IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 March 2024, accepted 8 April 2024, date of publication 11 April 2024, date of current version 18 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387462

RESEARCH ARTICLE

A Hybrid Network Analysis and Machine Learning Model for Enhanced Financial Distress Prediction

SABA TAHERI KADKHODA AND BABAK AMIRI
School of Industrial Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran
Corresponding author: Babak Amiri (babakamiri@iust.ac.ir)

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.

INDEX TERMS Financial distress prediction, financial analysis, network-based analysis, machine learning, classification, community detection.

I. INTRODUCTION

Financial distress prediction is a pivotal realm within corporate finance and economics, presenting an ongoing challenge for investors, businesses, and the broader financial ecosystem. It's a major challenge and frequently a sign of an imminent crisis. Predicting financial crises signals imminent challenges and holds profound implications for preserving shareholders' capital.

As the pulse of economic stability, predicting financial distress has spurred extensive research, evolving from early multivariate statistical models like the Altman Z-score model [1] to the contemporary frontier of machine learning (ML) algorithms. Altman's model was based on logistic regression and was widely used and developed by other researchers [2], [3], [4]. Applying machine learning algorithms for forecasting financial distress has gained significant prominence in corporate finance [5]. This trend has witnessed substantial exploration, revealing promising outcomes in predicting financial distress. Numerous studies have emphasized the efficacy of diverse machine-learning techniques across various industries and countries. Notable methods include support vector machines (SVMs), artificial neural networks (ANN), ensemble classifiers, deep learning models, and hybrid machine learning technologies. Beyond numerical factors, integrating textual and management considerations has emerged as a valuable avenue [6]. When equipped with advanced techniques like network analysis and community detection, such models offer a nuanced understanding of the intricate dynamics within financial ecosystems. Through the exploration of these methodologies, researchers aim to enhance the accuracy of predictions and contribute to the development of effective risk management strategies, fostering resilience and stability in the face of

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell.

economic uncertainties. Analyzing the companies' networks helps us make better and more accurate predictions of the companies' financial health.

This research introduces a network-based approach for forecasting financial distress using machine learning algorithms. We construct a comprehensive network of businesses by evaluating the similarity and correlation of vital financial variables among companies. Subsequently, a thorough analysis of this network allows us to extract novel information and insights, which are then integrated into our prediction model. Establishing this institutional network enables the extraction of valuable information, contributing to refining our prediction models. Furthermore, this method is particularly beneficial for businesses with restricted available data, as it accurately determines their position within the network, enhancing our ability to predict financial distress.

II. LITERATURE REVIEW

Financial distress prediction is a critical area of research in finance and economics, aiming to identify early warning signs of financial instability in companies and containing different dimensions. The scientific literature frequently associates the concept of financial distress with bankruptcy, insolvency, the likelihood of default, and patterns of failure. Financial distress is commonly defined as the state of a company facing challenges in meeting its financial obligations [3]. In parallel, it is essential to distinguish between insolvency, bankruptcy, and distressed assets. Insolvency represents a company's inability to meet its financial obligations, indicating a severe financial crisis. On the other hand, bankruptcy is legal, often following insolvency, where a company undergoes a legal process to resolve its outstanding debts. Distressed assets encompass assets of a company facing financial distress, signalling a potential need for intervention.

In recent years, many researchers have employed various methods to predict various aspects of a company's future financial status, including bankruptcy, insolvency, and distress. For this purpose, researchers have presented their models and used different indicators for prediction.

The early identification of a firm's potential failure was facilitated by introducing financial ratios and key indicators. The literature has put forth an extensive array of ratios for this purpose [7].

In this area, the Altman Z-score is a widely used model of financial distress, measuring financial distress inversely [8]. The significance of financial ratio analysis, encompassing profitability ratios, debt ratios, liquidity ratios, and cash flow ratios, has also been underscored in predicting financial distress [9]. Moreover, the examination of corporate social responsibility's influence on financial distress has been a subject of study, where indicators such as the Altman Z-score and ZM-Score play pivotal roles in assessing the financial health of entities encountering distress [10].

In financial distress prediction, diverse indicators have emerged as a focal point for researchers seeking insights into various facets of a company's future financial standing, encompassing bankruptcy, insolvency, and distress. Traditional models, notably featuring financial ratios and statistical approaches, have served as enduring pillars in this domain. Noteworthy financial ratios such as return on capital employed, cash flows to total liability, asset turnover ratio, fixed assets to total assets, debt to equity ratio, and firm size, as highlighted by [11] have proven instrumental in predicting financial distress. Additionally, [12] delves into the predictive capabilities of traditional distress prediction models, particularly in identifying early-stage financial distress for firms. Leveraging financial ratios such as leverage, liquidity, profitability, and activity ratios [13]. Moreover, the impact of leverage on corporate financial distress has been studied, revealing a positive effect on financial distress measured by the Z-score [14]. A diversification strategy has also been investigated, with evidence suggesting its potential to reduce the level of financial distress [15]. The literature also emphasizes the importance of ownership structure, with studies indicating its effects on the likelihood of financial distress [8].

In terms of bankruptcy prediction models, various approaches have been explored, including the use of cash flow ratios, logistic regression, and the Altman, Ohlson, Springate, and Zmejewski models [16], [17]. Furthermore, the predictive ability of chosen bankruptcy models has been studied to assess credit risk and predict the financial situation to indicate the probable bankruptcy of a company [18]. The bias of unhealthy small and medium-sized enterprises (SMEs) in bankruptcy prediction models has also been addressed, emphasizing the importance of using a financial health indicator to construct the estimation sample for accurate predictions [19].

