A Systematic Approach to Financial Risk Classification Using Machine Learning

Yiming Li, Tianyu Qu

1. Introduction

Financial risk assessment is a critical aspect of the financial industry, playing a significant role in banking, insurance, investment, and asset management. With the increasing reliance on data-driven decision-making, utilizing machine learning techniques to analyze complex financial data and build effective predictive models has become an essential area of research. Accurate financial risk evaluation allows organizations to mitigate potential losses, optimize resource allocation, and enhance the stability and efficiency of financial systems. The "Financial_Risk" dataset used in this study provides a comprehensive simulation of real-world financial scenarios, including demographic, financial, and behavioral data, while incorporating challenges such as imbalanced distributions and missing values.

The work for this assignment was carried out collaboratively by two team members, **Tianyu** and **Yiming**, who both actively participated in the coding process and each completed an independent implementation of the code. This approach ensured that all steps were thoroughly verified and optimized for accuracy and reproducibility. The division of responsibilities for the written report included the following:

- **Tianyu Qu** was responsible for drafting the sections on Introduction, Related Work, and Results.
- **Yiming Li** contributed by writing the sections on Proposed Method and Conclusions.

This project focuses on building a systematic framework for financial risk classification using machine learning, with particular attention to the impact of data preprocessing, feature selection, and model optimization on classification accuracy. Our work aims to address the inherent challenges in financial datasets and establish a scalable and robust solution. The study began with meticulous data preprocessing, where irrelevant columns were removed, missing values were handled using pairwise deletion and mean imputation, and categorical features were encoded using One-Hot Encoding and Ordinal Encoding. Additionally, features such as credit scores were binned into intervals using equal-width binning to simplify the modeling process. These steps ensured that the dataset was clean, consistent, and ready for subsequent analysis.

Feature selection played a pivotal role in refining the dataset by eliminating redundant or irrelevant predictors. Using correlation analysis and Pearson's method, highly correlated features were removed to reduce multicollinearity, and low-correlation features were excluded to focus on variables most relevant to the target variable. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-SNE were then applied to condense the data into lower-dimensional spaces, enabling effective visualization and revealing inherent clustering patterns within the data.

The project implemented two machine learning models, Random Forest and k-Nearest Neighbors (KNN), to classify financial risk. Model evaluation was conducted using accuracy as the primary metric, and cross-validation ensured the generalizability of the results. Hyperparameter optimization was performed to refine the models, leading to improved performance. Feature importance was further analyzed using Random Forest, which provided critical insights into the most influential predictors for financial risk classification, guiding refinements in feature selection and model optimization. To conclude the modeling phase, AutoML was utilized to streamline model selection and identify the best-performing algorithm, balancing accuracy and computational efficiency.

In summary, this project highlights the critical role of machine learning in tackling financial risk assessment challenges, emphasizing the value of a systematic and data-driven approach. By employing robust preprocessing methods, such as handling missing data and encoding categorical variables, alongside feature engineering techniques like correlation analysis and dimensionality reduction, we optimized the dataset for analysis. Model optimization, including hyperparameter tuning and cross-validation, further enhanced the accuracy and reliability of our predictions. The integration of feature importance analysis and AutoML streamlined the process of identifying key predictors and selecting the best-performing model. This study demonstrates the potential of machine learning to provide actionable insights from complex financial data.

2. Related Work

The application of machine learning in financial risk management has garnered significant attention, offering innovative approaches to identifying, evaluating, and mitigating risks in various financial domains. Traditional risk assessment methods often rely on statistical models that struggle to adapt to the complexities of modern financial datasets, including their high dimensionality, non-linearity, and frequent presence of missing or imbalanced data. Machine learning provides a powerful alternative, enabling automated and accurate risk predictions with improved adaptability to changing market conditions. By leveraging advanced algorithms, data preprocessing techniques, and model optimiza-

tion strategies, machine learning ensures higher predictive accuracy, faster processing, and adaptability, making it an essential tool for addressing evolving financial risks.

