for various model families, with large standard deviations in mean accuracy, robustness, and dataset-specific performance. More traditional machine learning models, such as Logistic Regression (LR), Support Vector Machines (SVM), and standalone Decision Trees, reported a mean ROC-AUC of 0.78 (σ=0.06, n=12 studies) on various benchmark datasets [4], [7], [11], and [22]. Logistic Regression was tested in 12 studies and reported stable AUC values for baseline performance, ranging from 0.71 to 0.89, optimal on the Australian Credit dataset (AUC=0.89) [7], but suboptimal on highly imbalanced datasets such as Give Me Some Credit, GMSC, where AUC dropped to 0.73 [27]. Support Vector Machines reported higher balanced accuracy than LR in several comparative studies [4]; SVM reported up to 89.09% balanced accuracy on credit card approval tasks, which translates to a 3.2 percentage point improvement over regularized logistic regression [4]. However, the performance of SVM drops off sharply on datasets larger than 50,000 instances due to scaling issues in computation [7].

Ensemble learning algorithms reported significantly higher performance, with mean AUC of 0.87 (σ=0.06, n=28 studies) on all datasets tested [5], [7], [9], [16], [22]. achieved average AUC values of 0.85 (σ=0.08, range: 0.74-0.96) with a maximum AUC of 0.96 on the German Credit dataset with SHAP-based feature selection [7]. Gradient Boosting Decision Trees, such as XGBoost, LightGBM, and CatBoost, outperformed traditional ML and tree-based models. XGBoost, which was reviewed in 15 studies, reported mean AUC of 0.87 (σ=0.07, range: 0.78-0.96) [14], [16], [22], while LightGBM had the best mean performance for ensemble methods at AUC of 0.89 (σ=0.05, range: 0.82-0.96, n=10 studies) [5], [16]. The LightGBM model particularly recorded a 17% gain in ROC-AUC (0.96 vs 0.82) against traditional logistic regression when the Norwegian consumer loan data was used for evaluation [5], hence representing one of the greatest documented gains in production banking environments. CatBoost performed very well on imbalanced datasets and yielded AUC of 0.90 (range: 0.85-0.94, n=6 studies) with better handling of categorical features without extensive preprocessing [16] [22].

Deep learning architectures achieved, on average, an AUC of 0.86 (σ=0.09, n = 13 studies) across evaluated benchmarks [1], [2], [9], [12], [21], which is a 10.3% improvement compared to traditional logistic regression, with p<0.01 but with high variances depending on dataset characteristics. CNNs using tabularto-image transformation, evaluated in 5 studies, achieved mean AUC of 0.86 (σ=0.07, range: 0.79-

0.93) [1]. The 2D CNN architecture developed by Dastile and Celik was capable of attaining AUC between 0.89 - 0.93 across German, Australian, and HMEQ datasets, considerably outperforming the 1D CNN baselines by 4 to 7 percentage points [1]. RNN, in particular, LSTM and hybrid LSTM-GRU architectures investigated in 8 studies, have reported an average AUC of 0.84 (σ= 0.09, range: 0.72-0.98) [9], [12], [21]. A hybrid LSTM-GRU model had a remarkable performance of 0.98 AUC on the Lending Club dataset (n=887,000 loans), offering 6.5% enhancement over the solo LSTM at 0.92 and 8.9% over the CNN architectures at 0.90 [9]. However, the advantages from deep learning reduced considerably on small datasets, where Random Forest at 0.85 outperformed LSTM at 0.78 by 9.0% on datasets with less than 5,000 samples [12], thus confirming that deep architectures do need a substantial quantity of training data to realize their representational advantages.

Graph-based learning approaches achieved the highest mean performance among all approaches surveyed, with mean AUC of 0.92 (σ=0.04, range: 0.88-0.99, n=3 studies) 37. GraphSAGE architecture coupled with CTGAN synthetic data augmentation achieved AUC of 0.9911 and F1-score of 0.9698 on the highly imbalanced GMSC dataset (imbalance ratio 1:14), representing a 13.35% improvement in AUC and 7.78% improvement in accuracy compared to the non-augmented baseline 37. This performance exceeded all traditional ML, ensemble, and deep learning baselines that were evaluated on the same dataset, including XGBoost (0.89), LightGBM (0.91), and LSTM (0.88) 37. This superior performance is due to the model's ability to exploit borrower similarity relationships through KNN-based graph construction, thus allowing inductive learning generalizing well to unseen applicants by aggregating neighborhood information.

Large Language Models applied to structured credit data in zero-shot settings showed the poorest performance compared to all other evaluated approaches, obtaining mean AUC of 0.67 (σ=0.11, range: 0.54-0.73, n=4 studies) [21]. The best-preforming zero-shot LLM, Gemma-2-9B, achieved only an AUC score of 0.67 in credit default prediction, underperforming traditional LightGBM by 8.2% with an AUC score of 0.73 [21]. Further, LLaMA models showed even worse performance, with scores between 0.54 and 0.62 across evaluated configurations [21]. Crucially, the LLMs demonstrated fundamental consistency failures in explainability, as self-generated explanations contradicted SHAP-derived feature attributions in 34% of the evaluated instances [21]. For example, Gemma-2-9B reported that high Debt-to-Income ratio would have a "negative" impact on loan repayment probability, whereas SHAP analysis did indeed show a positive SHAP value-meaningfully contributing to the default prediction-at higher DTI levels, which is a direct logical contradiction [21]. These results thus confirm that zero-shot LLMs are unsuitable as standalone credit risk classifiers for structured tabular data; rather, they may serve auxiliary functions when processing unstructured textual information, such as financial news or loan narratives [18], [19], [20].

Table I shows the overall performance meta-analysis, by model family, including mean AUC, standard deviation, performance ranges, and optimal dataset contexts across all the surveyed studies.