In conclusion, exploring various variables and indicators in the literature has provided valuable insights into the multifaceted landscape of financial distress prediction. Researchers have extensively examined factors influencing predictive models, from traditional financial ratios to diverse indicators encompassing ownership structure and corporate social responsibility. This foundational understanding sets the stage for the subsequent discussion on diverse models employed in the field, offering a comprehensive overview of the evolving strategies for financial distress prediction.

In addition to the indices used, the models employed in this field also play a significant role in predicting financial distress. Machine learning approaches have gained traction in financial distress prediction in recent years. The application of machine learning has further promoted studies on financial distress prediction and improved financial distress prediction accuracy [20]. In recent years, machine learning algorithms have become extensively utilized in predicting company financial difficulties [21]. Previous studies have applied the support vector machine prediction model in financial distress [22]. In addition, the rapid advancement of computers and software has given rise to other techniques such as data mining, machine learning, deep learning, and artificial

intelligence [23]. Furthermore, machine learning models have become a trend in quantitative finance, leading to a surge in interest in big data applications in the financial markets [24]. Reference [25] emphasized the focus on statistical and machine learning models for predicting financial distress. Machine learning methods, including artificial neural networks, support vector machines, and random forests, have demonstrated their effectiveness in forecasting financial trouble [26] machine learning methods, including random forest, support vector machines, and neural networks, effectively predicted financial distress and established early warning mechanisms for companies. Furthermore, [23] highlighted the emergence of machine learning techniques, including data mining, deep learning, and artificial intelligence, in financial distress prediction.

Moreover, comparing traditional and machine learning models has been a subject of interest. Several studies have highlighted the superiority of machine learning-based models over conventional methods in predicting corporate financial distress [11]. These machine-learning techniques include support vector machines, deep learning models, hybrid machine-learning technologies, genetic algorithms, and neural network models [20], [21], [22], [27], [28]. Reference [29] compared traditional methods such as logistic regression with machine learning models like random forest and neural networks to identify the model with the highest predictive accuracy of financial distress. This comparison sheds light on the effectiveness of machine learning approaches in financial distress prediction. Furthermore, [21] proposed combining financial management theory with machine learning algorithms to develop effective methods for predicting financial distress. Additionally, [24] and [30] we have enhanced the financial distress warnings evaluation system by incorporating ecological efficiency, indicating the potential for integrating diverse factors into distress prediction models.

The application of these models has been observed in various sectors, such as banking, real estate, and manufacturing, indicating the versatility of machine learning in financial distress prediction [16], [31], [32].

Various types of financial data have been commonly employed in machine learning models to predict financial distress. These models have been widely used due to their capability to model complicated features of financial data [33]. The types of financial data commonly employed in machine learning models for predicting financial distress include financial ratios, discriminant analysis, and linear discriminant analysis [1]. Additionally, deep learning-based models have been designed to predict a company's financial distress [25]. Furthermore, the authors of a study used a decision tree and classified regression tree of machine learning models to predict the financial fraud of listed companies in the United States [34]. Moreover, machine learning models have been used to predict the behavior of various aspects of financial markets given some input features [35]. These models have also been applied in developing the financial time series forecasting task and found to outperform other statistical models [36].

In summary, exploring machine learning models in financial distress prediction reveals a dynamic landscape of innovation and adaptability. As we transition to the discussion on the financial data utilized with these models, it becomes evident that the intricate interplay between advanced algorithms and diverse datasets, coupled with the critical aspect of feature selection, is crucial for achieving accurate predictions. Feature selection, drawing considerable attention from researchers, is essential in enhancing the predictive capabilities of machine learning models and contributes to the robustness of financial distress prediction strategies.

Feature selection involves identifying the most relevant variables, while feature engineering involves creating new features from the existing ones to enhance the model's predictive capability. Accurate financial distress prediction relies on the careful selection of suitable features. Machine learning methods, including support vector machines, artificial neural networks, and deep learning models, have been extensively employed to predict financial crises. These models require careful feature selection and engineering to achieve optimal performance. Emphasized the widespread use of machine learning algorithms in corporate financial distress prediction [32].

For instance, [37] specific financial ratios, such as return on capital employed, cash flows to total liability, asset turnover ratio, fixed assets to total assets, debt to equity ratio, and firm size, significantly contribute to distress prediction. In addition, [9] emphasized using diverse business analytics techniques, including statistical and artificial intelligence methodologies, to enhance the precision of financial crisis prediction models.

Hybrid machine-learning techniques have also been explored for financial distress prediction, aiming to establish effective prediction models focused on developing an effective financial distress prediction model using hybrid machine-learning techniques [22]. In addition, [22] introduced a novel framework for a financial early warning system that enhances its effectiveness by integrating the unconstrained distributed lag model with commonly employed financial distress prediction models, such as the logistic model and SVM.

Furthermore, the role of macroeconomic determinants in corporate financial distress has been investigated, highlighting the importance of systematic variable selection approaches to develop alternative models of financial distress [23].

Moreover, the incorporation of machine learning technologies, such as big data analysis, network analysis, and sentiment analysis, has been recognized as a favorable method for conducting systemic risk analysis in the financial industry [38] This suggests that machine learning aids in predicting individual company distress and contributes to assessing and measuring systemic risk in financial markets.

