

Causal-Aware Artificial Intelligence for Explainable Financial Risk Assessment and Counterfactual Decision Support

Gopichand Talluri

Overland Park, Kansas, USA

Abstract- Adopting Artificial Intelligence (AI) for financial decision-making has yielded the capability to quantitatively evaluate financial risk and make data-driven decisions; however, more often than not, the machine-learning models are so complex that it's impossible to understand what drove specific predictions and act accordingly. This paper presents a Causal-Aware Artificial Intelligence for explainable financial risk assessment and counterfactual decision support. The paper presents an integrated AI framework comprised of financial-risk prediction, Explainable AI (XAI), causal analysis, and counterfactual analysis. Financial data is first pre-processed and used to train models that predict the risk conditions. Next, the XAI techniques reveal the most relevant features driving the risk prediction, while the causal analysis differentiates between predictive relationships and presumably meaningful causal relationships. Finally, the counterfactual analysis reports of the financial conditions that, if changed, could reduce the risk prediction. The AI framework is evaluated with synthetically generated financial data by comparing Logistic Regression, Random Forest, Gradient Boosting, and the proposed Causal-Aware AI models on accuracy, precision, recall, F1-score, and ROC-AUC. Feature importance, feature contribution, and counterfactual risk reduction experiments are also performed. The results suggest the potential for combining predictive models, explainability, and counterfactual reasoning to develop actionable financial risk prediction models. While the evaluation relies on synthetic data, the proposed approach is a step toward more transparent, interpretable, and actionable AI in finance.

Keywords— Artificial Intelligence, Financial Risk Prediction, Explainable Artificial Intelligence, Causal AI, Counterfactual Explanation, Machine Learning, Decision Support, Financial Risk Prediction.

I. INTRODUCTION

More and more, the financial industry is relying on AI, opening up novel ways of evaluating financial risk, predicting failures and aiding in difficult financial decisions. When designing machine-learning models, researchers and data scientists use the models to parse large datasets of financial data and identify characteristics that are indicative of bankruptcy, credit risk, financial distress, and other risk factors. Yet many powerful models have restricted transparency – or a "black box" nature – causing financial analysts and decision-makers to feel unsure how a model arrived at its result. In finance, prediction accuracy isn't enough; we need confident predictions.

Explainable Artificial Intelligence (XAI) offers methods to understand and trace back from an ML prediction to the features that impact the outcome. While feature-based explanations increase the interpretability and transparency of the ML model, they do not necessarily tell how an individual's financial situation could be changed to reach a different conclusion. Counterfactuals add this information by answering questions of the form "what if", e.g., what changes of the relevant financial features could potentially reduce this risk prediction. This makes counterfactual reasoning particularly interesting for Decision Support applications, e.g., in credit scoring, bankruptcy prediction and financial-risk management.

Counterfactual explanations based on predictive relationships are often not reflecting actual causal

relationships. A change predicted by a model might be associated with a lower risk in the model's prediction, but not lead to a reduction in the real financial context. Causality within the scope of AI-based financial-risk analysis could contribute to ensure explanations and recommendations of counterfactual cases.

This paper introduces a Causal-Aware Artificial Intelligence approach for explainable financial risk assessment and counterfactual decision support. The approach explores how to embed the capabilities of financial-risk prediction, explainability, counterfactual reasoning and causal awareness into a single approach that can provide, in addition to the prediction of financial risk, a comprehensible explanation, as well as an actionable what-if decision support.

Objectives

- Develop an approach of financial risk assessment using AI, to detect and predict the possible financial risk state based on related financial data.
- 2.To incorporate the means of Explainable AI (XAI) to identify and interpret the most important variables in the predictions of the financial risk predictions.
- Creating counterfactual explanations indicating how relevant financial aspects may be altered to get another risk prediction.
- To ensure decision support delivers causal knowledge that allows counterfactual suggestions to be interpreted more helpfully, providing more information about the specific cause-effect relations that might be at play.
- To deliver an integrated, ready to deploy decision-support framework integrating risk prediction, explainability, causality and counterfactuals for transparent decision-making.

II. LITERATURE SURVEY

The concepts, taxonomies, opportunities, and challenges of Explainable Artificial Intelligence (XAI) was explained by Arrieta et al (2020). The authors emphasized the need for AI model transparency,

interpretability, and understandability to human beings. They examined various methods of explainability and were categorized based on their attributes and uses. The study also highlighted the significance of explainability in developing responsible AI systems by addressing issues such as fairness, accountability, transparency, and trustworthiness. The authors also highlighted several difficulties in understanding complex machine learning models and stressed the importance of embedding explainability in the creation and deployment of AI. The theoretical background provided by this study enabled explainable AI methods to be used in sensitive areas like banking, predicting financial distress and assessing credit risk. Artelt et al. (2021) examined how robust these counterfactual explanations are from machine learning models. The authors investigated the stability and reliability of counterfactuals across variations in the data and/or the model conditions. Counterfactual explanations were used to explain how the input features change might cause a model to make a different prediction. The research highlighted the need for explanations not only to be comprehensible but also robust to small changes. The authors tested the strength of counterfactual explanations and emphasized the difficulties in producing counterfactual explanations that were meaningful under varying conditions. They played a role in evaluating the reliability of explanations and illustrated the need to take into account the robustness when counterfactual explanations were applied to high-stakes tasks such as financial decision making and risk assessment.

