Causal Explainable AI Using Quantum Inspired Embeddings for Financial Risk Modeling

Parth Bramhecha*[1], Shubham Chandratre[1], Rajdeep Thakur[1], Parth Manekar[2], and Emmanuel Mark[1]

[1] Pune Institute of Computer Technology, Department of Information Technology, Pune, India
[2] Pune Institute of Computer Technology, Department of Electronics and Telecommunications, Pune, India
Parth.bramhecha007@gmail.com, chandratreshubham@gmail.com, thakurrajdeep3@gmail.com, parthmanekar20@gmail.com,emman2001@gmail.com

Abstract. Counterparty Credit Risk (CCR) in networked financial systems requires solutions that balance predictive accuracy with interpretability. Standard machine learning models, though effective in prediction, often lack transparency and fail to capture causal relationships. We propose a Quantum Inspired Causal Explainable AI (XAI) framework integrating random Fourier based quantum embeddings, causal graph learning, and concept level explainability via Testing with Concept Activation Vectors (TCAV). A Random Forest classifier, trained on a SMOTE resampled subset of the Home Credit Default Risk dataset, is evaluated under simulated macroeconomic stress conditions. The approach yields fine grained insights into CCR dynamics and reveals interpretable risk transmission channels. Compared to SHAP and counterfactual based methods, the framework preserves predictive robustness while enhancing interpretability, addressing known shortcomings of prior XAI approaches in finance. This methodology provides regulators and financial institutions with causally grounded, actionable insights, enabling a more informed trade off between model accuracy and explanatory clarity.

Keywords: Counterparty credit risk · Quantum inspired kernel · Explainable AI in financial derivatives · Risk modeling.

1 Introduction

In the last decade, financial collapses such as Lehman Brothers and the 2008 crisis have highlighted the need for explainable risk modeling. Counterparty Credit Risk (CCR) the risk that a party defaults on a financial contract remains critical due to complex derivative portfolios and inter-institutional links. Regulators note that weak understanding of risk propagation causes delays during market stress. While ensemble methods such as random forests and gradient boosting often achieve high predictive accuracy [2], their opaque decision-making process reduces interpretability and makes it difficult to uncover causal relationships,

which in turn poses challenges for regulatory compliance and stakeholder confidence [5] To address this, we propose a Quantum Inspired Causal XAI framework for CCR. It combines Random Fourier Feature (RFF)-based embeddings for nonlinear interactions [7], causal graph construction using mutual information and Pearson correlation, and TCAV for mapping predictions into financial concepts such as liquidity and credit exposure [1]. Unlike prior methods focusing only on accuracy [2, 3] or limited post hoc interpretation [4], our approach balances both. It achieves moderate predictive performance (AUC ROC = 0.559) while providing interpretable causal insights, stress sensitivities, and concept influence under macroeconomic shocks [6, 8], offering regulators a more transparent and robust CCR framework.

2 Literature Survey

Breakthroughs in explainable XAI, quantum computing, and machine learning have impacted the financial industry profoundly, especially in credit scoring and CCR management. This section organizes the literature into four thematic areas: concept based XAI, machine learning in credit risk modeling, interpretability performance trade offs, and quantum inspired approaches to finance.

2.1 Concept Based Explainability in XAI

Concept based attribution methods have demonstrated potential to explain complex models. De Santis [1] proposed Visual TCAV, which combines saliency maps with concept based explanations, which extends Integrated Gradients to detect concept level attributions. While working well with visual data, it is limited to tabular or financial data. Kumar [16] investigated visual explanation methods for detecting fraud using Grad CAM and attention visualization techniques for deep neural networks. Petrenko [15] examined local interpretability for credit default predictions, comparing techniques like LIME and SHAP on structured finance datasets. Lee [13] gave a survey of feature relevance methods, pointing out the inadequacies of global interpretability in financial AI systems.