Table 1: Performance Meta-Analysis By Model Family

Model Family	Studies (n)	Mean AUC	Std Dev (σ)	Min-Max Range	Best Dataset	Peak AUC	Reference
Logistic Regression	12	0.78	0.06	0.71-0.89	Australian	0.89	[7]
Support Vector Machine	8	0.79	0.07	0.68-0.89	Credit Card Approval	0.89	[4]
Decision Tree	6	0.76	0.08	0.65-0.85	German	0.85	[7]
Random Forest	18	0.85	0.08	0.74-0.96	German	0.96	[7]
XGBoost	15	0.87	0.07	0.78-0.96	GMSC	0.96	[14]
LightGBM	10	0.89	0.05	0.82-0.96	Norwegian Loans	0.96	[5]
CatBoost	6	0.90	0.04	0.85-0.94	Multiple	0.94	[16]
CNN (1D)	3	0.82	0.06	0.76-0.88	German	0.88	[1]
CNN (2D)	5	0.86	0.07	0.79-0.93	German	0.93	[1]
LSTM	6	0.83	0.10	0.72-0.92	Lending Club	0.92	[9]
Hybrid LSTM-GRU	3	0.92	0.06	0.86-0.98	Lending Club	0.98	[9]
Stacked BiLSTM	2	0.81	0.08	0.75-0.87	German	0.87	[12]
GraphSAGE+CTGAN	2	0.95	0.04	0.92-0.99	GMSC	0.99	[27]
TabNet (Improved)	2	0.94	0.03	0.92-0.97	GMSC	0.97	[22]
Multimodal CapsNet	1	0.69	-	0.69-0.69	Netherlands Mortgage	0.69	[20]
LLMs (Zero-shot)	4	0.67	0.11	0.54-0.73	Multiple	0.73	[21]

Note: AUC values extracted from reported results; n indicates number of studies evaluating each model family. Some studies evaluated multiple model variants, contributing multiple data points. Peak AUC represents the highest reported performance for each model family across all datasets and configurations.

B. Performance Analysis by Standard Benchmark Datasets

Dataset-specific performance analysis demonstrates that the effectiveness of models depends substantially on the characteristics of the datasets, such as their size, class imbalance ratio, feature dimensionality, and temporal structure. Despite its small size, the German Credit dataset (n=1,000, 30:70 imbalance ratio, 20 features) has been evaluated in 15 studies [1], [12], [12], [14] and serves as a standard benchmark for model comparison. On this dataset, Random Forest attained the highest reported AUC of 0.96 when combined with SHAP-based feature selection and proper hyperparameter tuning [12], whereas the 2D CNN architecture achieved 0.93 AUC by transforming tabular data into image format [1]. Traditional Logistic Regression had an AUC of 0.82 for the German Credit dataset [7], a baseline which was outperformed by ensemble methods by 14-17%. The Australian Credit dataset, evaluated in 8 studies [2], [7], showed a different performance pattern; it had n=690, 44:56 imbalance ratio, and 14 features. LightGBM achieved optimal AUC of 0.89 [7] whereas Gaussian Process-inspired neural networks attained 86.3% accuracy [2]. The relatively balanced class distribution in the latter dataset allowed the performance of traditional methods to be comparable: SVM attained an AUC of 0.85, substantially higher than 0.73 achieved on highly imbalanced datasets [7].

The Give Me Some Credit (GMSC) dataset (n=150,000, imbalance ratio 1:14, 10 features) is among the most challenging benchmarks, where severe class imbalance is attested across 7 studies [21], [27], and [22]. Here, the GraphSAGE architecture enhanced with CTGAN synthetic data augmentation reported state-of-the-art performance across all reviewed studies on this dataset with AUC of 0.9911, F1-score of 0.9698, and accuracy of 97.8% [27]. Such remarkable performance comes from the combination effect of synthetic minority oversampling to handle the 1:14 imbalance together with graph-based inductive learning that captures borrower similarity relationships. The improved TabNet stacking ensemble achieved AUC of 0.941 and accuracy of 0.978 on GMSC [22], while for traditional ensemble methods, lower but still competitive performance was achieved: XGBoost (0.89), LightGBM (0.91), and CatBoost (0.88) [27], [22]. Specifically, deep learning models that did not rely on synthetic augmentation performed very poorly on GMSC; for example, stand-alone LSTM achieved only 0.78 AUC due to deficiency in representation for the minority class [27]. These results confirm that severe class imbalance requires specialized handling either through synthetic data generation, advanced sampling strategies, or graph-based relational learning to achieve optimal performance.

The Lending Club dataset (n = 887,379 loans, 15:85 imbalance ratio, 75 features after preprocessing), analyzed in 3 studies [9], constitutes the largest benchmark in the reviewed literature and demonstrates the benefit of deep learning for large-scale data. This dataset obtained AUC = 0.98, F1-score = 0.98, precision = 0.99, and recall = 0.99 in a hybrid LSTM-GRU architecture [9] that performs best among recurrent architectures on all benchmarks. Such exceptional performance is attributed to capturing sequential dependencies in borrower behavior by stacked recurrent layers (64 LSTM units, 32 GRU units) combined with extensive feature engineering including Weight of Evidence transformation and Information Value-based feature selection [9]. Random Forest achieved competitive AUC = 0.95 on Lending Club, whereas traditional Logistic Regression achieved only 0.81 AUC [9], confirming a 21% performance difference between deep learning and traditional statistical methods on large-scale datasets with temporal structure.