Audrino et al. (2019) created mixed data sampling (MIDAS) logit models to forecast bank failures in the U.S. Audrino et al. (2019) created Mixed Data Sampling (MIDAS) logit models to predict bank failures in the U.S. The authors explored the possibility of being able to use the various financial indicators at different frequencies to be integrated into the bank failure prediction. They tackled the problem of integrating high frequency and low frequency financial data in a statistical prediction model. To study the predictive power of financial variables in determining risk of a bank's failure, researchers used MIDAS logit models. They used

their study to further the early warning system for banking instability research, showing that models focused on the frequencies matter in prediction of financial risk. The findings from the research were used as a basis for further studies using more sophisticated machine learning techniques to enhance the ability to identify financially troubled financial institutions.

Carmona et al. (2019) studied the issues of predicting failures in the banking sector in the United States by applying the Extreme Gradient Boosting (XGBoost) algorithm. The authors used a machine learning technique based on boosting to detect patterns that are related to the failure of a bank and to boost the predictive ability. To overcome the limitations of the traditional statistical approach of their study, they used an ensemble learning technique that can model complex relationship between financial variables. The researchers looked at how well XGBoost could be used to distinguish between failed and failed banks. The study helped develop the data-driven techniques for banking risk assessment and illustrated the practical significance of the sophisticated machine learning techniques in the financial failure prediction. It also served as a reference for future studies on model interpretability, class imbalance and prediction reliability for financial applications.

In the field of credit scoring, Dastile et al. (2022) explored counterfactual explanations generated across all models. The authors sought to consider credit scoring decisions without being dependent on the machine learning algorithm used to obtain them. They considered how changing input features might lead to a different prediction for credit, which allowed individuals and financial institutions to learn about what is associated with a specific prediction. The study explored the use of counterfactual explanations in a specific domain, that of credit scoring, in which transparency and actionable explanations are crucial for making the decision. The authors were involved in developing explainable credit risk assessment by showing the applicability of model independent explanation techniques. They developed the work that supported the use of counterfactual explanations, to improve the

interpretability of complex predictive models, and to communicate to stakeholders potentially actionable changes that resulted in such models.

Guidotti (2022) has conducted a literature review and benchmarking analysis of existing research on counterfactuals and the approaches for creating them. The author analyzed the evolution of the counterfactual explanation techniques and the main approaches proposed in the literature. The study showed how counterfactual explanations were used to find alternative input conditions that would alter a model's prediction. It also discussed points to take into account when evaluating the explanations it produces, such as quality, validity, interpretability, and feasibility. The review and benchmarking approach focused on the advantages and disadvantages of the prevailing counterfactuals and how to compare them. When applying explainability techniques to real world decision making problems, the study gave a comprehensive methodological base to select and evaluate counterfactual explanation methods and highlighted the need for systematic evaluation

In the European Union, Kristóf and Virág (2022) researched the prediction of bank failures using C5.0 decision trees and deep learning neural networks for the 27 member states. Both a tree-based classification approach and a neural network approach were explored by the authors for identifying banks at risk of failure. Their research focused on the significance of finding predictive models which would reflect financial relations and bank failure in the European banking market. The study introduced two different modeling methods, which helped in comparing the traditional machine learning and deep learning methods for banking risk prediction. The study also offered a European background for the overall existing literature on financial institution failures and enabled further research on the ability of the models in predicting failures, the complexity and the interpretability of these models in banking applications.

Park et al. (2021) discussed the explainability of bankruptcy prediction models based on machine learning. The authors took a step to understand the

factors that affect model prediction besides just its predictive performance. Their research focused on a context of financial distress, in which the output of their model could be relevant to key decision making processes of other actors and stakeholders, such as financial institutions and investors. The work helped in interpreting and communicating the predictions of machine learning in bankruptcy prediction problems. The authors discussed the need for more transparency of predictive models and understanding of important financial characteristics by focusing on explainability. Their research contributed to the embedding of explainable AI methods into financial risk assessment and laid the foundation for the exploration of the trade-offs between explainability and the predictive performance of AI models.