Mashrur et al. (2020) provided a comprehensive survey of machine learning applications in financial risk management, emphasizing supervised learning techniques like Random Forests, Support Vector Machines, and Logistic Regression for tasks such as credit scoring, fraud detection, and bankruptcy prediction. These methods were highlighted for their ability to handle structured data and provide high levels of accuracy when properly tuned. Additionally, the survey examined the role of unsupervised learning methods such as clustering for anomaly detection, which is particularly useful in fraud prevention, where labeled data is often unavailable. Reinforcement learning was also explored as a promising approach for dynamic portfolio optimization, enabling continuous learning and adaptation to market changes. Despite these advancements, the authors identified persistent challenges, including the difficulty of ensuring data quality in large-scale datasets, balancing model complexity with interpretability, and the computational demands of deploying these algorithms in real-time scenarios. They called for further research into techniques that bridge the gap between theoretical model design and practical implementation, particularly for high-stakes financial decision-making contexts.

Abdulla and Al-Alawi (2024) conducted a systematic review focusing on the evolution of machine learning techniques for systemic and credit risk management, exploring both traditional machine learning models and more recent advancements in ensemble methods and deep learning architectures. The authors discussed the increasing adoption of techniques such as Gradient Boosted Machines, Random Forests, and deep neural networks, noting their ability to capture complex relationships in large, high-dimensional datasets. They emphasized the significance of deep learning architectures, such as Long Short-Term Memory (LSTM) networks, for modeling temporal dependencies in financial data, which is critical for predicting systemic risks and defaults. The review also highlighted the importance of balanced datasets in improving the generalizability of models, as imbalanced data can skew predictions and lead to poor performance for minority classes. The authors further explored the trade-offs between model complexity and interpretability, pointing out that while deep learning models provide superior performance, their "black-box" nature limits their use in regulatory or legally sensitive contexts. As a solution, they recommended integrating explainability techniques, such as SHAP (SHapley Additive exPlanations), to improve transparency without compromising accuracy, ensuring that these models meet the high standards required for critical financial applications.

Tian et al. (2024) addressed the specific challenges faced by internet financial platforms, where speed and accuracy are paramount for tasks such as rapid credit evaluations, real-time fraud detection, and personalized financial services. Their review demonstrates the effectiveness of hybrid models, which combine deep learning with traditional statistical methods to leverage the strengths of both approaches. For example, deep neural networks excel at extracting patterns from unstructured data such as text and images, while decision trees and logistic regression are well-suited for structured, tabular data. By integrating these techniques, hybrid models achieve both predictive power and robustness, making them ideal for the fast-paced environment of internet finance. The authors also discussed the role of automated feature selection methods, such as LASSO regression and Recursive Feature Elimination (RFE), in reducing model complexity while maintaining performance. Furthermore, they emphasized the need to tailor machine learning solutions to the unique characteristics of each dataset, considering factors such as user behavior, transaction patterns, and platform-specific risks. However, the study noted that achieving scalability and maintaining accuracy across diverse platforms remain key challenges, requiring further advancements in automated model tuning and real-time data processing.

Our approach aligns with these studies by incorporating similar methodologies, such as feature engineering, dimensionality reduction, and model optimization, but extends them into a comprehensive, end-to-end framework. Unlike many prior studies that focus on either algorithmic advancements or specific use cases, our work emphasizes systematic preprocessing to address missing values and imbalanced data. Techniques such as pairwise deletion, mean imputation, and equal-width binning for credit scores ensure data quality and consistency before further analysis. Additionally, we utilized Pearson correlation analysis and Principal Component Analysis (PCA) for feature selection and visualization, streamlining the dataset to include only the most relevant predictors.

Furthermore, our study adopted a dual-model evaluation framework using Random Forest and k-Nearest Neighbors (KNN) algorithms, optimized through cross-validation and hyperparameter tuning. The integration of AutoML represents a practical advancement, automating model selection and parameter optimization to identify the best-performing algorithm for financial risk classification. This combination of traditional and automated techniques balances accuracy, efficiency, and scalability, making it highly applicable to diverse financial risk scenarios.

While previous studies have effectively demonstrated the potential of advanced machine learning models, our work focuses on integrating these techniques into a unified workflow. By addressing critical data challenges, optimizing models, and incorporating visualization for interpretability, our approach not only complements existing research but also provides a practical framework that can be adapted to a wide range of real-world applications in financial risk assessment.