Additionally, the use of specific financial ratios and indicators, such as return on equity (ROE), debt-to-equity ratio (DER), and current ratio (CR), in machine learning models has been found to significantly impact the prediction of financial distress conditions [39]. This emphasizes the importance of feature selection and the incorporation of relevant financial variables in developing accurate distress prediction models. Moreover, the literature indicates that machine learning models have demonstrated high prediction accuracy, ranging from 78% to 93%, in forecasting financial distress, surpassing the performance of traditional discriminant models [27]. This highlights the potential of machine learning models in providing robust and reliable predictions for financial distress, thereby assisting stakeholders in making informed decisions.

In recent years, the convergence of finance, network analysis, and machine learning has spurred the development of inventive methodologies for predicting and comprehending financial distress. Extensive applications of network analysis have emerged as a powerful tool to unravel the interconnectedness of financial challenges within diverse organizations or industries. This approach involves deploying statistical connection indicators, including correlation, granger causality, and tail dependency. These metrics delineate and scrutinize the intricate network of transmission effects among financial institutions, providing valuable insights into the dynamics of financial distress [40]. Market-based measures of interconnectedness have been used to investigate the relationship between financial stability and interconnectedness, with extensive literature reviews discussing various financial network models [41]. Additionally, qualitative approaches through literature reviews have been used to systematically analyze the network formed by financial distress literature in specific sectors [42]. Furthermore, network analysis has been utilized to predict bank distress and to provide support for measures of interconnectedness in early-warning models, aiming to capture the interconnectedness among financial entities that could trigger the formation of contagion channels [43], [44]. Assessing interconnectedness in financial institutions has been identified as an early warning indicator for distress in financial [45]. Several methodologies have been suggested in scholarly works to gauge the level of interconnection across financial institutions and systems [46].

Moreover, network models have played a crucial role in managing systemic risk by capturing the interconnectedness among financial entities, which could lead to amplifying shocks to the financial system [44]. The interconnectedness measures in financial networks are based on the topology of links between banks, insurers, and financial services companies [47]. Additionally, network analysis has been used to model financial distress propagation on customer-supplier networks, allowing the investigation of possible scenarios for the functioning of financial distress propagation and the assessment of economic health within the network [48].

In the context of specific sectors, such as the textile and garment industry, network analysis has been employed to predict financial distress and its impact, aiming to determine the effect of profitability, liquidity, and solvency on financial distress in these sectors [42], [49]. Furthermore, network analysis has been applied to predict financial distress in various sectors, including infrastructure, utilities, transportation, and retail trade, using financial ratios and macroeconomic variables as independent variables [45], [50], [51], [52].

Applying network analysis, researchers have been able to discern systemically essential banks and possible contagion paths, shedding light on the interconnectedness of financial institutions and the potential for contagion [53]. Furthermore, the study of contagion and systemic risk in financial networks has evolved from seminal works to a review of subsequent literature, indicating the continuous development and relevance of network analysis in understanding financial distress [54].

In tandem with the foundational studies discussed in the financial distress prediction, recent contributions have significantly shaped the landscape of credit prediction and financial network analysis. In this area [55] It introduced an innovative approach that explicitly considers feature interactions, offering a nuanced understanding of complex relationships within financial datasets. By incorporating contrastive learning techniques, the study sheds light on the subtle dependencies among features, contributing to more accurate credit predictions. This research significantly contributes to the evolving landscape of credit prediction methodologies by acknowledging and leveraging the power of feature interactions in financial modelling.

In the domain of shareholding networks, a notable exploration has been undertaken by [56]. This research employed a sophisticated Voting Game Approach to uncover and analyze control associations within shareholding networks. By employing game-theoretic principles, the study provided insights into the dynamics of control exerted by shareholders in corporate structures.

The use of complex network theory to model the structure of the financial system and analyze risk contagion, particularly in banking systems, has been widely adopted by [57], [58]. This approach has allowed for the systematic analysis of how network structure and bank characteristics affect solvency distress contagion risk in interbank networks [59]. As the economic and financial system becomes more complex, the use of complex networks to study systemic risk and risk contagion has become increasingly important [60].

The examination and prediction of situations in which the stability of actual financial networks is susceptible to contagion and the conditions under which it becomes widespread have been emphasized as crucial [48]. Additionally, the use of multilayer network analysis has been employed to model risk contagion in venture capital markets, providing insights into how risk can spread through connections between market participants and harm total market robustness [61].

Studies have also focused on the spatial network contagion of environmental risks among countries, demonstrating the applicability of network analysis beyond traditional

financial contexts [62]. Moreover, the warning of financial distress and bankruptcy has been investigated using financial data reported in financial statements, indicating the diverse applications of network analysis in understanding financial distress [30].

Statistical methods and machine learning have been extensively employed in prior research endeavors centered on bankruptcy and financial distress prediction. However, a notable gap exists, as no study has concentrated on creating a network structure among companies based on similarities in their financial data. It subsequently utilized this network analysis for predictive purposes regarding their future statuses. Establishing a network among companies based on limited financial indicators enables the examination of intricate relationships between them. Leveraging both network analysis and machine learning techniques, a heightened capacity emerges to forecast the probability of financial distress. This integrative methodology not only refines the accuracy of forecasting companies' future statuses but also allows for predicting potential bankruptcies, even when comprehensive financial information is limited. It facilitates precise predictions regarding a company's risk of financial distress and bankruptcy without necessitating an abundance of detailed financial data.