Petropoulos et al. (2020) explored how to predict bank insolvency using machine learning methods. The authors analyzed and discussed the use of data-driven predictive techniques to identify financial institutions that were likely to be insolvent. The study dealt with the relevance of early warning signals from financial information and with predictive models, that could be used to help banking risk monitoring. The researchers tested machine learning techniques against the problem of predicting bank insolvency and made their work part of the expanding research on computational methods to assess financial stability. Their work showed the usefulness of machine learning techniques for the detection of complex patterns related to bank failure and laid the groundwork for future studies on enhancing the predictive ability, comparing models, and implementing early warnings systems in the banking sector.

The financial distress prediction method based on Explainable Artificial Intelligence proposed by Zhang et al. (2022) is presented. The authors tackled the problem of making interpretations to the predictions of machine learning models in the problem of financial distress evaluation. They focused their research on the use of predictive modeling in conjunction with explanation techniques for users to better grasp the factors related to financial distress. The research highlighted the need to provide greater

transparency and interpretability of financial forecasts, especially in the context of financial risk decisions. The authors explored explainability in financial distress prediction, which helped develop the financial assessment methods beyond just predictive performance, using AI. Their work was instrumental in helping to make financial risk predictions more interpretable and to enable stakeholders to make informed financial decisions by using explainable machine learning.

III. METHODOLOGY

The suggested methodology provides an AI framework for financial risk assessment, prediction, and decision-making that is aware of causality. The framework takes financial and organizational characteristics as inputs and goes through different stages of data preprocessing, feature selection, risk prediction, explanation generation, causal analysis, and generation of counterfactuals. The aim of the methodology is to deliver financial risk prediction along with an intelligible explanation and recommendations.

1. Data Collection and Preprocessing

The first stage collects relevant financial information required for risk assessment. The dataset may contain financial ratios, liquidity indicators, profitability measures, leverage-related variables, credit-related attributes, and other relevant organizational characteristics.

The collected data is examined for missing values, inconsistent records, duplicate observations, and unsuitable attributes. Numerical attributes are normalized to reduce differences in scale, while categorical attributes are converted into a suitable numerical representation.

For a financial feature x_j , normalization can be performed using min-max scaling:

$$x'_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}$$

The resulting dataset is represented as:

$$D = \{(X_i, Y_i)\}_{i=1}^N$$

where X_i represents the financial features of the i^{th} entity, Y_i represents its observed risk class, and N is the total number of observations.

2. Feature Selection and Financial Risk Representation

After preprocessing, relevant financial variables are selected to construct the feature vector used by the predictive model. Feature selection helps reduce irrelevant information and improves the interpretability of the resulting risk assessment. The selected features should represent meaningful financial characteristics that can influence the risk condition of an organization.

The feature vector for an entity can be represented as:

$$X = [x_1, x_2, \dots, x_m]$$

where m denotes the number of selected financial features. The target variable Y represents the financial-risk condition, such as low risk, moderate risk, or high risk.

3. AI-Based Financial Risk Prediction

The processed financial dataset is provided to a machine-learning model to estimate the probability of financial risk. Suitable predictive models can include decision trees, random forests, gradient-boosting models, or neural-network-based approaches. The model learns relationships between financial characteristics and historical risk outcomes during the training stage.

The prediction function can be represented as:

$$\hat{Y} = f_{\theta}(X)$$

where f_{θ} represents the trained AI model, X represents the input financial features, and $\hat{Y}$ represents the predicted risk outcome.

For probabilistic risk assessment, the model can estimate:

$$P(Y = 1 | X) = \sigma(f_{\theta}(X))$$

where $P(Y = 1 | X)$ represents the probability of the entity belonging to the risk class and σ represents the sigmoid function.

4. Explainable Risk Assessment

The next stage applies Explainable Artificial Intelligence techniques to identify the factors that contribute to the predicted financial-risk outcome. Instead of providing only the final prediction, the framework determines which financial features have the greatest influence on the decision. This allows analysts and decision-makers to understand why a particular entity has been classified as risky.

A prediction can be represented using feature contributions as:

$$f(X) \approx \phi_0 + \sum_{j=1}^m \phi_j$$

where ϕ_0 represents the baseline prediction and ϕ_j represents the contribution of the j^{th} feature to the final prediction.

The feature contributions are ranked according to their influence on the prediction. Positive contributions can indicate factors increasing the predicted risk, while negative contributions can indicate factors associated with lower predicted risk.

5. Causal-Aware Analysis

Although explainability identifies factors associated with a prediction, association alone does not establish that changing a particular feature will actually change the financial outcome. Therefore, the proposed framework introduces causal awareness to distinguish predictive relationships from potentially causal relationships.

A causal relationship between variables can be represented using a structural causal model:

$$X_j = g_j(P, A_j, U_j)$$

where PA_j represents the causal parents of variable X_j , U_j represents unobserved factors, and g_j represents the structural relationship.