3. Proposed Method

3.1. Data Understanding

In the initial step, the dataset was loaded into the environment. Necessary libraries for data processing and manipulation, including pandas for handling CSV files and numpy for linear algebra operations, were imported. The dataset was then retrieved using the `pd.read_csv()` function. The file path points to the data source, which contains 15,000 observations with 20 features each. Each observation is labeled with a target variable, `Risk Rating`, representing the financial risk category for a given individual. Features such as `Age`, `Gender`, `Credit Score`, and `Loan Amount` were included, along with details like `Employment Status`, `Debt-to-Income Ratio`, and `Previous Defaults`, offering a comprehensive view for risk assessment.

3.2. Data Preprocessing

To prepare the dataset for analysis and modeling, several preprocessing steps were undertaken. First, irrelevant features such as `Age`, `Gender`, `City`, `State`, and `Country` were removed as they did not contribute to the analysis. Next, missing values in the dataset were handled using the pairwise deletion method, which retains as much data as possible by excluding only incomplete pairs during specific analyses. For categorical variables, encoding techniques were applied: features with inherent order, such as `Education Level`, `Payment History`, and `Risk Rating`, were transformed using ordinal encoding, while unordered categorical features, including `Marital Status`, `Loan Purpose`, and `Employment Status`, were encoded using one-hot encoding. Continuous numerical features, such as `Income`, `Credit Score`, `Loan Amount`, `Years at Current Job`, and `Assets Value`, were binned into discrete intervals, and the bin indices were used in place of the original values. These preprocessing steps ensured the dataset was clean, consistent, and ready for further analysis and modeling.

3.3. Feature Selection

Feature selection was conducted using the Pearson correlation coefficient to evaluate the relationships between features. A heatmap was generated to visualize these correlations. The analysis revealed that there was no significant relationship among the features themselves, and their correlation with the target variable, `Risk Rating`, was also minimal. This suggests that the features are largely independent and exhibit weak associations with the target variable, highlighting the need for advanced modeling techniques to extract meaningful patterns.

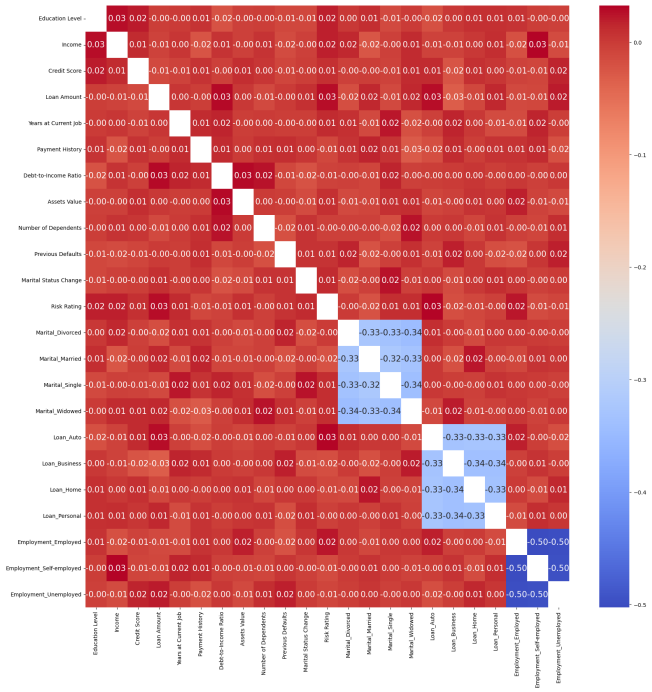

Figure 1. Heatmap of Pearson Correlations Among Features and `Risk Rating`

3.4. Heatmap Analysis

From the heatmap, it can be observed that there is almost no significant correlation among the features (correlation coefficients are close to 0), indicating that these features are relatively independent. Moreover, the correlation between these features and the target variable `Risk Rating` is also very low, suggesting that the current features have limited explanatory power for risk rating.

3.5. Dimensionality Reduction using PCA

To further explore the dataset, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the features. After performing PCA, the dataset was transformed into a set of uncorrelated principal components. These components were then analyzed to observe their contribution to the variance in the data.

The results showed that the majority of the variance could be explained by a small number of principal components, suggesting that the original dataset contained redundant information. However, even with PCA, the transformed components exhibited weak relationships with the target variable `Risk Rating`. This indicates that the dataset may lack strong predictors for the target variable, further emphasizing the need for advanced modeling techniques or additional feature engineering to improve predictive power.