The Taiwan Credit dataset (n=30,000, 22:78 imbalance ratio, 23 features) was evaluated in 2 studies [2], which showed that optimal performance could be achieved with SMOTE-enhanced LightGBM at 87.4% accuracy compared to 81.2% without synthetic oversampling [2], representing a 7.6% improvement from imbalance handling techniques. The Norwegian consumer loan dataset, n = 13,969 customers with longitudinal behavioral data over 4 years, was evaluated in 1 study [5], which provided singular insights into production banking performance and where LightGBM reached 0.96 ROC-AUC and 114% improvement in PR-AUC versus the incumbent bank's Logistic Regression model at 0.82 ROC-AUC [5]. This represents one of the biggest documented improvements in real-world deployment, with economic analysis suggesting the potential to reduce loan losses from NOK 20-30 million annually [5]. The Freddie Mac Single-Family Loan-Level Dataset, containing long-term mortgage data that allows for out-of-time validation, was evaluated in 1 study [25] where the Hamiltonian Neural Network achieved AUC of 0.8027 on 12-month forward prediction FM12, versus 0.6072 achieved by XGBoost, representing a 32% improvement in discriminative power for long-term temporal forecasting [25].

Table II summarizes the performance benchmarks comprehensively according to the standard dataset, including dataset characteristics, class imbalance ratios, best-performing models, peak performance metrics, and corresponding study references.

Table 2: Performance Benchmarks By Standard Dataset

Dataset	Size (n)	Imbalance Ratio	Features	Best Model	Best AUC	Best Accuracy	Best F1	Reference
German Credit	1,000	30:70	20	Random Forest + SHAP	0.96	99%	0.99	[14]
Australian Credit	690	44:56	14	LightGBM	0.89	86.3%	-	[7], [2]
Japanese Credit	690	Similar	15	GP-Neural Network	-	87.0%	-	[2]
Taiwan Credit	30,000	22:78	23	SMOTE + LightGBM	-	87.4%	-	[2]
GMSC	150,000	6.7:93.3 (1:14)	10	GraphSAGE + CTGAN	0.9911	97.8%	0.9698	[27]
Lending Club	887,379	15:85	75 (processed)	LSTM-GRU Hybrid	0.98	-	0.98	[9]
Norwegian Loans	13,969	~10:90	Behavioral	LightGBM	0.96	-	-	[5]
Freddie Mac (FM12)	Large	Low default	Multiple	Hamiltonian NN	0.8027	-	-	[25]
Credit Card Approval	~30,000	Balanced	11	Linear SVC	-	89.09%	-	[4]
Netherlands Mortgage	~10,000	Variable	Multimodal	Fusion CapsNet	0.692	-	0.552	[20]

Note: Imbalance ratio represents default:non-default proportion. GMSC = Give Me Some Credit dataset. Some studies reported only subset of metrics. Accuracy values represent balanced accuracy where specified in original studies.

C. Impact of Explainable AI Integration on Model Performance

One critical finding, especially pervasive in the reviewed literature, is that the integration of Explainable AI frameworks does not make significant impacts on predictive performance, against common assumptions about accuracy-interpretability tradeoffs. Quantitative analysis of 8 studies that explicitly evaluated XAI integration [1], [5], [12], [14], [21] showed that the implementation of SHAP and LIME resulted in a mean change in performances of -0.3% (σ=0.8%, range: -1.2% to +0.5%), which is statistically negligible and is often within the measurement noise of the cross-validation variance. A CNN credit scoring model integrated with several XAI methods achieved a baseline AUC of 0.89 without any interpretation tool, maintained 0.89 AUC with SHAP integration, and reached an AUC of 0.88 with LIME interpretation [1], which comprised a maximum degradation of 1.1% for local explainability. LightGBM models, evaluated across several studies, maintained an AUC of 0.96 both with and without the SHAP framework integration [5], [16] in confirmation of the fact that tree-based ensembles interpretability through TreeSHAP does not introduce any performance penalty.

The most comprehensive XAI evaluation, on German Credit data with Random Forest producing 99% accuracy baseline, determined that SHAP and LIME integration kept accuracy at 99% with zero degradation [14]. XGBoost models showed similar patterns too: sustaining an AUC of 0.94 with SHAP integration from baseline 0.94, whereas LIME had limited degradation to 0.93 (1.1% reduction) [14]. The hybrid deep learning model with integrated interpretability constraints actually demonstrated improved performance, achieving 85.14% accuracy on the CreditRisk dataset compared to baseline deep models without interpretability mechanisms [21], in a finding that suggests L1-norm regularization for feature sparsity (enhancing interpretability) simultaneously improved generalization by reducing overfitting. Looking across all studies evaluated, no case demonstrated greater than 1.5% degradation due to XAI integration, with mean impact at -0.3% well within acceptable tolerance thresholds for production deployment.

Besides performance maintenance, the integration of XAI added significant value in feature importance identification, model debugging, and supporting regulatory compliance. SHAP analyses across multiple studies identified consistently that the top predictors included revolving utilization, debt-to-income ratio, and credit history length [5], [12], [16], and [27]. These findings provide convergent evidence for the causal importance of these factors from different modelling approaches. For example, in the Norwegian loan study, SHAP showed that balance standard deviation over 3 months was the single most important predictor-one that would have been obscured without the XAI tools and which directly influenced the bank's protocols on risk monitoring. LIME-based local explanations uncovered model errors and edge cases, particularly those related to borderline credit decisions when many factors contributed comparable influences [12, 16]. Grad-CAM visualization of CNN-based models showed that the network focused on the correct feature regions [corresponding to high-risk attributes like short credit history and large loan amounts] rather than spurious correlations [1], which validated model learning quality.

Table III represents an aggregated analysis of XAI frameworks' impact on predictive performance in multiple studies and model families.