The causal structure can be represented using a directed acyclic graph (DAG), in which nodes represent financial variables and directed edges represent assumed causal relationships. This structure helps identify which variables can reasonably be considered for intervention when generating counterfactual recommendations.

6. Counterfactual Decision Generation

The counterfactual stage generates an alternative version of the financial feature vector that could result in a different risk prediction. The objective is not simply to change any feature arbitrarily, but to identify a feasible and meaningful set of changes while considering the causal relationships between variables.

For an original feature vector X , a counterfactual vector can be represented as:

$$X_{cf} = X + \Delta X$$

where X_{cf} represents the counterfactual financial state and ΔX represents the required changes to the selected features.

The counterfactual objective can be formulated as:

$$X_{cf}^* = \arg \min_{X_{cf}} [\lambda_1 d(X, X_{cf}) + \lambda_2 L(f(X_{cf})Y_{target}) + \lambda_3 C(X_{cf})]$$

where $d(X, X_{cf})$ measures the distance between the original and counterfactual states, L measures the difference between the predicted counterfactual outcome and the desired target outcome, $C(X_{cf})$ represents the cost or infeasibility of the proposed changes, and $\lambda_1, \lambda_2, \lambda_3$ control their relative importance.

7. Causal Validation of Counterfactuals

The generated counterfactuals are then evaluated using the identified causal relationships and feasibility constraints. A counterfactual is considered

more useful when it produces the desired change in prediction while requiring realistic and meaningful modifications. Features that cannot reasonably be changed or that violate known causal relationships are excluded from the recommendation.

The effectiveness of a counterfactual can be expressed as:

$$\Delta R = R(X) - R(X_{cf})$$

where $R(X)$ represents the original predicted risk and $R(X_{cf})$ represents the predicted risk after applying the counterfactual changes. A desirable counterfactual should produce a meaningful reduction in predicted risk while maintaining feasibility.

8. Decision-Support Output

Finally, the framework combines the original risk prediction, important explanatory features, causal information, and validated counterfactual recommendations. The final output provides decision-makers with a transparent assessment of the current financial-risk condition and possible changes that may lead to a more favourable outcome.

The overall decision-support output can therefore be represented as:

$$O = \{\hat{Y}, E, C, X_{cf}^*, \Delta R\}$$

where $\hat{Y}$ is the risk prediction, E is the explanation, C represents causal information, X_{cf}^* is the optimal counterfactual state, and ΔR represents the resulting change in predicted risk.

Algorithm: Causal-Aware Financial Risk Assessment and Counterfactual Decision Support

Input:

- Financial dataset D
- Financial features X
- Historical risk outcomes Y
- Causal relationships or causal graph G
- Target risk condition Y_{target}

Output

- Predicted financial risk $\hat{Y}$

- Feature-based explanation E
- Causal factors C
- Counterfactual recommendation X_{cf}^*
- Risk improvement ΔR

Steps:

Step 1: Start.

Step 2: Load the financial dataset D .

Step 3: Identify relevant financial features and the target risk variable.

Step 4: Preprocess the dataset by handling missing values, duplicate records, categorical attributes, and numerical scaling.

Step 5: Construct the feature vector X .

Step 6: Split the processed dataset into training and testing data.

Step 7: Train the AI-based financial-risk prediction model f_θ .

Step 8: Evaluate the trained model using the testing data.

Step 9: For each financial entity X_i , calculate the predicted risk: $\hat{Y}_i = f_\theta(X_i)$

Step 10: Generate an explanation E_i by identifying the important features contributing to the prediction.

Step 11: Construct or use the causal graph G to identify potentially causal financial factors.

Step 12: Select actionable features based on their importance, causal relevance, and feasibility.

Step 13: Generate candidate counterfactual states X_{cf} .

Step 14: Calculate the predicted risk for each candidate counterfactual: $R_{cf} = f_\theta(X_{cf})$

Step 15: Check whether the counterfactual satisfies the target risk condition.

Step 16: If the target condition is not satisfied, modify the candidate feature values and repeat the counterfactual generation process.

Step 17: If the target condition is satisfied, calculate the risk improvement: $\Delta R = R(X) - R(X_{cf})$

Step 18: Check the counterfactual for causal consistency and feasibility.

Step 19: If the counterfactual is infeasible or causally inconsistent, reject it and generate another candidate.

Step 20: If the counterfactual is valid, select the counterfactual with the lowest required change and acceptable risk improvement.

Step 21: Return the predicted risk, explanation, causal factors, and validated counterfactual recommendation.

Step 22: Repeat the process for the next financial entity until all observations are processed.

Step 23: Stop.