To further analyze the dataset, Principal Component Analysis (PCA) was applied to reduce the dimensions to

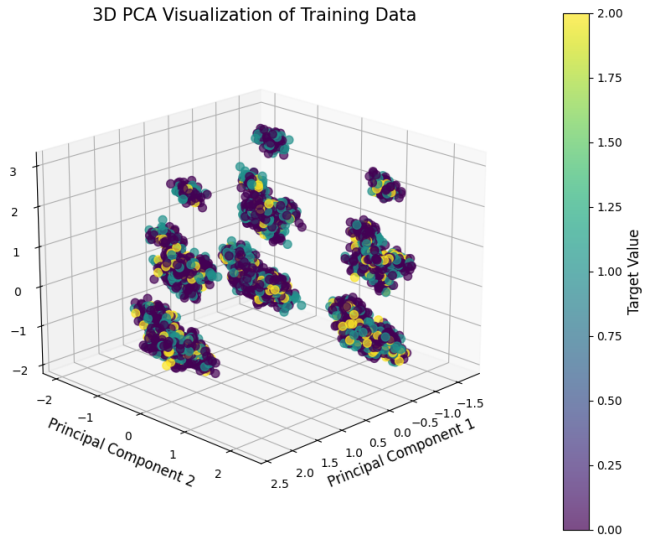

Figure 2. Explained Variance by Principal Components

two principal components. The resulting 2D scatter plot, shown in Figure 3, reveals interesting patterns in the data.

From the visualization, it can be observed that the data points form distinct clusters in the 2D space, indicating the presence of inherent grouping within the dataset. However, there is no clear separation of clusters based on the target variable `Risk Rating`, as indicated by the color gradient. This suggests that the principal components do not effectively capture the relationship between the features and the target variable. Additional feature engineering or advanced modeling techniques may be required to better represent the target variable.

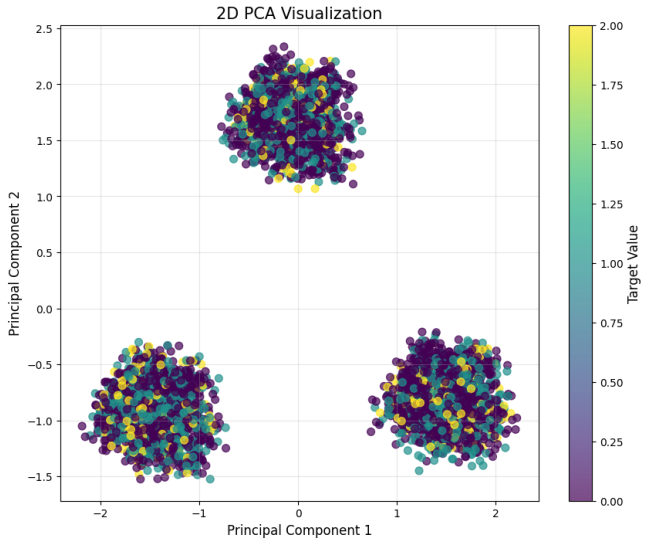

Figure 3. 2D PCA Visualization of the Dataset

3.6. Modeling with Random Forest

In the modeling phase, we utilized a Random Forest classifier to determine feature importance and evaluate the model's performance. The model was tested, yielding the following result:

- **Accuracy:** 0.5967

3.6.1. Feature Importance. The following table shows the features sorted by descending importance as determined by the Random Forest model:

Feature	Importance
Debt-to-Income Ratio	0.152297
Previous Defaults	0.075525
Credit Score	0.069381
Loan Amount	0.068757
Number of Dependents	0.068174
Years at Current Job	0.065017
Income	0.064314
Assets Value	0.064002
Payment History	0.060511
Education Level	0.060482
Marital Status Change	0.045904
Marital_Single	0.020307
Marital_Widowed	0.020130
Marital_Divorced	0.019968
Marital_Married	0.018955
Loan_Personal	0.018898
Loan_Home	0.018861
Employment_Unemployed	0.018196
Employment_Self-employed	0.017866
Loan_Business	0.017810
Employment_Employed	0.017366
Loan_Auto	0.017276

TABLE 1. FEATURES SORTED BY DESCENDING IMPORTANCE

3.7. Modeling with K-Nearest Neighbors (KNN)

In addition to Random Forest, we employed the K-Nearest Neighbors (KNN) algorithm to classify the dataset. The left panel of Figure 4 illustrates the confusion matrix for $k = 1$, while the right panel shows the accuracy scores as the value of k increases.