Therefore, the present study aims to delve more deeply into this phenomenon. The subsequent section will expound on the intricacies of forming networks among companies. Following this, we will undertake a detailed analysis of these networks. By harnessing the insights derived from this network analysis, our research will then progress to predicting the likelihood of financial distress for individual companies.

III. PROPOSED MODEL

In addressing the imperative need for accurate financial distress prediction, our proposed model seeks to bridge existing gaps in forecasting methodologies. Numerous studies have delved into this crucial area (financial distress prediction) and have used various methods, including statistical models, machine learning and deep learning models [3], [63]; a noticeable gap persists in integrating network analysis and machine learning techniques for a more comprehensive understanding of the underlying complexities. Traditional models often struggle to capture the intricate interdependencies between companies and fail to exploit the wealth of information embedded in the relationships within financial ecosystems. Our model aims to fill this gap by constructing networks based on company similarity and feature correlations, unlocking hidden patterns that may serve as early indicators of financial instability. In this section, we briefly propose the model and explain the steps of our process. This process is illustrated in Fig. 1.

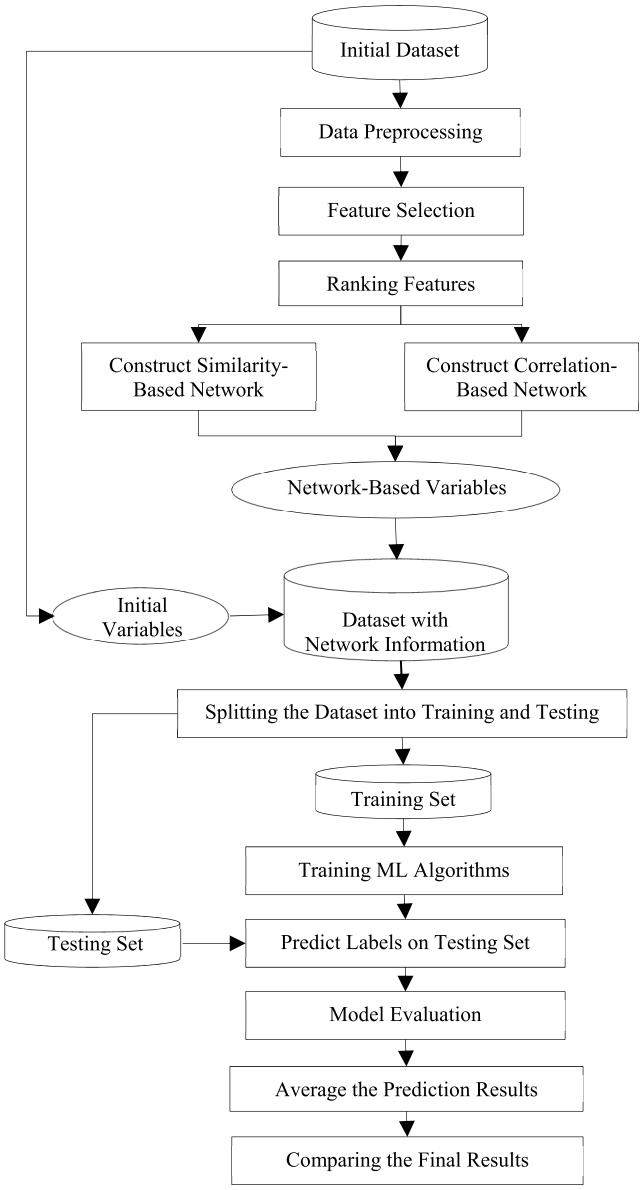

FIGURE 1. Flowchart of the proposed model process.

A. DATA PREPROCESSING AND FEATURE SELECTION

In the initial stage, meticulous handling of the dataset is essential. The data preprocessing phase involves a comprehensive analysis to enhance problem understanding. This includes addressing missing and duplicate data, evaluating target variable balance, and refining the dataset using machine learning techniques and feature correlation analyses. Identifying highly significant features is a pivotal outcome, forming the foundation for subsequent steps. We calculate feature correlation with the target variable to rank features based on their efficacy in prediction. After distilling the most influential features, the next step involves constructing two networks. One is based on the similarity of companies in these selected features, and the other is based on the correlation of companies in the most crucial feature. This strategic move aims to capture and emphasize the interrelationships between companies based on their commonalities, setting the stage for a more nuanced understanding of the underlying dynamics. Utilizing the top 5 features streamlines the analysis

and ensures a focused exploration of critical indicators that contribute significantly to the model's predictive capacity.

B. NETWORK CONSTRUCTION

Understanding the financial market as an intricate network enables us to scrutinize publicly listed companies' structure and developmental patterns.

Two different methods have been employed to construct a network among companies. In the first approach, the network is formed by assessing the similarity between companies. This involves gauging the similarity among companies in specific indicators mentioned in the preceding subsection, linking companies with comparable characteristics.

The second method employs the correlation between companies in specific indicators for network construction. This approach generates a correlation matrix among companies, and the distances between network points are subsequently determined based on this matrix. This method has previously been utilized to construct a financial network [64]. In the following subsection, we will elaborate on the methods employed for network formation. These strategies aim to comprehensively understand the relationships between companies, offering insights into their structural and correlational dynamics.