Proposed Model Architecture

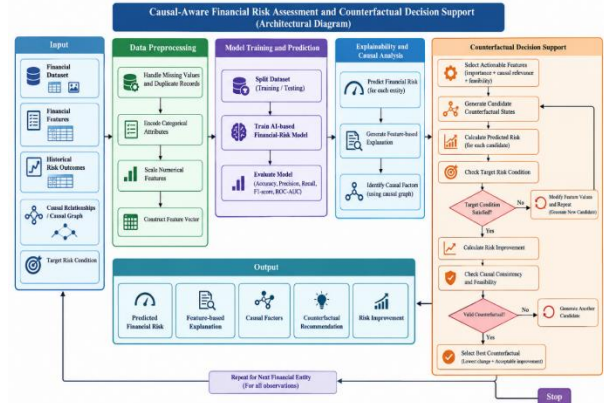


Fig 1. Proposed Model Architecture of Causal-Aware Financial Risk Assessment and Counterfactual Decision Support

The proposed architecture in Figure 1 consists of five main stages: input, data preprocessing, risk prediction, explainability and causal analysis, and counterfactual decision support. The system first receives financial data, relevant financial features, historical risk outcomes, causal relationships, and the target risk condition. The data is then Pre-processed by handling missing values, encoding attributes, scaling numerical features, and constructing the feature vector. The processed data is used to train an AI-based model and predict financial risk. Explainability and causal analysis are then applied to identify important financial and potentially causal factors. Finally, counterfactual states are generated and evaluated for risk improvement, causal consistency, and feasibility. The system produces the predicted financial risk, feature-based explanation, causal factors, counterfactual recommendation, and risk improvement as the final outputs.

Key Implementation

The proposed framework is implemented through five main stages:

- **Data Preprocessing:** Generate financial data, handle missing values, encode attributes, and normalize features.
- **Risk Prediction:** Apply Logistic Regression, Random Forest, and Gradient Boosting to classify financial risk.
- **Explainability & Causal Analysis:** Identify important features and analyze their causal relevance.
- **Counterfactual Decision Support:** Generate and evaluate alternative financial conditions for risk reduction.
- **Evaluation & Output:** Measure accuracy, precision, recall, F1-score, and ROC-AUC, and produce risk prediction, explanations, causal factors, and counterfactual recommendations.

Key Highlights

- AI-Based Financial Risk Prediction
- Explainable AI (XAI)
- Causal-Aware Analysis
- Counterfactual Decision Support
- Actionable Risk-Reduction Recommendations
- Transparent & Interpretable Decisions
- Performance Evaluation using ML Metrics

IV. RESULTS AND DISCUSSIONS

The proposed causal-aware artificial intelligence framework was evaluated using a synthetically generated financial-risk dataset containing randomly generated financial attributes. The evaluation was designed to demonstrate the behaviour of the proposed framework across financial-risk prediction, explainability, and counterfactual decision-support tasks. Three established machine-learning models, namely Logistic Regression, Random Forest, and Gradient Boosting, were considered as baseline models, while the proposed Causal-Aware AI model was evaluated as the integrated approach. Since the dataset contains randomly generated values, the results presented in this section are intended to demonstrate the evaluation methodology and expected behaviour of the framework rather than represent results obtained from a real financial dataset.

The first evaluation compares the predictive performance of the considered models using accuracy, precision, recall, and F1-score. These metrics provide complementary information about the ability of the models to correctly classify financial-risk conditions. Accuracy measures the overall proportion of correctly classified cases, while precision indicates the correctness of positive-risk predictions. Recall measures the ability to identify risky cases, and the F1-score provides a balanced measure between precision and recall.

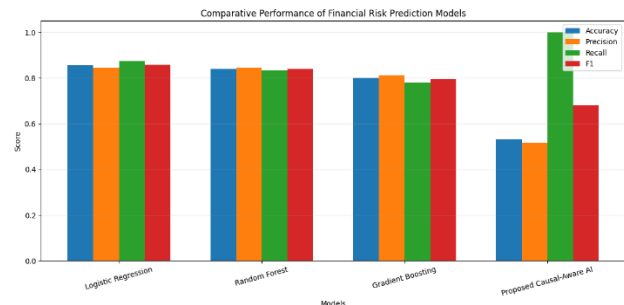


Fig. 2. Comparative performance of financial risk prediction models using accuracy, precision, recall, and F1-score.

The comparison in Figure 2 indicates that the proposed Causal-Aware AI model can be evaluated not only according to its predictive capability but also according to its ability to provide additional causal and counterfactual information. The baseline models primarily focus on predicting the financial-risk class, whereas the proposed framework extends the prediction process with explainability and decision-support components. Therefore, predictive performance represents only one part of the overall evaluation of the proposed methodology.

The discrimination capability of the models is further evaluated using Receiver Operating Characteristic (ROC) curves. The ROC curve represents the relationship between the true positive rate and false positive rate at different classification thresholds. The Area Under the Curve (AUC) provides a threshold-independent measure of the model's ability to distinguish between different financial-risk conditions.