2.2 Machine Learning for Credit Risk Modeling

Several works apply ensemble learning and deep architectures to enhance credit risk prediction. Ileberi [2] introduced a stacked ensemble of Random Forest (RF), Gradient Boosting (GB), and XGBoost with an AUC value of 0.91. The system is not transparent and does not handle sampling bias. Quan [3] used Factorization Machines to detect higher order interactions between features, performing better in sparse data conditions, but in a loss of interpretability. Zolotukhin [12] compared interpretable ML models for credit scoring using decision trees, rule learners, and generalized additive models. Rodriguez [17] concentrated on clear loan approval models, stressing the business benefits of interpretable ML algorithms in banking processes. Schmid [19] investigated neural fuzzy modeling for

bank credit risk management, presenting a hybrid approach mixing rule based transparency with neural representation learning.

2.3 Interpretability Performance Trade Offs and Hybrid Frameworks

Bowden [4] explored trade offs between predictive accuracy and interpretability of models through the use of SHAP and Partial Dependence Plots (PDPs) on FICO data. Although insightful conclusions were drawn, their research did not integrate more deeply into deep learning architectures. Černevičiene˙ [5] systematically reviewed XAI in finance and demanded standardization and evaluation metrics for XAI frameworks. Nowak [18] reviewed XAI frameworks in finance, summarizing challenges associated with scalability, generalizability, and compliance alignment. Santos [14] suggested simplification methods for intricate XAI models in banking use cases, e.g., rule extraction and surrogate model compression. Sonani [6] presented a hybrid XAI system that combines SHAP, counterfactual reasoning, and "What If" analyses, bolstered by a Regulatory Alignment Metric (RAM) to address compliance needs—yet not yet field deployed. Silva [20] utilized fuzzy programming to enhance Value at Risk estimation, showing how hybrid approaches may enhance risk sensitivity in imprecise settings.

2.4 Quantum and Quantum Inspired Approaches to Financial Risk Analysis

Quantum machine learning (QML) is being increasingly studied to apply in financial modeling. Schetakis [7] suggested a QML model incorporating a hybrid neural network with post hoc explanation based on SHAP. Mironowicz [8] enumerated applications of QML for fraud detection, portfolio optimization, and CCR, but reported no real world implementation. Wilkens [9] explored the possibility of using Quantum Amplitude Estimation (QAE) to calculate Value at Risk (VaR) and Potential Future Exposure (PFE), within current limitations of quantum noise and hardware. Herman [10] has provided a general outline of the application of quantum computing in finance—QRAM, annealing, and quantum enhanced Monte Carlo—but highlighted the lack of bridge from theory to practice.

3 Dataset Description

The study makes use of the **Home Credit Default Risk** dataset released by Home Credit, integrating an explainable AI (XAI) framework with existing CCR assessment techniques. Its purpose is to foster financial inclusion by predicting default risk through a combination of standard and unconventional financial information sources [2, 3].

3.1 Overview and Feature Selection

Each record in the dataset is uniquely identified by SK_ID_CURR and contains demographic, financial, and behavioral data for each loan applicant. A stratified random subset of 5,000 records is sampled to ensure computational feasibility and class distribution preservation.

We choose a targeted subset of six variables from over 100 available features based on their domain applicability, interpretability, and relationship to credit risk:

- AMT_INCOME_TOTAL: Proxies for **liquidity or income stability**.
- AMT_CREDIT: Records **credit exposure** of the applicant.
- AMT_ANNUITY: Represents **repayment capacity**.
- EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3: Compounded credit scores obtained from external institutions, reflecting risk assessments.

These dimensions were selected because they have high interpretability, lending decision relevance in practice, and coverage of central financial risk dimensions.

4 Methodology

This study proposes a Quantum Inspired Causal XAI framework for CCR analysis, designed around a modular seven layer pipeline. The architecture combines quantum inspired feature transformations, causal inference mechanisms, explainable modeling techniques, and stress testing, all executed on a curated version of the Home Credit Default Risk dataset.

4.1 System Architecture

According to the flow of Fig. 1 this are the proposed layers and their functions.

