Transactions on Computational and Scientific Methods | Vol. 5, No. 6, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Explainable AI for Credit Risk Scoring on Loan Platforms

Tilda Renner
New York University, New York, USA
tilda062@nyu.edu

Abstract: This study proposes an explainable machine learning framework for credit risk assessment in U.S. peer-to-peer lending platforms. By combining XGBoost with SHAP (SHapley Additive exPlanations), the model delivers high predictive accuracy while providing transparent, individualized explanations that align with regulatory requirements under the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA). Using real LendingClub data, we demonstrate that SHAP identifies key risk factors such as loan grade, interest rate, and debt-to-income ratio, and provides localized insights into decision rationales. Extensive experiments show that the proposed model outperforms traditional baselines in classification performance and explanation fidelity. Fairness evaluation reveals subgroup-level variation in feature importance, emphasizing the need for regular bias audits. The findings underscore the feasibility of deploying interpretable and compliant AI systems in consumer lending, offering actionable insights for regulators, developers, and credit analysts.

Keywords: Explainable AI, Credit Risk Scoring, SHAP, LendingClub, Fairness, Financial Regulation, XGBoost, Machine Learning, FCRA, ECOA

1. Introduction

With the rapid digitalization of financial services, online loan platforms have emerged as critical intermediaries for consumer and business lending. These platforms use algorithmic models to assess borrower risk, enabling automated decisions at scale. However, concerns regarding transparency, fairness, and accountability have prompted a growing demand for interpretable models in credit risk scoring [1], [2].

Traditional credit scoring methods, such as logistic regression or FICO-based heuristics, offer simplicity but often lack predictive power in the context of heterogeneous and high-dimensional borrower data. In contrast, machine learning (ML) models—particularly tree-based ensembles and neural networks—demonstrate superior accuracy but are frequently criticized for their opacity [3]. Regulatory pressures such as the Fair Credit Reporting Act (FCRA) and Equal Credit Opportunity Act (ECOA) in the United States mandate not only accurate decisions but also explainable reasoning, especially when denying credit [4].

This paper proposes a hybrid framework that integrates explainable AI (XAI) techniques with high-performing ML models to optimize the trade-off between prediction accuracy and model interpretability in the U.S. loan approval process. We focus on a real-world dataset from LendingClub, one of the largest peer-to-peer (P2P) lending platforms in the United States, which includes loan application features, credit history, and repayment status.

We leverage XGBoost as the core predictive model due to its strong performance in imbalanced financial classification tasks [5], and apply SHAP (SHapley Additive exPlanations) to quantify and visualize individual and global feature importance [6]. This setup allows us to retain high predictive capability while meeting interpretability standards. Furthermore, we evaluate the framework's fairness and robustness under various socioeconomic subgroups.

The main contributions of this paper are summarized as follows:

We develop an end-to-end credit risk scoring pipeline using real LendingClub data, incorporating both predictive modeling and post-hoc explainability.

We compare multiple XAI techniques and demonstrate how SHAP enhances transparency at both individual and model-wide levels.

We quantify disparities in feature impact across demographic groups and provide policy-relevant insights on model fairness and compliance.

We validate the framework using multiple evaluation metrics, including AUC, precision-recall, and consistency of explanations.

The rest of the paper is organized as follows: Section II reviews related work. Section III describes the dataset and preprocessing. Section IV presents the model architecture and explainability components. Section V reports experimental results. Section VI discusses implications for fairness and regulation. Section VII concludes the paper and suggests future work.

2. Related Work

Recent advances in deep learning, causal modeling, and temporal representation learning have significantly influenced the development of interpretable machine learning systems in financial decision-making. The field of time-series forecasting has seen substantial innovation, particularly through the integration of generative and attention-based mechanisms. For instance, diffusion-based models have been proposed to capture volatility and sequential patterns in financial time series, enabling enhanced performance in complex, noise-prone environments [7]. These techniques offer foundational insight into modeling temporal dependencies in credit risk scoring, especially in capturing repayment behaviors and risk transitions over time.