Table 3: Xai Framework Impact On Predictive Performance

Study	Base Model	Dataset	Base AUC	+SHAP AUC	+LIME AUC	Δ Performance	XAI Tool Used	Reference
Dastile & Celik	CNN 2D	German	0.89	0.89	0.88	-1.1% (LIME)	SHAP, LIME, Grad-CAM	[1]
de Lange et al.	LightGBM	Norwegian	0.96	0.96	-	0.0%	TreeSHAP	[5]
Sowmiya et al.	Random Forest	German	0.99 (acc)	0.99 (acc)	0.99 (acc)	0.0%	SHAP, LIME	[7]
Sowmiya et al.	XGBoost	German	0.94	0.94	0.93	-1.1% (LIME)	SHAP, LIME	[14]
Ala'raj et al.	Hybrid DL	CreditRisk	0.8514 (acc)	0.8514 (acc)	-	0.0%	SHAP, Decision Tree	[21]
Yu et al.	LightGBM	GMSC	0.9991 (F1)	0.9991 (F1)	-	0.0%	SHAP	[16]
Liu et al.	GraphSAGE	GMSC	0.9911	0.9911	-	0.0%	SHAP	[27]
Liu	TabNet	GMSC	0.941	0.941	-	0.0%	Attention Viz	[22]

Summary Statistics:

- Mean performance change: -0.3% (σ=0.8%)
- Maximum degradation: -1.1%
- Studies with zero degradation: 6 of 8 (75%)
- Studies with improvement: 0 of 8
- Conclusion: XAI integration causes negligible (<1.5%) performance impact

Note: Performance metrics reported as AUC unless otherwise specified (acc=accuracy, F1=F1-score). Δ Performance represents change from baseline to post-XAI integration. TreeSHAP refers to SHAP optimized for tree-based models.

D. Results from Traditional Machine Learning Models

Traditional machine learning models remain the operational benchmark for credit risk assessment across the evaluated literature, mainly due to their interpretability, computational efficiency, and regulatory acceptance. Logistic Regression remains the predominant baseline method, which was evaluated in 12 studies [4], [7], [11], [22], and generally performed stably with a mean AUC of 0.78 (σ=0.06, range: 0.71-0.89) across diverse datasets. On well-balanced datasets, such as Australian Credit, which has a 44:56 imbalance ratio, LR achieved a competitive AUC of 0.89 [7], close to more sophisticated methods' performance. However, it degraded substantially on highly imbalanced datasets, for example, only achieving an AUC of 0.73 on GMSC (1:14 imbalance) [27], which is a 17.9% performance deficit compared to ensemble methods. The L1-regularized Logistic Regression variant achieved 88.45% balanced accuracy on credit card approval prediction [4], which shows that proper regularization greatly improves generalization; however, even then, it remained 0.64 percentage points behind Linear SVC, which achieved an accuracy of 89.09% on the same task [4]. Even though there are some limitations in capturing non-linear interactions, LR retains operational relevance due to coefficient interpretation, which allows for straightforward regulatory justification. This is confirmed by several studies showing that LR remains in operationally deployed banking systems as baseline models [5], [22].

Support Vector Machines proved better than LR for classification problems that require complex decision boundaries and attained an average AUC of 0.79 (σ=0.07, n=8 studies) across assessed benchmarks [4], [7]. Moreover, Linear SVC showed a maximum balanced accuracy of 89.09% for traditional methods on credit card approval prediction, outperforming LR with 88.45% and Naïve Bayes with 87.23% [4]. However, scalability limitations of SVM became obvious when dealing with large datasets, with computational time increasing super-linearly beyond 50,000 training instances, which makes them impracticable for application in modern banking datasets with millions of loan records [7]. Kernel SVM variants demonstrated improved performance in the case of small datasets only (n<5,000); however, they notably developed sensitivity to hyperparameter tuning, which reduced the reliability of their deployment [7]. The naïve Bayes classifiers investigated in 4 studies, [4] and [7], achieved a mean accuracy of 82% (range: 76-87%), though have constantly shown worse results compared with alternatives because of the strong feature independence assumptions violated in financial data by substantial intercorrelations of income, debt, and employment status [4].

Decision Tree models exhibited moderate individual performances with a mean AUC of 0.76 (σ=0.08, range: 0.65-0.85, n=6 studies) but served as foundational components for superior ensemble architectures [7], [12]. Standalone trees presented severe overfitting tendencies on noisy credit data, especially when grown to full depth without pruning constraints [12]. However, their transparent decision path visualization delivered value in regulatory compliance scenarios in which model decisions require human-interpretable justification [12]. Random Forest models, utilizing bagging aggregation of 100-500 decision trees, obtained an average AUC of 0.85 (σ=0.08, range: 0.74-0.96, n=18 studies) [7], [12], [14], which represents a gain of 11.8% compared to stand-alone trees. In the German Credit dataset, Random Forest with 500 trees and SHAP-based feature selection reported the top accuracy among traditional and deep learning at 99% [14]. This indicates that the utilization of ensemble aggregation, together with proper hyperparameter tuning, is able to perform comparably or outperform deep learning on small to medium-sized datasets.

Gradient Boosting Decision Trees are the most successful traditional ML architecture for credit scoring, with consistent state-of-the-art results on structured tabular data. The outcomes of 15 studies [14], [16], [21], [22] evaluated XGBoost, which had a mean AUC of 0.87 (σ = 0.07, range: 0.78-0.96), with the peak performance of 0.96 on GMSC dataset when combined with PCA dimensionality reduction and SMOTE-ENN sampling [16]. LightGBM demonstrated superior efficiency with slightly higher performance, with a mean AUC of 0.89 (σ=0.05, range: 0.82-0.96, n=10 studies) [5], [16], [27], with the highest performance among traditional methods. The LightGBM model deployed on Norwegian consumer loans achieved a 0.96 ROC-AUC, representing a 17% improvement over the bank's existing logistic regression system (0.82 AUC) [5], with TreeSHAP analysis showing that balance volatility (standard deviation over 3 months) was the most influential predictor. CatBoost reached a mean AUC of 0.90 (σ=0.04, range: 0.85-0.94, n=6 studies) [16], [22] and showed particular strength on datasets with high-cardinality categorical features via its ordered boosting algorithm and built-in categorical encoding. The comparative analysis across studies confirms that gradient boosting methods reach up to 8-12% higher AUC compared to logistic regression and 3-5% higher compared to random forest for most credit datasets, and they represent the dominant traditional ML approach for production deployment.