1) CONSTRUCT THE NETWORK BASED ON SIMILARITY

To construct a network based on the similarity, we utilized the distances between companies in 5 feature values described in the previous section. We first normalized the dataset to avoid the undue influence of any single feature, considering their distinct ranges. This normalization ensures that each of the selected five features contributes equally to network formation, fostering a balanced and unbiased representation.

We obtained five matrices by calculating the distances between companies for each of the chosen features, each presenting the inter-company distances within a specific feature. Subsequently, a consolidated matrix emerges through the computation of the average distances within these matrices. This matrix (D), representing collective distances, is the foundation for network formation. In this matrix, $d_{i,j}$ is average distance between company i and j. In the subsequent stages, this matrix will be instrumental in delineating the relationships and connectivity patterns between companies based on their similarities in the selected features. This method ensures a fair consideration of all chosen features and fosters a holistic understanding of the nuanced interplay between companies in the network.

Then, we use the K-Nearest Neighbors (KNN) algorithm to construct the network. Using the distance matrix (D), each node will be connected to its K nearest neighbors. To obtain the best number of neighbors, we will try the algorithm with different K numbers in our dataset and choose the best number. This method to construct the network was earlier used to create a network between time series [33].

2) CONSTRUCT THE NETWORK BASED ON CORRELATION

In the second approach, we constructed financial market networks using a methodology grounded in correlation matrices. This involved calculating the correlation coefficient between pairs of companies based on a chosen financial indicator.

Upon computing the correlation coefficients, a matrix of dimensions N×N is obtained, where N represents the number of companies in consideration. Subsequently, the distance between any two companies can be determined using Equation 1. This method allows for a quantitative representation of company relationships, emphasizing their interconnectivity in the financial market.

$$d_{i,j} = \sqrt{2(1 - \rho_{i,j})} \tag{1}$$

In this Equation $\rho_{i,j}$ is the correlation coefficient between companies i and j.

The symmetric N×N distance matrix (D) illustrates the distances among N firms. Utilizing this distance matrix, a weighted matrix (W) can be formed to depict the intricate topology of the network. company i and j are linked within the network, and the weight of the link ($w_{i,j}$) can be calculated by Equation 2, signifying the strength of the connection between the two companies.

$$w_{i,j} = \exp(-d_{i,j}) \tag{2}$$

C. NETWORK ANALYSIS

Following the formation of two networks using the methods discussed in the previous subsection, the next step involves analyzing each of these networks. This analysis encompasses extracting seven key network features and clustering companies within the network through community detection.

In the following subsection, we explain these steps.

1) FEATURE EXTRACTION

This research employed network-based variables as predictors for forecasting financial distress. The network-based features extracted from these two networks are as follows.

- Degree Centrality
- Betweenness Centrality
- Closeness Centrality
- Clustering Coefficient
- Page Rank Centrality
- Average Neighbor Degree
- Clustering Coefficient Weighted

To better understand these features, we briefly define them in this section.

a: DEGREE CENTRALITY

Degree centrality, a widely utilized centrality metric, quantifies the number of direct connections associated with a specific node. It can be understood as representing the immediate risk of a node in capturing entities flowing through the network, like information. In examinations of weighted networks, the concept of centrality degree is commonly broadened to encompass the sum of weights [65].

b: BETWEENNESS CENTRALITY

Interactions between two non-directly connected nodes depend on other nodes within the set, especially those along the paths connecting the two nonadjacent nodes. These intermediary nodes, often called 'in-between nodes,' may influence interactions between the two nonadjacent nodes. The concept of betweenness is rooted in the idea of network paths. Reference [66] defines a network path as a sequence of nodes traversed by following connections from one to another throughout the network. A geodesic path represents the shortest route through the network from one node to another. The betweenness of a node is computed as the fraction of the shortest paths between pairs of nodes that traverse through this specific node [65].

c: CLOSENESS CENTRALITY

It measures a node's closeness to all other nodes in a network, considering the shortest path distance. Nodes with elevated proximity centrality demonstrate the capability to connect with other nodes in the network quickly [65].

d: CLUSTERING COEFFICIENT

The metric known as the clustering coefficient of a node within a network quantifies the likelihood that its neighboring nodes are connected. This measurement serves as an indicator of the tendency of nodes to form clusters or communities [67].

e: PAGE RANK CENTRALITY

Page rank serves as an algorithm utilized by Google to assess the ranking of web pages within search results. In the context of network analysis, Page rank centrality quantifies the importance of a node by considering the importance of nodes that have links directed towards it [68].

f: AVERAGE NEIGHBOR DEGREE

This evaluates the mean degree of a node's neighbors within the network, offering insights into the local structure surrounding that particular node [68].

g: CLUSTERING COEFFICIENT WEIGHTED

This represents an expansion of the clustering coefficient, factoring in the strength or intensity of connections among nodes. It gauges the inclination of nodes to create clusters, incorporating the weights of their connections [61].

After computing these 7 features for each network, we added them as new features to the original dataset. As a result of this process, we have a new dataset including financial indicators and network-centric features as predictor variables.

2) COMMUNITY DETECTION

During the community detection phase, we employ a label propagation algorithm to categorize companies based on common characteristics or patterns within the network. These identified clusters then act as supplementary labels or categories for the diverse attributes present in the dataset. The assignment of these categories can yield valuable insights, potentially enhancing the predictive accuracy of the models. In this step we used a community detection algorithm for two networks (networks based on similarity and correlation). The number of clusters and nodes in each cluster are different in each network. In the next section, we present each network cluster characteristic.