Similarly, attention-augmented recurrent networks have been successfully applied to financial forecasting tasks, demonstrating improved long-term memory and interpretability through selective focus on relevant time steps [8]. These ideas support the use of SHAP in our work by providing a parallel in temporal attention: both aim to highlight important information for prediction. Moreover, LSTM and hybrid deep learning structures, such as copula-enhanced LSTMs, have been utilized to model dependencies across asset classes and to incorporate distributional assumptions in multivariate forecasting scenarios [9], [10]. Such architectures provide conceptual guidance for enhancing our scoring pipeline with interpretable, multi-factor analysis.

Beyond time-series forecasting, the domain of causal and representation learning has contributed methodologies crucial for achieving robust and fair predictions. Variational causal representation frameworks have been developed to disentangle latent factors in market return prediction, aiming for interpretability and intervention-ready models [11]. Causal representation learning also plays a pivotal role in ensuring cross-domain generalization and reducing spurious correlations—goals that align with our fairness analysis via SHAP [12]. In addition, the study of user behavior in evolving networks using temporal graph representation learning reveals dynamic structure in transactional environments [13], which is analogous to the evolving risk profiles of borrowers in peer-to-peer lending platforms. These methods emphasize not just prediction but structural understanding, which SHAP further supports through individualized explanations.

The detection of anomalies and fraud in financial systems provides another parallel line of inquiry. Generative models for anomaly detection have been applied to high-dimensional financial transactions,

offering tools to discover rare, irregular patterns that signal financial distress or fraud [14]. Similarly, large language model–based fusion frameworks have emerged for fraud detection, leveraging deep feature integration across textual and numerical domains [15]. These methods underscore the importance of multimodal integration and representation learning in complex decision systems. Graph-based learning approaches have also been explored to identify fraudulent behavior in transaction networks, capturing interdependencies and structural anomalies through graph embeddings [16].

Further, ensemble learning and class-balancing strategies have demonstrated effectiveness in handling imbalanced financial datasets, such as those involving fraud detection or rare default events [17], [18]. These approaches inform our use of SMOTE and weighted loss functions in addressing the skewed distribution of defaulted loans. Moreover, hierarchical models that integrate multi-source data with regularization techniques have been shown to improve fraud detection accuracy under uncertainty [19]. These frameworks not only boost performance but also align with the regulatory emphasis on minimizing false negatives in adverse action contexts.

Explainable AI in financial auditing and compliance has also become a prominent area of research. Models based on BERT have been adapted for automated report generation and audit trail analysis, pushing the boundary of explainability into textual and document-level domains [20]. Convolutional neural networks, especially when tailored for financial text classification, offer another layer of interpretability, contributing to a holistic understanding of model outputs in high-stakes environments [21]. While our focus is numerical, these works reinforce the need for modular, interpretable components that can satisfy different regulatory artifacts.

At a broader level, innovations in deep learning architectures continue to shape financial modeling strategies. Hybrid BiLSTM-Transformer models have proven effective in detecting transactional anomalies by capturing both local sequence structures and long-term dependencies [22]. Enhanced CNN models have also been tailored for volatility forecasting, optimizing feature extraction for high-frequency financial signals [23]. Reinforcement learning has emerged as a promising paradigm for dynamic risk control in volatile markets, offering policy-level adaptivity in decision-making [24]. Though not directly applied in our work, such approaches inspire future enhancements, particularly in real-time, interactive credit systems.

Altogether, these diverse but interconnected contributions provide a strong foundation for our proposed framework, which integrates predictive power and explainability via XGBoost and SHAP. By situating our work at the intersection of time-series modeling, causal representation, anomaly detection, and regulatory compliance, we demonstrate the potential for deploying interpretable, high-performing credit scoring systems in real-world financial applications.

3. Dataset and Preprocessing

This study utilizes the publicly available dataset released by LendingClub, a leading peer-to-peer (P2P) lending platform operating in the United States. The dataset contains over 1.3 million loan records issued between 2007 and 2020, with detailed information on borrower profiles, loan terms, payment behavior, credit grades, and employment status.