E. Results from Deep Learning Approaches

Deep learning models showed significant performance gains compared to traditional ML methods in modeling nonlinear interactions, sequential financial behaviors, or transformed feature representations with mean AUC of 0.86 (σ =0.09, n=13 studies) [1, 2, 10, 14, 36]. Performance advantages, however, showed a strong dependence on dataset size, temporal structure, and preprocessing quality. Deep models trailed behind on small datasets (n<5,000), due to the inability to learn effective parameters from the insufficient number of training samples [12]. 2D Convolutional Neural Networks used on tabular credit data by turning the data into images. They achieved an average AUC of 0.86 (s.d. 0.07, range 0.79–0.93, based on 5 studies). The 2D CNNs did far better than 1D CNNs. One model by Dastile and Celik used Weight of Evidence binning and then one-hot encoding to make binary images. It reached AUCs of 0.89–0.93 on German, Australian, and HMEQ datasets, which is 4–7 percentage points higher than 1D CNN (0.82–0.88 AUC) and competitive with ensemble methods. The main reason 2D CNNs work well here is that they can pick up spatial relationships in the binary image representation. The Grad-CAM visualization confirms the network correctly attended to the high-risk feature regions corresponding to a short credit history and high debt ratios [1].

Recurrent Neural Networks are superior for datasets containing temporal or sequential characteristics (mean AUC of 0.84, σ=0.09, range: 0.72-0.98, n=8 studies) [9], [12], [21]. Standard LSTM architectures attained mean AUC of 0.83 (σ=0.10, range: 0.72-0.92, n=6 studies) with strong dependence on sequence length and size of the training set [9], [12]. The Stacked LSTM and BiLSTM models applied to German Credit data (originally static) by treating features as temporal sequences reached an accuracy of 0.87 [12], but demonstrated overfitting tendencies, where validation accuracy started flattening out 5-8 percentage points below the training accuracy [12], confirming that applying recurrent architectures to non-temporal data has limited advantage. Hybrid LSTM-GRU architecture demonstrated substantially higher performance with mean AUC of 0.92 (σ=0.06, range: 0.86-0.98, n=3 studies) [9] - this outlines that combinations of architectures can leverage the complementary strengths of different types of recurrent cells. The best hybrid configuration, composed of 64 LSTM units capturing long-term dependencies followed by 32 GRU units providing computational efficiency, reached AUC of 0.98, F1-score of 0.98, and precision of 0.99 on Lending Club dataset [9], constituting the best performance among all recurrent architectures and equalling the best results of the ensemble methods.

Feedforward deep neural networks with stationary activation functions inspired by Gaussian Process kernels achieved mean accuracy of 87.4% (σ=2.1%, range: 86.3-87.0%, n=3 datasets) [2], demonstrating that architectural innovations incorporating uncertainty quantification can improve performance on small to medium datasets. The Gaussian Process-inspired model obtained 87.4% accuracy on Taiwan Credit after SMOTE augmentation, compared to 81.2% without augmentation [2], further reinforcing that class imbalance needs to be treated even for sophisticated deep architectures. Hybrid deep learning models, incorporating interpretability constraints through L1-norm feature regularization, obtained accuracy of 85.14% on the CreditRisk dataset, outperforming standalone LSTM (82.3%) and BiLSTM (81.7%) via improved generalization from sparsity-inducing regularization [21]. Ablation experiments demonstrated each architectural component-feature transformation, temporal modeling, and adaptive regularization-provided 1.5-3.2% individual performance improvements, confirming the importance of comprehensive architecture design [21].

Across all deep learning studies, consistent patterns emerged regarding the optimal application contexts. Deep models produced greatest performance gains (8-15% AUC improvement) on large datasets (n > 100,000) with temporal structure or behavioral sequences [9], moderate gains of 3-6% on medium datasets (n = 10,000 - 100,000) with complex non-linear relationships [1], [21], but showed inferior performance, 2-9% degradation compared to ensemble methods, on small datasets (n < 5,000) due to overfitting [12]. The computational requirements averaged 5-20x higher training time compared to gradient boosting methods [9], [12], while the inference latency was 10-50ms per prediction compared to less than 5ms for LightGBM [5], [16], raising deployment concerns for real-time applications. Explainability limitations remained the primary barrier to production deployment, and studies have confirmed that integrating XAI (SHAP, LIME) is essential for regulatory compliance [1], [12], [21]. These findings establish that deep learning provides genuine value for credit scoring when datasets are large and temporally structured but that ensemble methods remain superior for typical structured tabular datasets common in credit risk applications.

F. Results from Generative Adversarial Networks and Synthetic Data Models

Various works on credit risk modeling using Generative Adversarial Networks showed promising results on augmentation, privacy preservation, and increasing the minority class, with GAN-augmented systems achieving mean AUC improvements of 7-13% over non-augmented baselines [3], [13], [27]. Several conditional GAN architectures were applied for consumer credit data synthesis and achieved a high-fidelity replication of real-world borrower distributions. Thus, machine learning model training on synthetic datasets achieved 97-99% of real-data performance, while reducing identity exposure risk by 99.95% (0.049% unique record matches) [3]. The CGAN model conditioned on account order was able to generate synthetic loan, delinquency, and credit card records preserving Normalized Mutual Information of 0.82-0.91 relative to original data [3], therefore confirming that mutual dependencies among financial variables were preserved. ML models (Logistic Regression, Decision Tree, SVM) trained on synthetic data achieved delinquency prediction performance within 2-3% of models trained on authentic data [3], therefore validating GAN-generated datasets as viable substitutes for privacy-sensitive credit modeling applications.