This strategic categorization contributes to a more nuanced understanding of the dataset's structure and fosters the potential for refining and optimizing predictive models based on shared characteristics among companies. After this step, we add cluster number as a definite feature to the dataset.

D. MACHINE LEARNING MODELS

In this step, we have a new dataset including financial indicators and new features obtained from the network.

Now, we can use different classification algorithms, such as Support Vector Machines and Decision Trees, to predict financial distress. In this phase, to compare different approaches to network construction, we use features from the similarity-based network and the correlation-based network separately and analyze the obtained results. The following subsection introduces machine learning models and evaluation matrices used in this phase.

1) CLASSIFICATION MODELS

Utilizing classification algorithms allows us to harness the existing data for both model training and prediction. In this section of the study, we have incorporated five widely recognized and extensively used machine learning models. The subsequent paragraphs provide a succinct overview of each model.

a: LOGISTIC REGRESSION

Logistic regression is a foundational model for binary classification tasks. It assesses the relationship between the dependent binary variable and the independent variables, making it well-suited for predicting financial distress where the outcome is binary either distressed or not [69].

b: K-NEAREST NEIGHBORS (KNN)

KNN, a non-parametric algorithm, classifies or predicts an instance based on its proximity to the K nearest neighbors within the feature space. The algorithm's output is influenced by the most predominant class among its K nearest neighbors in the context of classification, or it computes the average of their values in the case of regression [69].

c: SUPPORT VECTOR MACHINE (SVM)

Support vector machines are potent models for classification and regression tasks. SVMs excel in delineating decision boundaries in high-dimensional spaces, making them advantageous in scenarios where the data may not be linearly separable [70].

d: DECISION TREE

Decision trees are versatile models adept at handling complex relationships within data. These hierarchical structures recursively split the dataset based on features, providing interpretable results and identifying significant predictors for financial distress [70].

e: RANDOM FOREST

Random Forest is an ensemble learning model that combines multiple decision trees to enhance predictive accuracy. By aggregating the results of various trees, Random Forest mitigates overfitting and yields robust predictions, making it valuable in the intricate landscape of financial distress prediction [69].

In this phase, we employ these algorithms in the dataset in several scenarios and then evaluate them with metrics we will introduce in the following subsection. The results will be presented and analyzed in the following sections.

2) EVALUATION

Diverse metrics such as accuracy, precision, recall, F-Score, and ROC curves are utilized to assess the models and compare the efficiency of different approaches. The performance of the implemented models is computed for each scenario. This section involves evaluating and comparing the results derived from the model. A detailed explanation of the outcomes obtained from the model on the subsequent dataset will be provided. We will introduce the metrics mentioned above to enhance comprehension of the model evaluation stage.

a: ACCURACY

This statistic measures the overall efficiency of our model, calculated as the proportion of accurately predicted instances (true positives and true negatives) to the total number of data points using Equation 3. In this context, TP denotes true positives, and TN denotes true negatives. Essentially, accuracy represents the proportion of accurate predictions [71].

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ data} \tag{3}$$

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 [71].

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

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 [71].

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

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 [72].

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

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

A. DATASET

Iranian Small and Medium-sized Enterprises (SMEs) play a vital role in the country's economy, contributing to job creation, innovation, and economic growth. As these SMEs navigate various financial challenges, predicting and preventing financial distress becomes crucial for their sustainability. This study presents a comprehensive dataset focused on Iranian SMEs, aiming to enhance our understanding of the factors influencing financial distress.

Our study exclusively incorporated audited financial statements with unqualified (clean) opinions to uphold the highest data quality standards. This stringent criterion was applied to guarantee the reliability and credibility of the financial data under scrutiny. By exclusively including audited statements with clean opinions, we aimed to mitigate potential inconsistencies and enhance the overall trustworthiness of the dataset. This meticulous selection process reflects our commitment to maintaining data integrity and upholding the robustness of our analytical approach.

The dataset under investigation in this research encompasses a diverse range of financial and non-financial variables, providing a holistic view of the financial health of Iranian SMEs. This dataset includes variables related to 422 Iranian SMEs across various periods. Within this dataset, the financial indicators for each company are available for up to 14 different periods. Additionally, an index indicating the financial distress status of each company is accessible for each period. This is a score between -8.6 and 128.4 in our dataset. To have a better understanding of this variable Table 1 represents some statistical variables related to the financial distress index. To define the target variable

TABLE 1. Financial distress index descriptive statistics.

Data	Mean	STD	Min	Max
All Companies	1.04	2.65	-8.6	128.4
Healthy Companies	1.13	2.64	-0.49	128.4
Financially Distressed Companies	-1.41	2.69	-8.6	-0.5

as binary, a company is considered financially healthy if the financial distress index is more significant than -0.5; otherwise, it is supposed to be in financial distress. According to this approach, 136 financially distressed companies are against 286 healthy ones in the dataset.

In addition to the financial distress index, this dataset includes 83 financial and non-financial variables. These features pertain to the preceding era and might be utilized to forecast the organisation's financial problems. All of these features are continuous variables except x80, which is categorical.

To facilitate a clear presentation of the data, we have categorized these variables into nine distinct categories: earnings and per share analysis, Efficiency and Turnover Ratios, Financial Structure and Leverage, Growth and Sustainability, Industry and Non-Operating Factors, Liquidity and Solvency Ratios, Operational Efficiency and Expenses, Profitability Ratios, and Risk and Coverage Ratios.