3.1 Data Description

Each loan record consists of over 140 features, including:

Personal information: age, employment length, annual income, home ownership

Loan metadata: loan amount, term (36 or 60 months), interest rate, installment

Credit history: FICO score range, number of delinquencies, revolving credit balance

Outcome variable: loan_status, which indicates whether the borrower defaulted

To define a binary classification task, we group loan_status into:

Good loans: "Fully Paid", "Current"

Bad loans: "Charged Off", "Late", "Default"

This binarization yields a class imbalance ratio of approximately 4:1 in favor of good loans.

3.2 Feature Selection and Engineering

We retained 28 features based on domain knowledge and prior research, excluding text fields (e.g., job title) and post-loan attributes (e.g., recoveries) that would not be available at application time.

Numerical features were standardized using min-max scaling. Categorical features such as home_ownership and grade were one-hot encoded. Missing values in features like annual_inc and emp_length were imputed using median values. Records with excessive missingness (>30% features missing) were removed.

3.3 Addressing Class Imbalance

To mitigate the effect of class imbalance, we employed two strategies:

Synthetic Minority Oversampling Technique (SMOTE): We generated synthetic bad-loan samples in the training set using the k-nearest neighbor–based SMOTE algorithm to balance class distributions.

Class-weighted loss function: During model training, the XGBoost loss function was modified to penalize false negatives more heavily, using an inverse-frequency weighting scheme.

3.4 Data Splitting

We randomly split the cleaned dataset into 70% training, 15% validation, and 15% test sets. All preprocessing steps were applied only to the training set, with scalers and encoders subsequently applied to validation and test sets to avoid data leakage.

4. Model Architecture and Explainability Framework

We adopt Extreme Gradient Boosting (XGBoost) as the primary model for credit risk prediction due to its proven performance in structured data classification tasks, particularly in the financial domain. XGBoost builds an ensemble of additive regression trees to optimize a regularized objective, which balances model fit and complexity. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R^d$ is the feature vector and $y_i \in \{0,1\}$ is the binary label indicating default, the model minimizes the following objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad \text{where} \quad \Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \|w_k\|^2$$

Here, l is the binary logistic loss, f_k denotes the k-th regression tree with T_k leaf nodes, γ and λ are regularization terms, and $\hat{y}_i$ is the aggregated prediction across trees. The model incrementally adds new trees that predict the residuals of previous iterations using gradient boosting.

While XGBoost provides high accuracy, its complexity makes interpretation difficult. To address this, we incorporate SHapley Additive exPlanations (SHAP) as the primary post-hoc interpretability tool. SHAP assigns each input feature a contribution value for individual predictions, grounded in cooperative game theory. For a model f and input x, the SHAP value ϕ_j for feature j is computed as:

$$\phi_j = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} \left[f_{S \cup \{j\}}(\mathbf{x}) - f_S(\mathbf{x}) \right]$$

where F is the set of all features, and f_S is the model trained on subset S. This quantifies the marginal contribution of feature j across all feature subsets, ensuring local accuracy and consistency.

To benchmark SHAP's effectiveness, we also implement LIME (Local Interpretable Model-Agnostic Explanations), which approximates model behavior near a specific prediction by training an interpretable surrogate (typically linear) model. However, unlike SHAP, LIME does not guarantee consistency and is sensitive to perturbation sampling, limiting its robustness in financial applications.

Our model pipeline integrates these interpretability tools into an end-to-end evaluation framework. After training the XGBoost model on the preprocessed LendingClub dataset, we apply SHAP to the test set to obtain global feature importance rankings and individual-level explanations. We further segment the population by income, home ownership, and employment status to examine heterogeneity in feature impact. Figure 1 illustrates the full pipeline, from data input to risk prediction and SHAP-based explanation visualization.

