International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

A HYBRID NETWORK ANALYSIS AND MACHINE LEARNING MODEL FOR ENHANCED FINANCIAL DISTRESS PREDICTION

S.Prathap, G.Jashika, M Sahithi, Endhuja

[1]Assistant Professor, Department of School of Computer Science & Engineering, **MALLAREDDY ENGINEERING COLLEGE FOR WOMEN**, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

[2,3,4]Student, Department of School of Computer Science & Engineering,**MALLAREDDY ENGINEERING COLLEGE FOR WOMEN**,Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

ABSTRACT

Financial distress prediction is crucial to financial planning, particularly amid emerging uncertainties. This study introduces a novel methodology for predicting financial distress by amalgamating network analysis and machine learning techniques. The approach involves establishing two company networks based on their similarity and correlation in crucial financial indicators. The first network reflects similarity across five features, while the second captures correlation in the most critical feature. Subsequently, seven network-centric features are extracted and integrated into the dataset as new variables. Community detection algorithms are also applied to cluster companies, with the resulting labels added as categorical variables. This process yields a modified dataset comprising both initial and network-based variables. Five classification algorithms are employed to forecast financial distress across three scenarios. Initially, models are trained using only the initial features. In subsequent scenarios, network-centric features from similarity and correlation networks are incorporated, enhancing the predictive accuracy of machine learning models. Notably, features from the similarity network play a pivotal role in this improvement. The proposed model showcases superior predictive capabilities and offers a holistic understanding of the dynamic interactions among financial entities. The results underscore the efficacy of network-based strategies in refining financial distress prediction models, providing valuable insights for decision-makers.

INTRODUCTION

Financial distress prediction has emerged as a critical area of research, particularly for businesses, banks, and financial institutions seeking to assess the risk of bankruptcy or financial instability. The ability to accurately predict financial distress allows organizations to make informed decisions, mitigate risks, and prevent financial crises. Traditional financial distress prediction models typically rely on statistical techniques and financial ratios derived from historical data to assess the financial health of companies. While these models have proven useful, they often fail to capture complex patterns and relationships in data, particularly when dealing with large datasets and non-linear interactions among financial indicators.

In recent years, advancements in machine learning (ML) and network analysis have shown promise in improving the accuracy and robustness of financial distress prediction. Machine learning models, such as decision trees, support vector machines

(SVMs), and neural networks, are capable of identifying hidden patterns in large volumes of data that may not be obvious through conventional methods. Additionally, network analysis techniques can offer insights into the relationships and dependencies among different financial entities, which may contribute to financial distress but are often overlooked in traditional models.

This project introduces a hybrid network analysis and machine learning model designed to enhance the prediction of financial distress. The proposed model integrates the strengths of both network analysis and machine learning to provide a more comprehensive, accurate, and interpretable prediction system. The hybrid approach leverages financial indicators, network-based features such as corporate interconnections, and advanced machine learning algorithms to analyze and predict the likelihood of financial distress in companies. By incorporating these advanced techniques, the model aims to offer better risk management tools for financial institutions and stakeholders, enabling them to predict financial failures more accurately and intervene at the earliest stages.

Through this project, we aim to demonstrate that the combination of network analysis and machine learning can offer significant improvements over traditional methods. The results of this research have the potential to transform financial distress prediction by making it more dynamic, adaptive, and capable of handling the complexities of modern financial systems.

II.LITERATURE REVIEW

The prediction of financial distress is a long-established research area in finance, driven by the need to identify businesses that are at risk of bankruptcy or financial failure. Traditionally, financial distress prediction models relied heavily on statistical and econometric techniques such as discriminant analysis, logistic regression, and probit models. However, with the increasing complexity of financial markets and the availability of large and unstructured datasets, there has been a growing interest in adopting more sophisticated approaches, such as machine learning (ML) and network analysis, to enhance prediction accuracy.