Features categorized under ''Earnings and Per Share Analysis'' delve into earnings and per share metrics, offering valuable insights into the company's profitability on a per-share basis. ''Efficiency and Turnover Ratios'' gauge how effectively a company utilizes its assets while managing inventory and receivables. Features under ''Financial Structure and Leverage Ratios'' evaluate the company's financial structure and leverage, scrutinizing the balance between debt and equity in its capital structure. ''Growth and Sustainability'' features assess growth rates, providing insights into the company's sustainability and long-term viability. Considering non-industry factors and industry-specific metrics in this category adds contextual relevance to financial distress predictions. ''Liquidity and Solvency Ratios'' measure a company's capacity to meet short-term obligations, offering indicators of financial stability. ''Operational Efficiency and Expenses'' features evaluate overall operational efficiency, expenses, and financial health. ''Profitability Ratios'' assess the company's profit generation and operational cost management, shedding light on operational efficiency and overall profitability. Finally, ''Risk and Coverage Ratios'' concentrate on the company's ability to manage risks, fulfill obligations, and uphold financial health.

In the appendix section, we have included a detailed table presenting all the variables in the dataset along with their respective categories. This table is a valuable reference for readers seeking an in-depth understanding of the dataset structure and variable categorization.

TABLE 2. Selected features.

Feature	Absolute Value of Correlation	Rank
x75	0.791193	1
x76	0.784320	2
x68	0.783543	3
x78	0.782831	4
x77	0.741128	5

TABLE 3. Descriptive statistics.

Variable	Full Sample			
	Mean	STD	Min	Max
x75	27138	11179	7958	34501
x76	174	79	228	25
x68	165	75	25	215
x78	20	2	16	22
x77	19	3	12	22

In the following subsection, we will explain employing the proposed model's steps into this dataset and describe it more during each step.

B. DATA ANALYSIS

As we mentioned in section III, in the first step we prepared the dataset to analyze. In this process, we handled missing values and dropped duplicated records. Then to calculate the similarity, first, we must choose the most essential features in financial distress prediction. As mentioned in section III We calculate the correlation coefficient between all 83 and the target variables. Then we choose 5 variables with the highest absolute value of correlation coefficient.

These features including:

- *x75*: Cash Flow Per Share
- *x76*: Cash Flow to Sales
- *x68*: Inventory/Current Liability
- *x78*: Cash Flow to Equity
- *x77*: Cash Flow to Liability

Table 2 represents the absolute value of correlation between selected features and target variables.

To have a better analysis of these variables Table 3 reports descriptive statistics on these features in the whole dataset. Table 3 shows the average, standard deviation, maximum, and minimum values of the mentioned variables for all companies in the dataset. To have a better comparison between healthy and financially distressed companies we calculate these descriptive statistics for each category (healthy and financially distressed companies) separately. The results of this analysis are reported in Tables 4 and 5.

In the next steps, we use these 5 features to calculate the similarity between companies and construct the first network based on the average distance between companies in these features. Then, we use the correlation coefficient in the most important feature (x75) to construct the second network.

TABLE 4. Descriptive statistics for distressed companies.

Variable	Positive Cases			
	Mean	STD	Min	Max
x75	14451	9568	7958	34501
x76	84	63	228	25
x68	80	63	25	215
x78	18	2	16	22
x77	15	3	12	22

TABLE 5. Descriptive statistics for healthy companies.

Variable	Negative Cases			
	Mean	STD	Min	Max
x75	33171	5282	7969	34501
x76	217	39	228	26
x68	206	36	25	215
x78	21	1	16	22
x77	20	1	12	22

TABLE 6. Results of the algorithms with different KS.

K	Number of Cluster	Algorithm/Accuracy				
		Logistic Regression	KNN	SVM	Decision Tree	Random forest
10	14	94%	85%	94%	88%	89%
25	6	90%	92%	90%	89%	91%
50	4	95%	94%	95%	92%	94%

As mentioned in the previous section to form the network-based on similarity, we used KNN and algorithm. To find the best number of neighbors in this approach we try different numbers of K and construct the network and then use 5 classification algorithms that were introduced in section III. To choose the best K we compared the best accuracy obtained of the models. Table 6 presents the results of this approach.

As illustrated in Table 6, the algorithm exhibits generally acceptable accuracy across all scenarios. Furthermore, the optimal accuracy for all algorithms is achieved when the number of neighbors is set to 50. Additionally, the configuration of the number of clusters and the members within each cluster is such that, particularly in large datasets, community detection yields significant insights into each company's status.

On the other hand, an increase in the number of neighbors leads to computational complexity in the algorithm, reducing the number of clusters. Consequently, it becomes challenging to obtain substantial information through this approach. Therefore, several neighbors set to 50 is chosen as an appropriate value, forming the basis for constructing the network.

In the next step, we construct another network based on correlation between companies in x75 which is the most crucial feature in prediction. We calculate the correlation matrix and then create the D matrix a distance between nodes in the network. We also calculate W which is the matrix of weight of each link in the network. The formulation to get these matrices were explained in section III.

To avoid the excessive complexity of the network, a connection is established between two nodes only if the distance between them ($d_{i,j}$) is less than 1.5. If the distance exceeds 1.5, we consider there is no connection between the two points, preventing overlapping relationships.