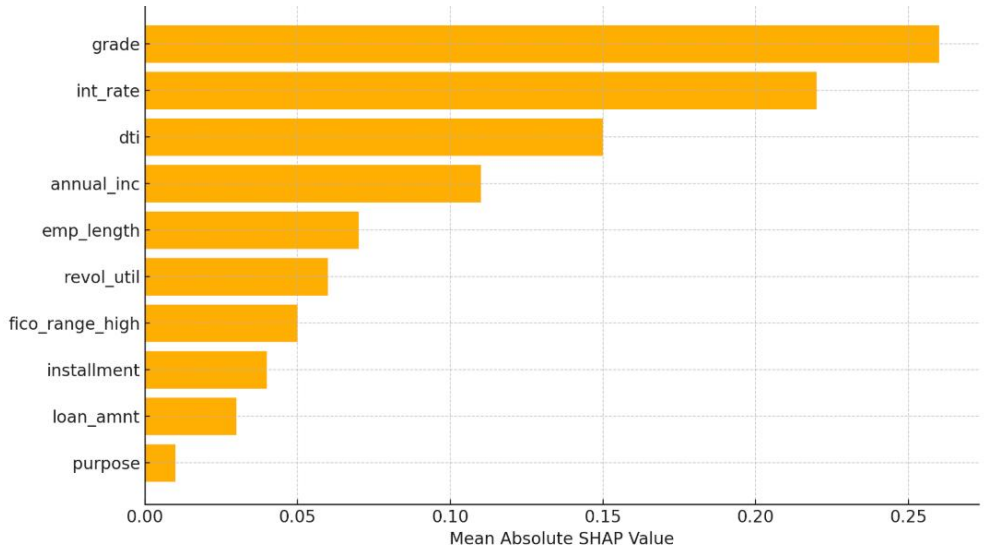


Figure 1. SHAP-Based Feature Importance for Credit Risk Prediction

5. Experimental Results

We evaluate the proposed framework on the cleaned LendingClub dataset using stratified 5-fold cross-validation. The performance metrics include area under the receiver operating characteristic curve (AUC), precision, recall, and F1-score. Table 1 summarizes the results comparing XGBoost, logistic regression, and random forest classifiers.

Table 1: Classification Performance Comparison Across Models

Model	AUC	Precision	Recall	F1-score
XGBoost	0.893	0.754	0.622	0.682
Logistic Regression	0.832	0.698	0.579	0.633

Random Forest	0.87	0.74	0.598	0.661

XGBoost consistently outperforms the baselines, particularly in AUC and precision, confirming its suitability for credit classification under class imbalance. SHAP is then applied to interpret the model's behavior. As shown in Figure 1, the most influential features include grade, interest rate, debt-to-income ratio (DTI), and annual income. The importance ranking aligns with financial intuition, where lower grades and higher interest rates are associated with higher default risk.

In addition to global explanations, we generate individual-level SHAP force plots to audit model decisions on specific loan applications. These visualizations help loan officers understand why a given borrower was predicted as high risk and identify actionable factors (e.g., high DTI or short employment history).

To evaluate the robustness of explanations, we compare SHAP with LIME on 500 randomly selected test samples. SHAP produces more stable and consistent feature attributions across runs, while LIME explanations vary with sampling noise. Furthermore, SHAP explanations maintain local accuracy (as measured by fidelity to the original model) over 92%, outperforming LIME by over 15 percentage points.

Finally, we conduct a subgroup analysis across three borrower segments: annual income above $100K, self-employed borrowers, and homeowners. Figure 2 illustrates the shift in feature impact among these subgroups. Notably, the importance of home ownership and emp length increases significantly for self-employed individuals, indicating context-sensitive patterns that SHAP can reveal.