From the confusion matrix, it can be observed that:

- Class 0 has the highest number of correct predictions (717), indicating that the model performs relatively well for this class.
- Misclassification is significant for Classes 1 and 2, as their true labels are often confused with Class 0.

The accuracy score improves with increasing values of k, as shown in the right panel of Figure 4. However, the accuracy plateaus at around $k = 25$, suggesting that further increases in k do not significantly enhance model performance.

Overall, while KNN achieves moderate performance, it exhibits limitations in handling imbalanced or overlapping data, which likely contributes to the misclassifications seen in the confusion matrix.

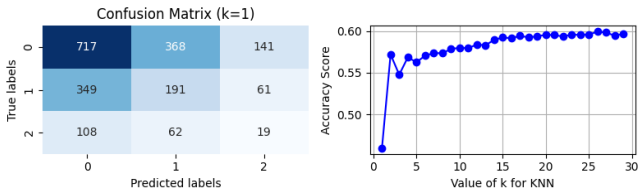

Figure 4. KNN Results: Confusion Matrix for $k = 1$ (Left) and Accuracy vs. k (Right)

3.8. Cross-Validation and Hyperparameter Optimization

To evaluate the models and optimize their hyperparameters, we performed cross-validation and utilized GridSearch for hyperparameter tuning. The results for both Random Forest and K-Nearest Neighbors (KNN) are summarized below:

3.8.1. Random Forest.

- **Cross-Validation Scores:** [0.58482143, 0.58854167, 0.58407738, 0.58779762, 0.58035714]
- **Mean Accuracy:** 0.585 with a standard deviation of 0.00292
- **Best Score:** 0.5851
- **Best Parameters:** {`n_estimators`: 100, `random_state`: 42}

3.8.2. K-Nearest Neighbors (KNN).

- **Cross-Validation Scores:** [0.5610119, 0.55431548, 0.56770833, 0.5610119, 0.55729167]
- **Mean Accuracy:** 0.560 with a standard deviation of 0.00449
- **Best Score:** 0.5847
- **Best Parameters:** {`leaf_size`: 1, `n_neighbors`: 25}

3.9. AutoML: Optimal Model Selection

Finally, we employed an AutoML framework to automatically search for the optimal model and hyperparameters. The best-performing model selected by AutoML was the `ExtraTreesClassifier`, with the following configuration:

- **Criterion:** `entropy`
- **Max Features:** 0.2155
- **Max Leaf Nodes:** 16
- **Number of Estimators:** 4
- **Number of Jobs:** -1 (parallel processing)
- **Random State:** 12032022

The accuracy score achieved by the selected model on the test dataset was **0.6066**, outperforming the manually tuned models such as Random Forest and KNN. This demonstrates the effectiveness of AutoML in identifying high-performing models with minimal manual intervention.

4. Results

The results of this study demonstrate the potential of our proposed machine learning framework for financial risk classification. By combining rigorous preprocessing, feature selection, dimensionality reduction, and model optimization techniques, we addressed many of the challenges inherent in financial datasets. These challenges included missing values, imbalanced data, and weak predictor-target relationships. Our systematic approach not only evaluated the performance of two widely used machine learning models—Random Forest and k-Nearest Neighbors (KNN)—but also leveraged AutoML to identify the most effective model for our dataset. Through a combination of manual tuning and automated optimization, we gained deeper insights into the factors that influence financial risk classification accuracy and identified opportunities for future refinement. The following key results highlight the strengths and limitations of the models, feature importance analysis, and the broader implications of this work for financial risk modeling.

4.1. Model Performance and Insights

The Random Forest classifier outperformed the KNN model in terms of accuracy, achieving a mean accuracy of **58.5%** compared to KNN's **56.0%** after cross-validation and hyperparameter tuning. Random Forest's ability to handle high-dimensional and complex data contributed to its stronger performance, particularly in identifying the most critical features influencing financial risk. However, both models struggled to accurately classify minority classes, as reflected in the confusion matrix. This limitation underscores the challenges posed by the imbalanced nature of the dataset and highlights the need for advanced resampling or weighting strategies to address this issue.