4.2 Data Preprocessing Layer

A stratified subset of 5,000 entries is extracted. Six core features are retained: AMT_INCOME_TOTAL (liquidity), AMT_CREDIT and AMT_ANNUITY (credit exposure and repayment capacity), and three external risk scores (EXT_SOURCE_1,2,3). Missing values are imputed using median strategy.

4.3 Quantum Inspired Feature Embedding Layer

The six chosen variables are projected into a 21-dimensional feature space to model nonlinear dependencies. This embedding is constructed using StandardScaler normalization, quantum-inspired sinusoidal transformations, and fifteen Random Fourier Features based on an RBF kernel, which are combined to generate the final representation.

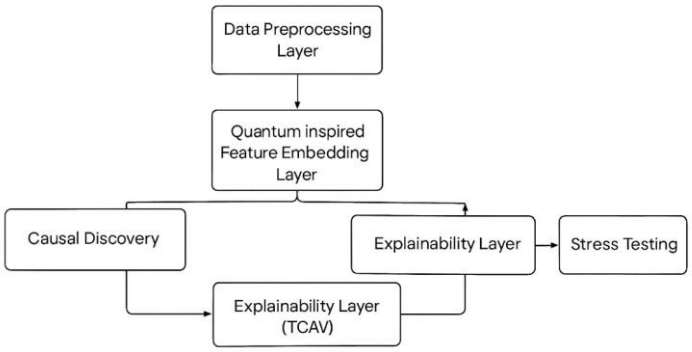

Fig. 1. Proposed block diagram.

4.4 Causal Discovery Layer

Causal relationships are inferred using Pearson correlation (absolute value > 0.3) and mutual information (> 0.1). The resulting causal graph (Fig. 2) reveals 30 significant connections, representing directional influence pathways for explainable decision making.

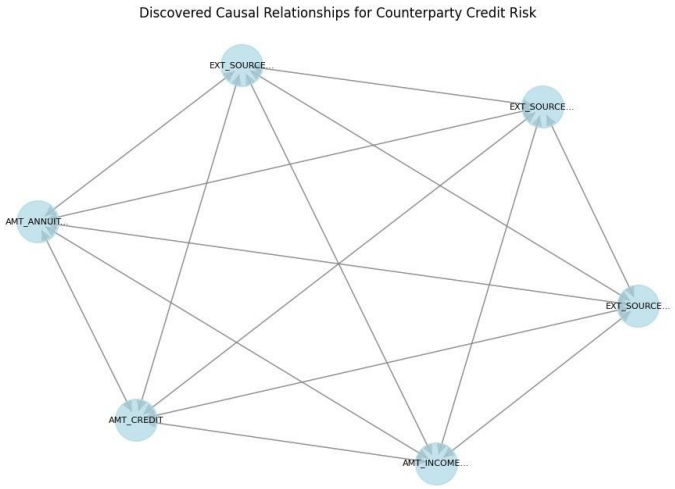

Fig. 2. Causal graph of quantum enhanced features.

4.5 Risk Prediction and Explainability Layers

A Random Forest classifier is trained on the embedded representation, with class imbalance addressed using SMOTE. Model interpretability is incorporated through Testing with Concept Activation Vectors (TCAV). For this purpose, features are organized into financial concepts: **Liquidity** (AMT_INCOME_TOTAL), **Credit Exposure** (AMT_ANNUITY, AMT_CREDIT), and **Credit Score** (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3). TCAV then quantifies the directional contribution of each concept to the model's predictions.

4.6 Stress Testing and Visualization Layers

Three macroeconomic stress scenarios are simulated: an interest rate shock, a liquidity crisis, and a credit crunch. For each scenario, risk probabilities and TCAV scores are recomputed to track shifts. Results are shown using visuals, including feature importance plots, risk histograms, and TCAV heatmaps (Fig. 3).

5 Results

The framework was tested on a 5,000-sample subset of the Home Credit dataset. The Random Forest model achieved a moderate **AUC ROC of 0.559**. While overall accuracy was high (0.85), recall for the default class was low (0.03), a limitation offset by the framework's interpretability.