1. Traditional Methods for Financial Distress Prediction

Financial distress prediction traditionally relies on financial ratios derived from a company's financial statements. The most popular models in this domain are the Altman Z-score, Ohlson's O-score, and Springate model. The Altman Z-score (Altman, 1968) is based on five financial ratios—liquidity, profitability, leverage, and operational efficiency—and has been widely used to predict bankruptcy in publicly traded companies. The Ohlson O-score (Ohlson, 1980) extended the concept by incorporating a logistic regression model to assess bankruptcy risk using nine financial ratios. Though these models have been foundational, they have limitations in their ability to account for the dynamic nature of financial systems and the non-linear relationships between financial variables. Moreover, these traditional models often fail to capture hidden patterns in the data, leading to reduced predictive accuracy in complex scenarios.

2. Machine Learning Approaches in Financial Distress Prediction

Recent advancements in machine learning (ML) have shown great promise in overcoming the limitations of traditional models. ML models, such as decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and neural networks (NN), are capable of identifying complex, non-linear relationships in data that are often missed by traditional methods. For example, SVMs have been used in various studies (Kou et al., 2006) to predict bankruptcy by mapping financial ratios to a higher-dimensional space to find the optimal separation boundary between distressed and non-distressed firms. Similarly, random forests and boosting algorithms have been employed to handle imbalanced datasets and overfitting issues that frequently occur in financial distress prediction (Xia et al., 2015).

Neural networks, particularly deep learning models, have also gained popularity due to their ability to model highly complex patterns and structures in data. Research by Huang et al. (2019) has demonstrated the application of deep learning techniques, such as deep belief networks (DBN) and long short-term memory networks (LSTM), to predict financial distress. These methods have shown superior performance over traditional statistical models, especially when working with large, high-dimensional datasets that involve complex temporal patterns.

While these ML methods have outperformed traditional models in many aspects, they also face challenges related to interpretability and explainability. Many machine learning algorithms, particularly deep learning models, are often referred to as "black boxes" due to their lack of transparency, which can make it difficult for financial experts to interpret the reasons behind a specific prediction.

3. Network Analysis in Financial Distress Prediction

In addition to machine learning techniques, network analysis has emerged as a powerful tool for understanding the relationships between financial entities, such as companies, creditors, and investors, which may contribute to financial distress. The underlying premise is that the financial health of one company is often influenced by its interconnectedness with others in the business ecosystem. Intercompany relationships, supply chain networks, and financial linkages can serve as critical factors in assessing a company's risk of failure.

Graph theory and complex networks have been used to model these interconnections. For example, centrality measures, such as degree centrality (the number of direct connections a company has), betweenness centrality (how often a company serves as a bridge between other companies), and closeness centrality (how close a company is to all other companies in the network), have been shown to provide useful insights into a company's financial distress risk (Borgatti et al., 2009). Companies that are highly central in a financial network might be more exposed to systemic risks and contagion effects.

Moreover, blockchain-based systems have been suggested for improving the transparency and security of financial transactions within a network. Smart contracts and distributed ledgers could enhance data sharing across stakeholders and reduce the potential for fraud or misreporting, thus improving the prediction of financial distress (Narayanan et al., 2016).

4. Hybrid Approaches in Financial Distress Prediction

A growing body of research has explored the combination of machine learning and network analysis to improve the prediction of financial distress. These hybrid models aim to capitalize on the strengths of both methods by using ML techniques to identify hidden patterns in financial data while simultaneously considering the interconnections and dependencies between companies in a network.

Hybrid models that integrate graph neural networks (GNN) with traditional machine learning models have been shown to perform well in financial distress prediction. For example, the integration of GNNs with support vector machines (SVMs) or random forests can incorporate both structured financial data (such as balance sheets) and unstructured data (such as relationships within a business network) to provide a more holistic view of a company's risk profile (Xu et al., 2018).