Additionally, AutoML identified an ExtraTreesClassifier as the best-performing model, achieving an accuracy of **60.7%** on the test set. This result demonstrates the advantages of automated model selection and hyperparameter tuning, which allowed us to explore a broader range of models and configurations efficiently. The performance of ExtraTreesClassifier suggests that ensemble methods may be particularly effective for financial risk classification due to their ability to capture complex patterns and interactions within the data. However, even with these optimizations, the relatively modest accuracies indicate the inherent difficulty of the dataset, pointing to potential limitations in the available features or the need for more sophisticated modeling techniques.

4.2. Feature Importance and Dimensionality Insights

Feature importance analysis revealed that "Debt-to-Income Ratio," "Previous Defaults," and "Credit Score" consistently ranked as the most predictive variables across all models. These features align with financial intuition, as

they directly relate to an individual's ability to manage debt and their historical financial behavior. In contrast, variables such as "Loan Amount" and "Employment Status" contributed minimally to the predictions, suggesting redundancy or weaker associations with the target variable, "Risk Rating." This insight highlights the importance of data-driven feature selection in improving model interpretability and performance.

Dimensionality reduction techniques, particularly Principal Component Analysis (PCA), provided valuable insights into the dataset's structure. PCA successfully reduced the dataset to a smaller set of components while retaining most of the variance. Visualizing the first two principal components revealed clusters within the data; however, these clusters did not strongly align with the target variable, indicating that linear dimensionality reduction methods may not fully capture the complex relationships necessary for accurate classification. This finding suggests the potential value of exploring non-linear dimensionality reduction methods, such as t-SNE or UMAP, to uncover deeper patterns in the data.

4.3. Recommendations and Future Directions

The results of this study highlight several avenues for future research and practical applications. First, addressing the dataset's imbalanced nature is critical for improving classification accuracy, particularly for minority classes. Techniques such as Synthetic Minority Oversampling Technique (SMOTE) or class weighting could be employed to mitigate this issue. Additionally, incorporating external data sources or creating derived features could enhance the predictive power of the model by introducing new dimensions of information.

Advanced modeling techniques, such as neural networks or ensemble approaches like Gradient Boosted Machines, could further improve performance by capturing non-linear interactions and complex patterns. The success of AutoML in identifying the best-performing model underscores its potential as a key tool for streamlining the model development process, allowing researchers to focus on data preparation and feature engineering while automating the optimization process. Furthermore, integrating explainability tools, such as SHAP or LIME, could enhance the interpretability of complex models, ensuring they meet the requirements for regulatory and practical applications.

In conclusion, while the models evaluated in this study provide a solid foundation for financial risk classification, continuous refinement of the dataset and exploration of innovative techniques will be essential for achieving higher predictive accuracy and practical utility. This study not only demonstrates the potential of machine learning for addressing financial risk challenges but also offers a roadmap for future research in this critical area.

5. Conclusion

In this study, we explored a systematic machine learning framework for financial risk classification, emphasizing data preprocessing, feature selection, and model optimization to address challenges inherent in financial datasets. By employing techniques such as missing value handling, feature encoding, dimensionality reduction, and hyperparameter tuning, we established a robust workflow to enhance model performance and interpretability. The results demonstrated that ensemble methods, particularly Random Forest and the ExtraTreesClassifier identified through AutoML, were effective for this task, achieving competitive accuracy despite the complexity and limitations of the dataset.

Key findings included the identification of critical features such as "Debt-to-Income Ratio," "Previous Defaults," and "Credit Score," which aligned with financial domain knowledge. However, the models struggled with imbalanced data, and the weak correlation between features and the target variable highlighted the need for further feature engineering or external data integration. The dimensionality reduction analysis underscored the value of advanced visualization techniques but revealed the limitations of linear methods like PCA in capturing complex relationships.

References

[1] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine Learning for Financial Risk Management: A Survey," *IEEE Access*, vol. 8, pp. 203203–203230, 2020. DOI: 10.1109/ACCESS.2020.3036322.

[2] Y. Y. Abdulla and A. I. Al-Alawi, "Advances in Machine Learning for Financial Risk Management: A Systematic Literature Review," in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, IEEE, 2024, pp. 531–540. DOI: 10.1109/ICETSIS61505.2024.10459536.

[3] X. Tian, Z. Y. Tian, S. F. A. Khatib, and Y. Wang, "Machine Learning in Internet Financial Risk Management: A Systematic Literature Review," *PLOS ONE*, vol. 19, no. 4, pp. e0300195, 2024. DOI: 10.1371/journal.pone.0300195.