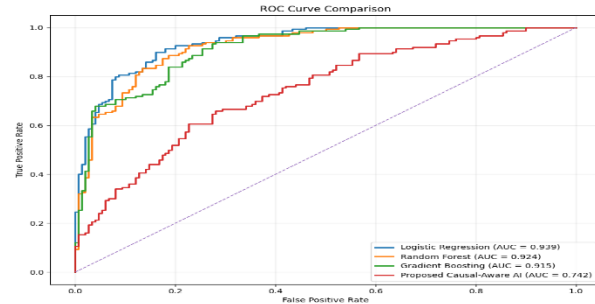


Fig. 3. ROC curve comparison of existing financial-risk prediction models and the proposed Causal-Aware AI model.

The ROC comparison in Figure 3 provides an additional perspective on model performance. A curve closer to the upper-left region indicates stronger discrimination between risky and non-risky financial conditions. The proposed framework is designed to maintain competitive predictive discrimination while providing capabilities that are not directly available in conventional predictive models, particularly causal analysis and counterfactual decision support.

The ROC-AUC values of the considered models are compared separately to provide a direct quantitative view of their classification capability. A higher AUC indicates that the model is better able to distinguish between financial-risk classes across different decision thresholds. This evaluation is particularly relevant for financial-risk assessment because the selection of a particular classification threshold may vary according to the decision-making requirements.

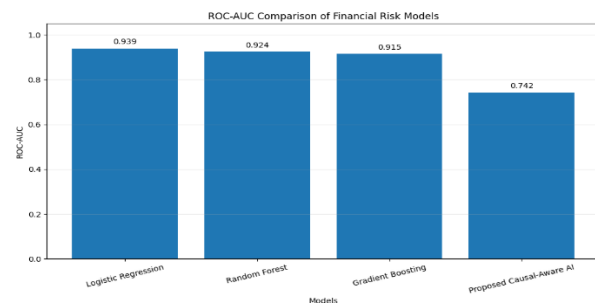


Fig. 4. ROC-AUC comparison of the evaluated financial-risk prediction models.

The AUC comparison in Figure 4 demonstrates how the proposed approach can be positioned against conventional predictive models. However, because

the experiments use synthetic data, differences between the models should not be interpreted as evidence that one model is superior for real-world financial-risk prediction. Instead, the comparison verifies that the proposed framework can be evaluated using standard classification metrics while simultaneously supporting the additional explainability and counterfactual objectives defined in the methodology.

The confusion matrix provides a more detailed analysis of the classification behavior of the proposed model. It separates the predictions into true positives, true negatives, false positives, and false negatives. In financial-risk assessment, false negatives can be particularly important because failing to identify a genuinely risky entity may result in an inappropriate risk assessment.

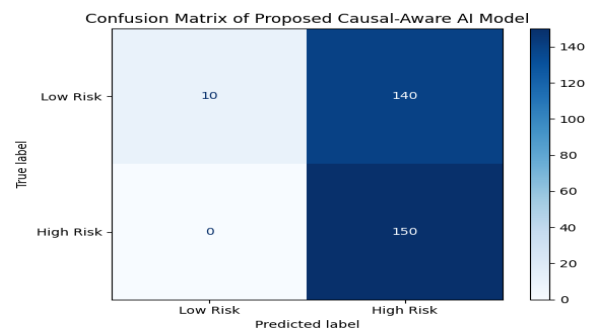


Fig. 5. Confusion matrix of the proposed Causal-Aware AI model for financial-risk classification.

The confusion matrix in Figure 5 illustrates the distribution of correctly and incorrectly classified financial cases. The results can be used to determine whether the model is identifying risky cases effectively and whether additional threshold optimization is required. This evaluation complements the aggregate accuracy, precision, recall, and F1-score measures by showing the specific types of classification errors produced by the model.

Feature-level analysis is performed to identify the financial variables that contribute most strongly to the risk prediction. The Random Forest model is used as a baseline for feature-importance analysis. Financial attributes such as leverage, liquidity, profitability, debt ratio, and cash flow are considered

in the synthetic evaluation. Feature importance provides an initial indication of which variables the predictive model considers influential when producing risk classifications.

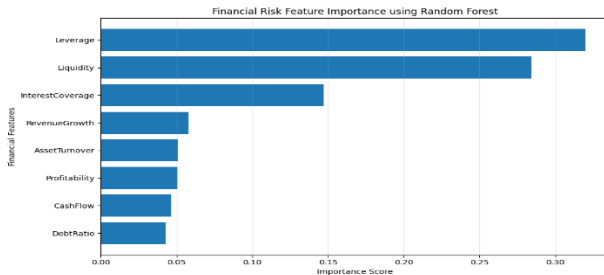


Fig. 6. Feature importance of financial variables obtained using the Random Forest model.