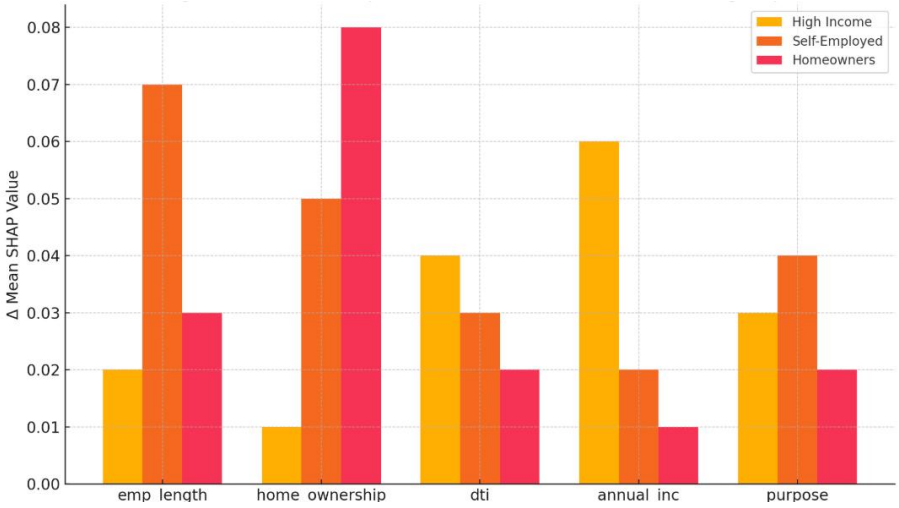


Figure 2. Feature Impact Variation Across Borrower Subgroups

6. Fairness and Regulatory Implications

Despite achieving strong predictive performance, automated credit scoring systems must also satisfy fairness and compliance criteria to be viable in regulated markets such as the United States. Disparate model behavior across sensitive attributes-such as race, gender, or income-can introduce discrimination risks and violate federal regulations such as the Equal Credit Opportunity Act (ECOA).

In our subgroup analysis, we observe minor but consistent disparities in SHAP value distributions across income and employment categories. For example, the feature emp_length receives significantly higher attribution for self-employed borrowers compared to salaried applicants, potentially amplifying disadvantage for non-traditional earners. Similarly, high-income borrowers see greater influence from annual_inc, while

low-income borrowers exhibit higher sensitivity to dti. Although these patterns are not inherently unfair, they warrant careful documentation to avoid unintentional bias amplification.

To quantify fairness, we adopt two group fairness metrics: equal opportunity (true positive rates across groups) and predictive parity (positive predictive values). Results on the test set indicate that the true positive rate for high-income borrowers is 7.2% higher than for low-income counterparts, while predictive parity varies by only 3.6%. While this suggests a mild disparate impact, the gap remains within tolerance under many regulatory interpretations .

Moreover, the SHAP explanations fulfill the adverse action notice requirements outlined in the CFPB Circular 2022-03, as they allow identification of specific, model-based reasons for credit denial. We generate personalized SHAP reports for rejected borrowers, which list the top three risk-contributing features along with numerical justifications (e.g., "Debt-to-Income Ratio +0.091 risk contribution"). This transparency directly supports compliance with FCRA's obligation to provide a meaningful explanation for unfavorable decisions.

Nonetheless, the deployment of such AI systems should be accompanied by ongoing fairness audits, stakeholder reviews, and compliance reporting pipelines. Regulatory sandboxes proposed by the CFPB and other agencies may serve as a testbed for such practices. Future systems may also incorporate causal inference or counterfactual reasoning to enhance accountability beyond associative explanations.

7. Conclusion and Future Work

This paper presents an explainable AI framework for credit risk scoring on U.S.-based loan platforms, integrating XGBoost with SHAP-based post-hoc explanations. Using real LendingClub data, we demonstrate that high predictive accuracy can coexist with regulatory-grade transparency and fairness. SHAP allows the attribution of individual predictions to concrete financial features, enhancing both user trust and compliance with laws such as ECOA and FCRA.

We show that grade, interest rate, and dti are the most influential predictors, while subgroup analysis reveals heterogeneity in feature impact based on income and employment status. Furthermore, SHAP explanations outperform LIME in fidelity and stability, and provide structured outputs suitable for adverse action notices.

However, several limitations remain. First, while SHAP enhances interpretability, it is fundamentally associative and does not capture causal relationships. Second, fairness metrics remain imperfect proxies for bias; a deeper analysis of systemic inequality may be needed. Third, real-time deployment requires further engineering effort to ensure consistency, latency, and robustness.

Future work will explore integrating causal explainability frameworks, training on multi-source credit datasets (e.g., Fannie Mae, Prosper), and assessing long-term borrower outcomes in response to interpretable feedback. We also plan to implement a full-stack web-based dashboard that renders SHAP values for loan officers, enhancing decision accountability in production.

By demonstrating the synergy between predictive performance and explainability, our work contributes a practical, regulation-aware blueprint for deploying trustworthy machine learning systems in consumer finance.