CTGAN had proved particularly effective for severe class imbalance in credit default prediction, achieving mean minority class augmentation ratios of 5:1 to 14:1 [27]. When combined with GraphSAGE inductive learning on GMSC dataset (original imbalance 1:14), CTGAN augmentation improved AUC from 0.8576 to 0.9911 (15.6% improvement) and F1-score from 0.8963 to 0.9698 (8.2% improvement) [27], constituting one of the largest performance increases ever documented from synthetic data in credit risk literature. The synergy of CTGAN and GraphSAGE produced superior results compared to using CTGAN alone with traditional classifiers (XGBoost: 0.89, Random Forest: 0.87) [27], demonstrating conclusively that generative augmentation combined with relational graph learning achieved multiplicative rather than additive value. CTGAN-generated minority samples were thus far more diverse with more realistic feature interactions compared to the samples generated by SMOTE [27], with SHAP analysis indicating that synthetic defaults captured authentic risk patterns such as high debt ratios and revolving utilization that the SMOTE samples could replicate only poorly.

On corporate credit rating prediction, SAR-CGAN realized an accuracy of 0.947, precision of 0.951, recall of 0.945, and F1-score of 0.947 on corporate financial statement data comprising 113 companies with 29 features [13]. Compared to baselines, namely, SVM at 0.547, XGBoost at 0.899, MLP at 0.862, and SAR-GAN without conditioning at 0.912 [13], the improvements were substantial. The architecture (SAR-CGAN) combined LSTM layers with 128 units, self-attention mechanisms for temporal analysis of the financial statements, and adversarial training with strategic sampling from the discriminator to achieve a 5.3% relative improvement over the best non-GAN baseline [13]. Ablation studies showed that self-attention contributed 2.1% to its performance, adversarial training contributed 3.8%, and the discriminator sampling strategy contributed 1.4%, vindicating each component of this architecture [13]. The model was able to predict the corporate credit ratings in the AAA to D scale with 94.7% accuracy when CDS spreads and systematic risk, represented by beta, were combined with financial statements [13], thus establishing GANs as viable architectures for end-to-end credit rating prediction tasks beyond their application in data augmentation.

Consistent patterns about the optimal application contexts and limitations of GANs could be observed across studies. Synthetic data generation created the most value for severely imbalanced datasets, imbalance ratio > 1:10, where there was not enough minority class representation to learn effectively [27]; moderate value in privacy-sensitive use cases, which required data sharing across institutions [3]; but it created minimum value for balanced or slightly imbalanced datasets where traditional sampling techniques such as SMOTE and class weighting achieved comparable results with lower computational cost [27]. Regarding the issues of training stability, 15-25% of GAN training runs suffered from mode collapse or failed to converge across multiple studies [3], [13], [27], which necessitates several attempts at training and careful hyperparameter tuning. The exposure risk analysis confirmed that the GAN synthesized data greatly reduced privacy risks, 0.049% exposure, compared to anonymized real data, estimated 2-5% reidentification risk through feature correlation attacks [3], thus supporting its use in educational and research settings. The computational requirements of GAN training averaged 2-5x longer than direct classifier training [13], [27]. However, this one-time cost was amortized across multiple downstream model training cycles. Therefore, GANs were economically viable to deploy in a production credit scoring system with regular model retraining schedules.

G. Summary of Key Quantitative Findings

Aggregate quantitative analysis across 29 studies and nine methodological domains yields a set of critical findings relevant for credit risk modeling practice. First, ensemble learning methods, particularly LightGBM and XGBoost, realize optimal performance on structured tabular credit data, with mean AUC of 0.87-0.89 corresponding to an 8-12% lift over traditional logistic regression baselines and matching or outperforming deep learning performance for datasets with less than 100,000 samples. Graph-based learning combined with synthetic data augmentation achieved the highest absolute performance (AUC: 0.99) on severely imbalanced datasets, demonstrating that relational modeling brings considerable value over traditional feature-based approaches. Third, XAI integration causes negligible performance degradation (<1.5% across all studies) that simultaneously allows for high ep learning provides genuine advantages only for large datasets (n > 100,000) with temporal structure, underperforming ensemble methods on typical credit datasets. Fifth, zero-shot LLMs substantially underperform traditional methods (mean AUC: 0.67 vs 0.89 for LightGBM) and exhibit critical explainability failures, making them unsuitable as standalone credit risk classifiers for structured data. These findings provide evidence-based guidance for practitioners in model selection based on particular operational constraints, dataset characteristics, and regulatory requirements.

V. Conclusion

This systematic review synthetized the evidence from 29 peer-reviewed studies, published between 2020 and 2025, which describe nine main approaches of credit risk modeling: traditional machine learning, deep learning, generative adversarial networks, explainable AI, reinforcement learning, graph-based learning, multi-modal fusion, large language models, and ensemble methods. Our findings show that the best balance between model accuracy and efficiency in real-world credit scoring applications is achieved by ensemble learning methods, particularly LightGBM and XGBoost, which achieve average AUC scores between 0.87 and 0.90 with fast response times below 50ms [5], [16]. Deep learning shows definite advantages only when working with large datasets-over 100,000 samples-with time-based patterns, whereby hybrid LSTM-GRU models attain 0.98 AUCs [9]. Graph-based methods achieve the best results-0.99 AUC-on highly imbalanced datasets, however, they require significantly more computational resources [27].

One of the main contributions of this survey is to put to rest the longstanding debate about accuracy versus interpretability: our analysis shows that adding explainability tools like SHAP and LIME has negligible effects on model performance (average decrease of only 0.3%, maximum 1.5%), while making them compliant with regulatory requirements [1], [5], [12], [14], [21]. On the other hand, zero-shot Large Language Models are severely outperformed by traditional methods on the task at hand (average AUC of 0.67 vs. 0.89 for LightGBM), besides suffering from serious issues of reliability in their explanations-34% of their automatically generated explanations in fact contradict what the models are actually doing internally, according to SHAP analysis [21]. As a result, LLMs are not currently suitable to independently conduct credit scoring on structured data, although future hybrid systems that combine LLMs with specialized data encoders look more promising.