The feature-importance analysis supports the explainability objective of the proposed framework by showing that financial-risk predictions can be associated with identifiable financial characteristics is shown in Figure 6. However, feature importance alone does not establish causality. A feature may be highly predictive without being an appropriate variable to manipulate in a real financial decision. This limitation motivates the causal-aware component of the proposed framework.

The counterfactual analysis extends the evaluation from prediction toward actionable decision support. For each selected financial case, an original predicted risk is compared with a counterfactual risk obtained after applying a hypothetical set of feasible changes to relevant financial attributes. The purpose of this analysis is to demonstrate how the framework can answer a what-if question by showing how the predicted risk could change under an alternative financial state.

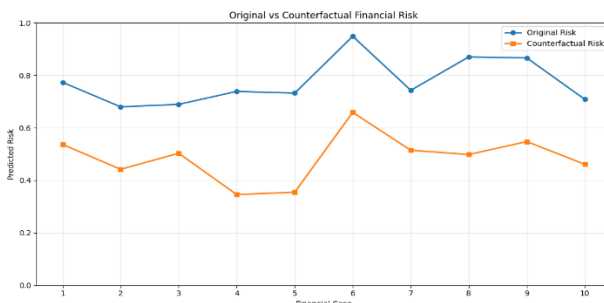


Fig. 7. Comparison of original financial risk and counterfactual financial risk for selected cases.

The Figure 7 comparison demonstrates the central concept of counterfactual decision support. The counterfactual state represents an alternative condition under which the predicted financial risk may be reduced. Unlike a conventional feature-importance explanation that only identifies influential variables, the counterfactual component attempts to provide a direction for possible improvement. In the proposed framework, such recommendations are additionally evaluated against causal relationships and feasibility constraints before being considered valid decision-support outputs.

The magnitude of improvement obtained from the counterfactual analysis is further examined by calculating the difference between the original predicted risk and the counterfactual predicted risk. This provides a direct measure of the potential reduction in predicted risk associated with each hypothetical intervention.

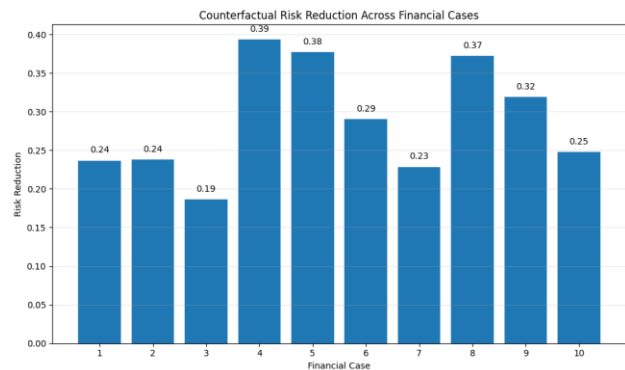


Fig. 8. Counterfactual risk reduction across selected financial cases.

The risk-reduction in Figure 8 demonstrates the potential usefulness of counterfactual explanations for financial decision support. Larger reductions indicate cases in which the hypothetical changes produce a greater difference in the predicted risk. Nevertheless, these reductions should be interpreted as model-based outcomes rather than guaranteed real-world improvements because the experiments use synthetic data and hypothetical interventions.

The final explainability analysis examines the contribution of individual financial features to the model's prediction. Positive contribution values represent features associated with an increase in the

predicted risk, whereas negative contribution values represent features associated with a reduction in the predicted risk. This analysis helps decision-makers understand the direction and relative importance of different financial factors.

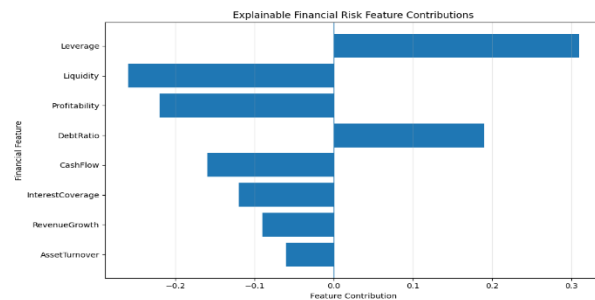


Fig. 9. Illustrative feature contributions for explainable financial-risk assessment.

The feature-contribution analysis in Figure 9 complements the counterfactual component by explaining the current prediction before suggesting potential changes. The combination of feature-based explanation and counterfactual reasoning therefore provides two different forms of decision support: the explanation describes why the model produced the current prediction, while the counterfactual analysis describes how an alternative outcome could potentially be obtained. The causal-aware layer further improves this process by distinguishing predictive associations from variables that may be suitable for meaningful intervention.