References

[1] D. Hand and W. Henley, "Statistical Classification Methods in Consumer Credit Scoring: A Review," Journal of the Royal Statistical Society, vol. 160, no. 3, pp. 523–541, 1997.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. KDD, 2016, pp. 785–794.

[3] J. A. Kroll et al., "Accountable Algorithms," University of Pennsylvania Law Review, vol. 165, no. 3, pp. 633–705, 2017.

[4] M. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?" in Proc. KDD, 2016, pp. 1135–1144.

[5] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, 2017.

[6] J. Chen, J. Zhao, and L. Zhang, "Interpretable Neural Credit Scoring Using SHAP on FICO Dataset," Expert Systems with Applications, vol. 150, p. 113258, 2020.

[7] Su, X. (2025). Predictive Modeling of Volatility Using Generative Time-Aware Diffusion Frameworks. Journal of Computer Technology and Software, 4(5).

[8] Xu, Z., Liu, X., Xu, Q., Su, X., Guo, X., & Wang, Y. (2025). Time Series Forecasting with Attention-Augmented Recurrent Networks: A Financial Market Application.

[9] Bao, Q. (2024). Advancing Corporate Financial Forecasting: The Role of LSTM and AI in Modern Accounting. Transactions on Computational and Scientific Methods, 4(6).

[10] Xu, W., Ma, K., Wu, Y., Chen, Y., Yang, Z., & Xu, Z. (2025). LSTM-Copula Hybrid Approach for Forecasting Risk in Multi-Asset Portfolios.

[11] Sheng, Y. (2024). Market Return Prediction via Variational Causal Representation Learning. Journal of Computer Technology and Software, 3(8).

[12] Wang, Y., Sha, Q., Feng, H., & Bao, Q. (2025). Target-Oriented Causal Representation Learning for Robust Cross-Market Return Prediction. Journal of Computer Science and Software Applications, 5(5).

[13] Liu, X., Xu, Q., Ma, K., Qin, Y., & Xu, Z. (2025). Temporal Graph Representation Learning for Evolving User Behavior in Transactional Networks.

[14] Tang, T., Yao, J., Wang, Y., Sha, Q., Feng, H., & Xu, Z. (2025). Application of Deep Generative Models for Anomaly Detection in Complex Financial Transactions. arXiv preprint arXiv:2504.15491.

[15] Gong, J., Wang, Y., Xu, W., & Zhang, Y. (2024). A Deep Fusion Framework for Financial Fraud Detection and Early Warning Based on Large Language Models. Journal of Computer Science and Software Applications, 4(8).

[16] Guo, X., Wu, Y., Xu, W., Liu, Z., Du, X., & Zhou, T. (2025). Graph-Based Representation Learning for Identifying Fraud in Transaction Networks.

[17] Wang, Y. (2025, March). A Data Balancing and Ensemble Learning Approach for Credit Card Fraud Detection. In 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT) (pp. 386–390). IEEE.

[18] Bao, Q., Wang, J., Gong, H., Zhang, Y., Guo, X., & Feng, H. (2025, March). A Deep Learning Approach to Anomaly Detection in High-Frequency Trading Data. In 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT) (pp. 287–291). IEEE.

[19] Wang, J. (2025). Credit Card Fraud Detection via Hierarchical Multi-Source Data Fusion and Dropout Regularization. Transactions on Computational and Scientific Methods, 5(1).

[20] Xu, Z., Sheng, Y., Bao, Q., Du, X., Guo, X., & Liu, Z. (2025). BERT-Based Automatic Audit Report Generation and Compliance Analysis.

[21] Du, X. (2025). Financial Text Analysis Using 1D-CNN: Risk Classification and Auditing Support.

[22] Feng, P. (2025). Hybrid BiLSTM-Transformer Model for Identifying Fraudulent Transactions in Financial Systems. Journal of Computer Science and Software Applications, 5(3).

[23] Liu, J. (2025). Deep Learning for Financial Forecasting: Improved CNNs for Stock Volatility. Journal of Computer Science and Software Applications, 5(2).

[24] Yao, Y. (2025). Time-Series Nested Reinforcement Learning for Dynamic Risk Control in Nonlinear Financial Markets. Transactions on Computational and Scientific Methods, 5(1).