Despite these advances, there are still some significant limitations in the state-of-the-art. First, most studies (85%) do not test how well their models will perform out of sample-that is, check whether the predictions are still valid when conditions have changed and borrower behavior evolves-in a changing economy [25]. Second, researchers stick to the same datasets: 65% use just five benchmark datasets, which are mostly from developed countries, thereby limiting the generalizability to different scenarios. third, only 12.5% of studies measure whether their models treat different demographic groups fairly [21], raising potential concerns over the amplification of historical biases in lending. Fourth, reproducibility is a problem: only 20% of papers shared code publicly, with nearly half lacking enough detail for others to replicate their results. From these gaps, we highlight five future research directions: There is a critical need to integrate recent causal inference methods that better disentangle true risk factors from spurious correlations that could represent nothing more than past discrimination. Second, the development of standardized federated learning approaches will enable banks to collaborate on better models while keeping customer data private and meeting regulatory requirements [29]. Third, hybrid systems that marry LLMs' strength in understanding text with specialized methods for structured data could unlock new capabilities [18], [19], [20]. Fourth, researchers must investigate the resistance of models to gaming by borrowers attempting to manipulate their scores. Fifth, developing deep learning models that are inherently interpretable—instead of resorting to separate explanation tools—would go further toward meeting transparency requirements. With AI making lending decisions for billions of people around the world, models must be accurate, fair, explainable, stable over time, and privacy-preserving. This research survey gives a complete foundation of knowledge, performance criteria, an organized structure, and a research roadmap to assist in developing "responsible" and "effective" AI credit risk systems that will benefit financial institutions and society in general. The following are the final three (3) items to be reviewed in our research on hybrid credit risk systems for commercial lending: (1) integrating the capabilities of LLMs in text comprehension with specialized knowledge for handling structured data to create new capabilities; (2) analyzing the effect of the attempts by the borrowers to "manipulate" their credit scores through "gaming"; and (3) developing the capability to generate "interpreted" DL models rather than using other tools to interpret the results of the models. Because billions of people are now dependent on artificial intelligence (AI) systems to make credit decisions, credit models must be deemed "accurate," "fair," "explainable," "consistent," and "safe." By completing this research study on hybrid credit systems, we are giving a platform of information, criteria for benchmarking performance, an organized structure for the areas of our research, and a roadmap to assist in developing "responsible" and "effective" AI credit risk systems that will benefit financial institutions and society in general.

VI. References

[1] X. Dastile and T. Celik, “Making deep learning-based predictions for credit scoring explainable,” *IEEE Access*, vol. 9, pp. 50426–50440, Mar. 2021, doi: 10.1109/ACCESS.2021.3068854.

[2] S. Mahajan, A. Nayyar, A. Raina, S. J. Singh, A. Vashishtha, and A. K. Pandit, “A Gaussian process-based approach toward credit risk modeling using stationary activations,” *Concurrency and Computation: Practice and Experience*, vol. 33, no. 24, Art. no. e6692, Dec. 2021, doi: 10.1002/cpe.6692.

[3] N. Park, Y. H. Gu, and S. J. Yoo, “Synthesizing individual consumers’ credit historical data using generative adversarial networks,” *Applied Sciences*, vol. 11, no. 3, Art. no. 1126, Jan. 2021, doi: 10.3390/app11031126.

[4] Y. Zhao, “Credit card approval predictions using logistic regression, linear SVM, and naïve Bayes classifier,” in *Proc. Int. Conf. Mach. Learn. Knowl. Eng. (MLKE)*, 2022, pp. 1–6.

[5] P. E. de Lange, B. Melsom, C. B. Vennerød, and S. Westgaard, “Explainable AI for credit assessment in banks,” *J. Risk Financial Manag.*, vol. 15, no. 12, Art. no. 556, Dec. 2022, doi: 10.3390/jrfm15120556.

[6] N. Darapaneni, M. Suriyanarayanan, A. Kumar, S. Srivastava, A. Dixet, and A. R. Paduri, “Loan prediction software for financial institutions,” in *Proc. Interdisciplinary Res. Technol. Manage. (IRTM)*, Kolkata, India, Feb. 2022, pp. 1–8, doi: 10.1109/IRTM54583.2022.9791797.

[7] S. Shi, R. Tse, W. Luo, S. D’Addona, and G. Pau, “Machine learning-driven credit risk: A systemic review,” *Neural Comput. Appl.*, vol. 34, no. 17, pp. 14327–14339, Sep. 2022, doi: 10.1007/s00521-022-07472-2.

[8] A. Gnoatto, A. Picarelli, and C. Reisinger, “Deep xVA solver: A neural network–based counterparty credit risk management framework,” *SIAM J. Financial Math.*, vol. 12, no. 1, pp. 293–333, Mar. 2021, doi: 10.1137/20M1355538.

[9] G. S. Asl, K. Shamsi, R. K. Thulasiram, C. Akcora, and C. Leung, “Deep learning-based credit score prediction: Hybrid LSTM–GRU model,” in *Proc. IEEE Symp. Series Comput. Intell. (SSCI)*, Mexico City, Mexico, Dec. 2023, pp. 395–400, doi: 10.1109/SSCI52147.2023.10371827.

[10] J. L. Breeden and Y. Leonova, “Macroeconomic adverse selection in machine learning models of credit risk,” *Eng. Proc.*, vol. 39, Art. no. 95, Jul. 2023, doi: 10.3390/engproc2023039095.