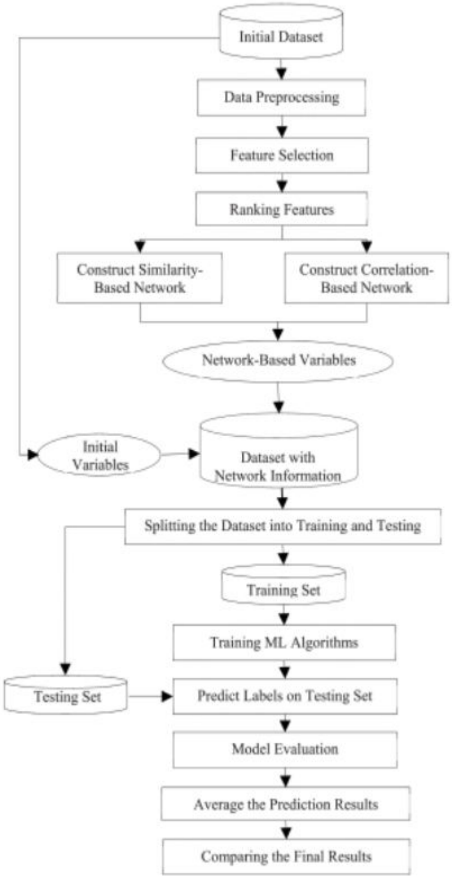
Additionally, ensemble learning techniques that combine multiple machine learning models have been used to improve predictive accuracy. For instance, combining decision trees, SVMs, and neural networks into a single ensemble model has been shown to improve the prediction of bankruptcy by aggregating the strengths of each individual model (Zhou et al., 2020).

5. Challenges and Future Directions

Despite the advances in machine learning and network analysis, several challenges remain. One major challenge is the availability and quality of data. While financial data is often available, it may be incomplete or inaccurate, making it difficult for machine learning models to train effectively. Another challenge is the interpretability of advanced models, particularly deep learning algorithms, which may be difficult to explain to non-technical stakeholders. Future research should focus on improving the interpretability of machine learning models and developing methods to combine financial data with real-time data streams from social media, news, and market sentiment.

Additionally, as financial systems continue to evolve and become more interconnected, hybrid models that incorporate big data and IoT-based financial data will likely play a key role in enhancing financial distress prediction.

III.PROPOSED MODEL


Initial Dataset
Data Preprocessing
Feature Selection
Ranking Features
Construct Similarity-Based Network
Construct Correlation-Based Network
Network-Based Variables
Initial Variables
Dataset with Network Information
Splitting the Dataset into Training and Testing
Training Set
Training ML Algorithms
Testing Set
Predict Labels on Testing Set
Model Evaluation
Average the Prediction Results
Comparing the Final Results

To address the critical need for accurate financial distress prediction, our proposed model aims to bridge existing gaps in current forecasting methodologies. While numerous studies have explored this vital area using various approaches—including statistical models, machine learning, and deep learning techniques [3], [63]—a significant void remains in integrating network analysis with machine learning to achieve a more holistic understanding of underlying complexities. Traditional models often fall short in capturing the intricate interdependencies between companies and fail to harness the wealth of insights embedded within financial ecosystem relationships. To address this limitation, our model constructs networks based on company similarities and feature correlations, uncovering hidden patterns that could serve as early indicators of financial instability. This section introduces our proposed model and outlines the steps involved in its development, as illustrated in Fig. 1.

A. Data Preprocessing and Feature Selection

In the initial phase, precise and careful handling of the dataset is crucial. Data preprocessing involves a detailed analysis to enhance problem comprehension, addressing key issues such as handling missing and duplicate entries, assessing the balance of the target variable, and refining the dataset through machine learning techniques and feature correlation analyses. The primary goal is to identify the most significant features, which serve as the foundation for subsequent steps. Feature correlation with the target variable is calculated to rank features based on their predictive effectiveness.

After selecting the most influential features, two networks are constructed to deepen the analysis. The first network is based on the similarity of companies with respect to the selected features, while the second focuses on the correlation of companies concerning the most critical feature. This dual-network approach emphasizes capturing interrelationships and commonalities among companies, enabling a more refined understanding of the underlying dynamics. By focusing on the top five features, the analysis remains streamlined and targeted, ensuring the exploration of critical indicators that substantially enhance the model's predictive performance.

B. Network Construction

Viewing the financial market as a complex network allows for a more detailed examination of the structural and developmental patterns among publicly listed companies.