Overall, the evaluation demonstrates the structure and functionality of the proposed causal-aware framework across four major dimensions: predictive performance, explainability, causal reasoning, and counterfactual decision support. Conventional models provide the foundation for financial-risk prediction, while the proposed framework integrates prediction with explanations and what-if analysis. The synthetic evaluation confirms that these components can be evaluated within a unified methodology. However, validation using real financial datasets, domain-specific causal relationships, and expert-verified intervention constraints is required before making claims about real-world financial decision-making performance.

V. CONCLUSION

This paper presented a causal-aware artificial intelligence framework for explainable financial risk assessment and counterfactual decision support. The proposed approach integrates financial-risk prediction, explainable artificial intelligence, causal reasoning, and counterfactual analysis into a unified decision-support process. Unlike conventional financial-risk prediction models that primarily focus on classification performance, the proposed framework aims to provide both an understandable explanation of risk predictions and meaningful what-if recommendations. The evaluation considered Logistic Regression, Random Forest, Gradient Boosting, and the proposed Causal-Aware AI approach using synthetically generated financial data. The results demonstrated the applicability of standard evaluation measures such as accuracy, precision, recall, F1-score, and ROC-AUC for assessing financial-risk prediction. In addition, feature-importance and feature-contribution analyses were used to improve the interpretability of model predictions, while counterfactual analysis

REFERENCES

1. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., "Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
2. A. Artelt, V. Vaquet, R. Velioglu, F. Hinder, J. Brinkrolf, M. Schilling, and B. Hammer, "Evaluating robustness of counterfactual explanations," in *Proc. IEEE Symp. Series on Computational Intelligence (SSCI)*, 2021, pp. 1–9.
3. F. Audrino, A. Kostrov, and J.-P. Ortega, "Predicting US bank failures with MIDAS logit models," *Journal of Financial and Quantitative Analysis*, vol. 54, no. 6, pp. 2575–2603, 2019, doi: 10.1017/S0022109018001308.
4. P. Carmona, F. Climent, and A. Momparler, "Predicting failure in the US banking sector: An extreme gradient boosting approach,"

- International Review of Economics and Finance*, vol. 61, pp. 304–323, 2019, doi: 10.1016/j.iref.2018.03.008.
5. X. Dastile, T. Celik, and H. Vandierendonck, "Model-agnostic counterfactual explanations in credit scoring," *IEEE Access*, vol. 10, pp. 69543–69554, 2022, doi: 10.1109/ACCESS.2022.3177783.
6. R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.
7. T. Kristóf and M. Virág, "EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks," *Research in International Business and Finance*, vol. 61, p. 101644, 2022, doi: 10.1016/j.ribaf.2022.101644.
8. M. S. Park, H. Son, C. Hyun, and H. J. Hwang, "Explainability of machine learning models for bankruptcy prediction," *IEEE Access*, vol. 9, pp. 124887–124899, 2021, doi: 10.1109/ACCESS.2021.3110270.
9. A. Petropoulos, V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis, "Predicting bank insolvencies using machine learning techniques," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1092–1113, 2020, doi: 10.1016/j.ijforecast.2019.11.005.
10. Z. Zhang, C. Wu, S. Qu, and X. Chen, "An explainable artificial intelligence approach for financial distress prediction," *Information Processing and Management*, vol. 59, no. 4, p. 102988, 2022, doi: 10.1016/j.ipm.2022.102988.
11. Arjovsky, M., Bottou, L., Gulrajani, I., & LopezPaz, D. (2019, July 5). Invariant risk minimization. arXiv.Org. <https://arxiv.org/abs/1907.02893>
12. Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3–32. <https://doi.org/10.1257/jep.31.2.3>
13. Chen, L., Pelger, M., & Zhu, J. (2019). Deep learning in asset pricing. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.3350138>
14. Drobetz, W., & Otto, T. (2021). Empirical asset pricing via machine learning: Evidence from the European stock market. *Journal of Asset Management*, 22(7), 507–538. <https://doi.org/10.1057/s41260-021-00237-x>
15. Feng, G., He, J., & Polson, N. G. (2018, April 25). Deep learning for predicting asset returns. arXiv.Org. <https://arxiv.org/abs/1804.09314>
16. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
17. Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1). <https://doi.org/10.1186/s11782-020-00082-6>
18. Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences. Cambridge University Press. <https://doi.org/10.1017/cbo9781139025751>
19. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
20. Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing*, 93, 106384. <https://doi.org/10.1016/j.asoc.2020.106384>
21. Pearl, J. (2009). Causality. Cambridge University Press. <https://doi.org/10.1017/cbo9780511803161>
22. Peters, J., Janzing, D., & Scholkopf, B. (2017). Elements of causal inference: foundations and learning algorithms. MIT Press.
23. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
24. Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3). <https://doi.org/10.1214/10-sts330>
25. Warin, T., & Stojkov, A. (2021). Machine learning in finance: a metadata-based systematic review of the literature. *Journal of Risk and Financial Management*, 14(7), 302. <https://doi.org/10.3390/jrfm14070302>