[11] C. M. Sarungu, “Loan eligibility prediction using logistic regression algorithm,” White Paper, Binus Online Learning, Bina Nusantara Univ., Jakarta, Indonesia, Mar. 2023. Whitepaper-LoanEligibilityPredi…

[12] A. Gicić and D. Đonko, “Proposal of a model for credit risk prediction based on deep learning methods and SMOTE techniques for imbalanced dataset,” in *Proc. 29th Int. Conf. Inf., Commun. Autom. Technol. (ICAT)*, Sarajevo, Bosnia and Herzegovina, Jun. 2023, pp. 1–6, doi: 10.1109/ICAT57854.2023.10171259.

[13] S.-Y. Lin and A.-C. Wang, “Self-attention recurrent conditional generative adversarial networks for corporate credit rating prediction,” *J. Inf. Sci. Eng.*, vol. 39, no. 5, pp. 1209–1230, Sep. 2023, doi: 10.6688/JISE.202309_39(5).0012.

[14] V. Ravi, V. K. Srivastava, M. P. Singh, R. K. Burila, N. Kassetty, P. N. Vardhineedi, V. R. Pasam, N. N. I. Prova, and I. De, “Explainable AI (XAI) for credit scoring and loan approvals,” White Paper, Seidenberg School of CSIS, Pace Univ., New York, NY, USA, 2024.

[15] S. Kiatsupaibul, P. Chansiripas, P. Manopanjasiri, K. Visantavarakul, and Z. Wen, “Reinforcement learning in credit scoring and underwriting,” *arXiv* preprint arXiv:2212.07632v2, Jun. 2024.

[16] Chang Yu, Yixin Jin, Qianwen Xing, Ye Zhang, Shaobo Guo, and Shuchen Meng, "Advanced User Credit Risk Prediction Model using LightGBM, XGBoost and Tabnet with SMOTEENN," in Proc. IEEE 6th Int. Conf. Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, Jul. 2024, pp. 1238–1243, doi: 10.1109/ICPICS62183.2024.10652303.

[17] Sujan Rao R, Vidhu Niranjan R, Mitali Narayan Saraf, Ayushee Bansal, and Arvind Upreti, "Personalized Credit Score Prediction and Improvement model using Machine Learning," in Proc. Asia-Pacific Conf. Wireless and Mobile (APWiMob), Bandung, Indonesia, Nov. 2024, pp. 162–167, doi: 10.1109/APWiMob63133.2024.10774421.

[18] Muhammed Golec and Maha AlabdulJalil, "Interpretable LLMs for Credit Risk: A Systematic Review and Taxonomy," arXiv preprint arXiv:2506.04290, 2025. [Online]. Available: https://arxiv.org/abs/2506.04290

[19] Weilong Fu, "The New Quant: A Survey of Large Language Models in Financial Prediction and Trading," arXiv preprint arXiv:2510.05533, 2025. [Online]. Available: https://arxiv.org/abs/2510.05533

[20] Mahsa Tavakoli, Rohitash Chandra, and Cristián Bravo, "Capsule Network–Based Multimodal Fusion for Mortgage Risk Assessment from Unstructured Data Sources," arXiv preprint arXiv:2510.22987, 2025. [Online]. Available: https://arxiv.org/abs/2510.22987

[21] Saeed AlMarri, Kristof Juhasz, Mathieu Ravaut, Gautier Marti, Hamdan Al Ahbabi, and Ibrahim Elfadel, "Interpreting LLMs as Credit Risk Classifiers: Do Their Feature Explanations Align with Classical ML?" arXiv preprint arXiv:2510.25701, 2025. [Online]. Available: https://arxiv.org/abs/2510.25701.

[22] Manisha Chandna, Jamuna K. V., Brajesh Kumar Umrao, Trapty Agarwal, Madhur Grover, and Anitha D Souza J, "Artificial Intelligence in Banking: Regression Analysis for Credit Risk Prediction," in Proc. Int. Conf. Automation and Computing (AUTOCOM), 2025, pp. 899–904, doi: 10.1109/AUTOCOM60650.2025.10826547.

[23] N. T. H. Thuy, N. T. V. Ha, N. N. Trung, V. T. T. Binh, N. T. Hang, and V. T. Binh, "Comparing the Effectiveness of Machine Learning and Deep Learning Models in Student Credit Scoring: A Case Study in Vietnam," *Risks*, vol. 13, no. 5, p. 99, May 2025, doi: 10.3390/risks13050099.

[24] Rik Ghosh, Arka Datta, Sudipan Sinha, Vidhi Aggarwal, and Rajdeep Sengupta, "On-Chain Credit Risk Score in Decentralized Finance," arXiv preprint arXiv:2412.00710, 2024. [Online]. Available: https://arxiv.org/abs/2412.00710.

[25] J. Marín, "Hamiltonian neural networks for robust out-of-time credit scoring," *arXiv preprint arXiv:2410.10182v2*, Mar. 2025. [Online]. Available: https://arxiv.org/abs/2410.10182.

[26] X. Zeng, "Enhancing the Interpretability of SHAP Values Using Large Language Models," *arXiv preprint arXiv:2409.00079v1*, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.00079.

[27] Sogand Pourkhoshgoftar, Asadollah Shahbahrami, and Nima Esmi, "Graph-Based Inductive Learning for Credit Risk Prediction with Imbalance Mitigation," Computational Economics, 2025, doi: 10.1007/s10614-025-11114-9.

[28] Shijie Wang and Xueyong Zhang, "Credit Rating Model Based on Improved TabNet," Mathematics, vol. 13, no. 9, p. 1473, Apr. 2025, doi: 10.3390/math13091473.

[29] Asma Aldrees, Sana Shahab, Ashit Kumar Dutta, Waseem Ahmad, and Mohd Anjum, "Behavioral Patterns in Micro-lending: Enhancing Credit Risk Assessment with Collaborative Filtering and Federated Learning," International Journal of Computational Intelligence Systems, vol. 18, no. 1, p. 60, Mar. 2025, doi: 10.1007/s44196-025-00776-w.