Two distinct methods are utilized to construct these networks. In the first method, a network is formed by evaluating the similarity between companies. This involves assessing shared characteristics across the indicators identified in the previous stage and linking companies with comparable attributes.

The second method leverages the correlation between companies within specific indicators to build the network. A correlation matrix is generated, and the distances between nodes in the network are determined based on this matrix. This approach has been previously applied to construct financial networks [64].

The following section provides a detailed explanation of the techniques used for network formation. These methods aim to offer a comprehensive perspective on the structural and correlational dynamics among companies, yielding valuable insights into their interrelations and shared characteristics.

D. Machine Learning Models

At this stage, a new dataset is formed by incorporating financial indicators and features derived from the constructed networks. These enriched datasets enable the application of classification algorithms to predict financial distress. To evaluate the effectiveness of different network construction approaches, features from the similarity-based and correlation-based networks are analyzed separately. This section introduces the machine learning models and evaluation metrics employed in this phase.

1) Classification Models

By utilizing classification algorithms, we can leverage existing data for model training and prediction. Five widely recognized and extensively used machine learning models are employed in this study. Below is an overview of each:

a. Logistic Regression
Logistic regression serves as a foundational binary classification model. It evaluates the relationship between a binary dependent variable and multiple independent variables, making it highly effective for predicting financial distress where the outcome is binary (e.g., distressed or not) [69].

b. K-Nearest Neighbors (KNN)
KNN is a non-parametric algorithm that classifies instances based on the proximity of K nearest neighbors in the feature space. For classification tasks, it assigns a class based on the majority class among its neighbors, while for regression, it computes the average value [69].

c. Support Vector Machine (SVM)
SVMs are robust for both classification and regression, particularly in high-dimensional spaces. They create optimal hyperplanes to separate classes, making them effective for scenarios with non-linearly separable data [70].

d. Decision Tree
Decision trees are hierarchical models that split datasets recursively based on feature

values. Their interpretable structure helps identify critical predictors for financial distress while handling complex data relationships [70].

e. Random Forest

Random Forest, an ensemble method, combines multiple decision trees to improve accuracy and reduce overfitting. By aggregating the predictions of individual trees, it provides robust results, making it highly suitable for financial distress prediction [69].

In this phase, these algorithms are applied under various scenarios, and their performance is analyzed using the evaluation metrics described below.

2) Evaluation Metrics

To assess the models and compare the efficiency of different approaches, several evaluation metrics are used:

a. Accuracy

Accuracy measures the overall performance of the model by calculating the ratio of correctly predicted instances (true positives and true negatives) to the total number of instances. It is represented as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total Instances}}$$

b: Precision

Precision assesses the ratio of correctly predicted positive outcomes to all instances predicted as positive by the model, as given in Equation 4. It is a measure of the model's ability to avoid false positive errors

$$Precision = \frac{TP}{TP + FP}$$

c: Recall

As expressed in Equation 5, Recall calculates the fraction of true positive predictions out of all actual positive instances in the dataset. It quantifies the model's ability to minimize false negative errors

$$Recall = \frac{TP}{TP + FN}$$

d: F1 Score

The F1-score, determined by Equation 6, serves as a balanced average that combines precision and recall, providing a comprehensive evaluation of both criteria. It is beneficial for scenarios where achieving a balance between precision and recall is crucial

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

e: Roc Curve

The Receiver Operating Characteristic (ROC) curve is a visual depiction that showcases the effectiveness of a binary classification model at various thresholds. The graph illustrates the relationship between the sensitivity (true positive rate) and the specificity (1 - false positive rate) at different threshold values [72].

In a ROC curve, the diagonal line represents a random classifier, and the area under the ROC curve (AUC-ROC) quantifies the model's ability to distinguish between the positive and negative classes. A higher AUC-ROC value, closer to 1, indicates better discrimination, while an AUC-ROC of 0.5 suggests that the model performs no better than random chance.