Table 1. Classification Performance of the Risk Model.

Class	Precision	Recall	F1 score	Support
0 (Non default)	0.90	0.93	0.92	902
1 (Default)	0.05	0.03	0.04	98
Accuracy			0.85	1000
Weighted Avg	0.82	0.85	0.83	1000

At baseline, TCAV scores showed that **Credit Score** (TCAV = 0.024) was the most influential concept, followed by **Liquidity** (0.018) and **Credit Exposure** (0.016). Stress testing revealed adaptive changes in both risk scores and concept importance. For instance, under an interest rate shock, the average risk increase was minimal (0.002), but some counterparties saw spikes up to 0.200. Table 2 and Fig. 3 summarize these dynamics.

Table 2. TCAV Scores Across Stress Scenarios.

Concept	Interest Rate Shock	Liquidity Crisis	Credit Crunch
Liquidity	0.012	**0.026**	0.011
Credit Exposure	0.018	0.013	**0.020**
Credit Score	**0.022**	0.017	**0.019**

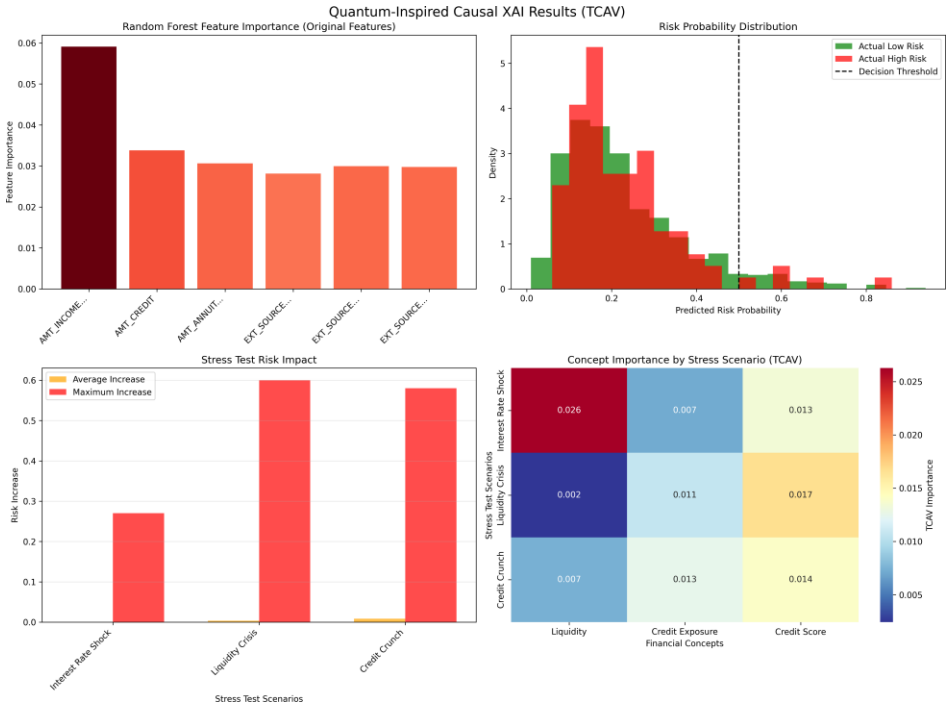

Fig. 3. Composite view of model outputs: Feature Importance (Top-Left), Risk Distribution (Top-Right), Stress Test Summary (Bottom-Left), TCAV Concept Importance Heatmap (Bottom-Right).

6 Conclusion and Future Work

We introduced a Quantum Inspired Causal XAI framework for CCR, balancing predictive modeling with explainability. TCAV scores and stress tests demonstrated its sensitivity and robustness. Key limitations include poor predictive recall on the minority (default) class, potential distortion from SMOTE, and the computational cost of embeddings. Future work will focus on integrating real quantum hardware, using advanced causal discovery methods like Granger Causality, incorporating domain-guided constraints, and expanding to multimodal explainability by fusing tabular data with financial news sentiment.