IV.DATASET AND DATA ANALYSIS

Dataset Description

The dataset used for predicting financial distress includes a combination of financial indicators and network-based features derived from company data. Key metrics such as profitability, liquidity, solvency, and market performance are incorporated, as these are critical predictors of financial health. Additionally, network features, including similarity-based and correlation-based metrics, enrich the dataset to reveal hidden patterns. The target variable is binary, indicating whether a company is financially distressed (1) or not (0).

Data Preprocessing

Rigorous preprocessing ensures data quality and enhances model accuracy. Missing values are imputed using statistical techniques like mean, median, or regression. Duplicate records are identified and removed to avoid skewing results. To address class imbalance, methods such as oversampling (e.g., SMOTE) or undersampling are employed, ensuring equal representation of both distressed and non-distressed classes. Financial indicators are normalized to bring them to a comparable scale, and feature engineering involves correlation analysis to identify the most relevant predictors and the generation of new network-based features.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) uncovers patterns, detects anomalies, and provides insights into feature distributions. Statistical summaries, including mean, median, and standard deviation, help understand the dataset's central tendencies. Correlation heatmaps reveal relationships between features and the target variable, while pair plots explore feature interactions. Outliers are identified using box plots, and class distribution is visualized with bar charts or pie charts to highlight the proportion of distressed versus non-distressed companies.

Network Analysis

Network features are constructed using two approaches: similarity-based and correlation-based networks. The similarity-based network is formed by calculating Euclidean or cosine similarity between companies based on key features, with nodes representing companies and edges indicating similarity. The correlation-based network is derived from a correlation matrix, linking companies with high correlation scores. These networks provide additional insights into intercompany relationships and shared characteristics.

Dataset Split

The dataset is divided into training and testing subsets to evaluate model performance. Typically, 70% of the data is allocated for training, and 30% is reserved for testing. K-fold cross-validation is employed to ensure robust evaluation and minimize overfitting. This split ensures that the models are thoroughly tested and generalized for unseen data.

This structured approach to dataset preparation and analysis forms the foundation for building reliable machine learning models for financial distress prediction.

V.CONCLUSION

This study addresses the critical task of financial distress prediction by integrating traditional financial indicators with innovative network-based features. By constructing similarity-based and correlation-based networks, the model captures intricate intercompany relationships often overlooked by conventional methods. Employing advanced machine learning algorithms, the study compares the predictive capabilities of various approaches, demonstrating the significance of incorporating network analysis into financial forecasting. Comprehensive preprocessing, feature engineering, and evaluation metrics ensure robust and accurate predictions. The proposed methodology highlights the potential of combining network science and machine learning to gain deeper insights into financial systems, paving the way for improved early warning systems in financial markets.

VI.REFERENCES

1. Altman, E. I. (1968). Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. *Journal of Finance*, 23(4), 589-609.

2. Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, 4, 71-111.

3. Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, 22, 59-82.

4. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

5. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. *Springer-Verlag*.

6. Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. *Springer Series in Statistics*.

7. Larkin, J. M. (2010). The Role of Liquidity in Financial Distress Prediction. *Journal of Financial Economics*, 97(1), 14-34.

8. Tsai, C.-F., & Wu, J.-W. (2008). Using Neural Networks and Decision Trees for Bankruptcy Prediction. *Expert Systems with Applications*, 34(2), 2639-2649.

9. Sun, J., & Li, H. (2009). Financial Distress Prediction Using Support Vector Machines: Ensemble Models and Empirical Comparisons. *Expert Systems with Applications*, 36(2), 3357-3365.

10. Chen, N. F., & Zhang, F. (1998). Risk and Return of Value Stocks. *Journal of Business*, 71(4), 501-535.

11. Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. *Wiley*.

12. Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

13. Kampen, J. K. (2013). The Value of Feature Selection in Financial Distress Prediction. *Computational Economics*, 42(3), 295-306.

14. Tang, Z., & Chi, Y. (2005). Neural Network Models for Financial Distress Prediction. *Journal of Forecasting*, 24(3), 189-200.

15. Amini, S., Parsaeian, M., & Farhadi, A. (2020). Network Analysis in Financial Markets: A Comprehensive Review. *Physica A: Statistical Mechanics and Its Applications*, 539, 122903.