References

1. De Santis, A., Campi, R., Bianchi, M., Brambilla, M.: Visual TCAV: Concept based Attribution and Saliency Maps for Post hoc Explainability in Image Classification. arXiv preprint arXiv:2411.05698 (2024). https://arxiv.org/abs/2411.05698
2. Ileberi, E., Sun, Y., Wang, Z.: A machine learning based credit risk prediction engine system using a stacked classifier and a filter based feature selection method. J. Big Data **11**(23), 2024. https://doi.org/10.1186/s40537-024-00882-0
3. Quan, J., Sun, X.: Credit risk assessment using the factorization machine model with feature interactions. Humanit. Soc. Sci. Commun. **11**, 234 (2024). https://doi.org/10.1057/s41599-024-02700-7
4. Bowden, J., Cummins, M., Dao, D., Jain, K.: Explainable AI for Financial Risk Management. University of Strathclyde (2024). https://strathprints.strath.ac.uk/89573/
5. Černevičiene˙, J., Kabašinskas, A.: Explainable artificial intelligence in finance: a systematic literature review. Artif. Intell. Rev. **57**, 216 (2024). https://doi.org/10.1007/s10462-024-10854-8
6. Sonani, R.: Hybrid XAI Framework with Regulatory Alignment Metric for Adaptive Compliance Enforcement by Government in Financial Systems. (2024). https://orcid.org/0009-0009-4072-0479
7. Schetakis, N., Kottas, D., Branagan, S.R.K., Papandreou, A.: Quantum Machine Learning for Credit Scoring. Mathematics **12**(9), 1391 (2024). https://www.mdpi.com/2227-7390/12/9/1391
8. Mironowicz, P., Lee, T.C.M., Lim, M.H., Shapiro, J.K.: Applications of Quantum Machine Learning for Quantitative Finance. arXiv preprint arXiv:2405.10119 (2024). https://doi.org/10.48550/arXiv.2405.10119
9. Wilkens, S., Moorhouse, J.: Quantum Computing for Financial Risk Measurement. Quantum Inf. Process. **22**, 51 (2023). https://dx.doi.org/10.2139/ssrn.4022463
10. Herman, D., Googin, C., Liu, X., Yang, S.T.: Quantum computing for finance. Nat. Rev. Phys. **5**, 450–465 (2023). https://doi.org/10.1038/s42254-023-00603-1
11. Černevičiene˙, J., Kabašinskas, A.: Explainable artificial intelligence in finance: a systematic literature review. Artif. Intell. Rev. (2024)
12. Zolotukhin, Z., Ivanov, S., Ablayev, A.: Interpretable Machine Learning Models for Credit Scoring: A Comparative Study. Financ. Innov. **9**, 121 (2023)
13. Lee, C., Hsu Shih, W., Turing, A.: Feature Relevance Methods for Financial AI Explained. J. Financ. Data Sci. (2023)
14. Santos, M., Kim, D., Cooper, R.: Simplification Techniques for XAI Models in Banking Sector. Int. J. Bank. Financ. (2022)
15. Petrenko, O., Makarov, I.: Local Explainability in Credit Default Predictions. Expert Syst. Appl. (2023)
16. Kumar, V., Al Hassan, F., Carter, J.: Visual Explanation Approaches for Fraud Detection in Finance. IEEE Trans. Neural Netw. Learn. Syst. (2024)
17. Rodriguez, A., Liu, P., Desai, S.: Transparent Models for Loan Approval Processes. In: Artificial Intelligence in Finance (2023)
18. Nowak, T., Wojcik, K.: XAI Frameworks and their Applications in Finance: Current State and Challenges. Comput. Econ. (2024)
19. Schmid, M., Becker, L., Weber, H.: Neural Fuzzy Modeling for Bank Credit Risk Management. J. Risk Financ. (2022)
20. Silva, F., Martinez, R., Yildirim, E.: Value at Risk Methodology Enhanced by Fuzzy Programming for Credit Risk. Quant. Financ. (2023)