After constructing these networks, we start to analyze them and extract 7 key network-centric features that we introduced in the section III we added these new features to the dataset: degree centrality, betweenness centrality, closeness centrality, clustering coefficient, page rank centrality, average neighbor degree, and weighted clustering coefficient.

We also use community detection and cluster 422 companies into 4 clusters in a similarity-based network and 4 clusters in a correlation-based network. We add the cluster labels as categorical features. Then, we implement the 5 classification algorithms mentioned in the previous section in the dataset. In this process, we split the dataset into training and testing sets in a ratio of 80% to 20%.

To compare and analyze the efficiency of network-centric features in prediction, initially, we use the original dataset and implant the algorithms. Then, we use the network-centric features of each network in addition to the original features. We evaluate the results with matrices introduced in section III and compare the results of these 3 scenarios in the next section. This stage aimed to explore whether discerning a company's position within its network of companies could offer adequate insights for predicting its financial status.

A heatmap reflecting the correlation among the network-based features was generated to enhance comprehension of the data employed in the proposed model. Furthermore, to delve into the correlations more deeply, features exhibiting substantial positive or negative correlations were individually depicted in a scatter plot in Fig. 2 and Fig. 3.

Figure 2 represents the similarity-based and Figure 3 represents the results of the correlation-based network. In Fig. 2, Page Rank Centrality and Average Neighbor Degree features exhibit a high correlation with a coefficient of 0.97. Additionally, a significant negative correlation between Page Rank Centrality and Weighted Clustering Coefficient is evident in this figure the correlation confidence between these features is -0.95. In addition, there is a significant negative correlation between Average Neighbor Degree and Closeness Centrality with the target variable. This aspect will be further analyzed in the next section. In the correlation-based network, Degree Centrality and Average Neighbor Degree have a high correlation coefficient of 0.98. and Average Neighbor Degree have a high correlation with a coefficient of 0.92 with Closeness Centrality. In addition, like similarity-based network, there is a negative correlation between the target and Average Neighbor Degree. The correlation between the target variable and other features will be investigated and analyze in the next section. Furthermore, we will evaluate and contrast the outcomes derived from the machine learning algorithms across the three scenarios above.

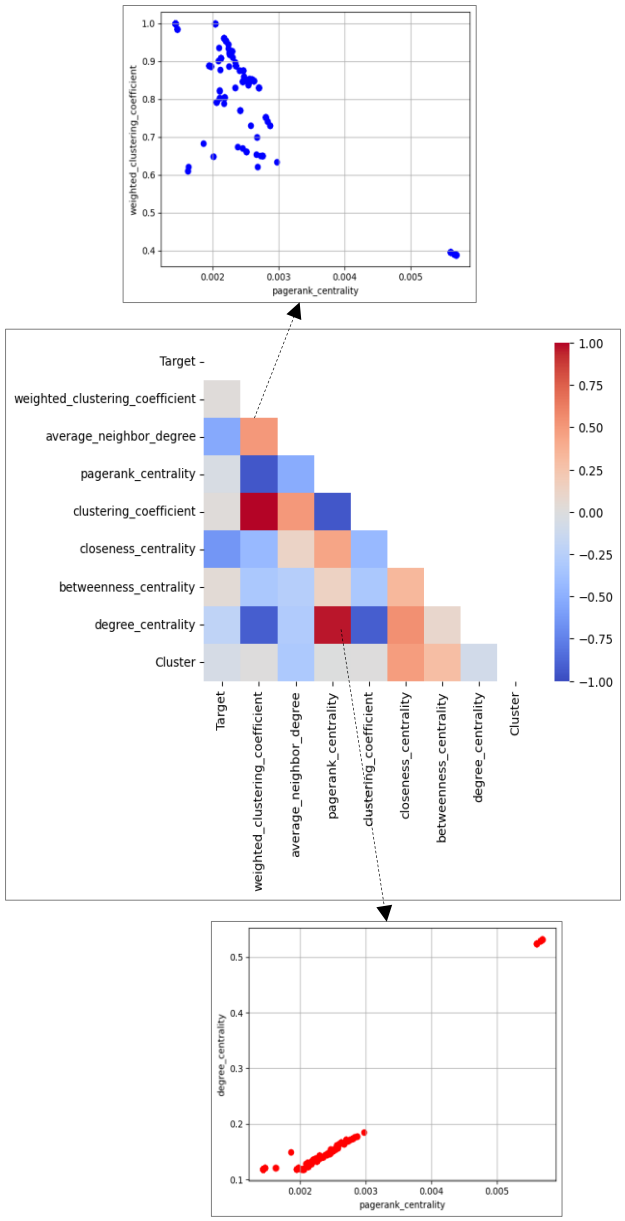

FIGURE 2. Correlation and scatter plots of the 2 highest correlated relationships in the similarity-based network.

V. RESULTS AND DISCUSSION

In this section, initially, we delve deeper into the results obtained from network construction and then compare the results of machine learning algorithms in 3 cases. Fig. 4 and Fig. 5 illustrate the similarity-based and correlation-based network. Different node colors represent each company's cluster.

When we use community detection to cluster the companies in both networks, we have 4 clusters, but the number of companies in each cluster is different. Table 7 represents the number of companies in each cluster. In addition, Fig. 6-13 illustrate each cluster's subnetworks. In these subnetworks, healthy companies and companies facing financial distress are differentiated by blue and red colors.

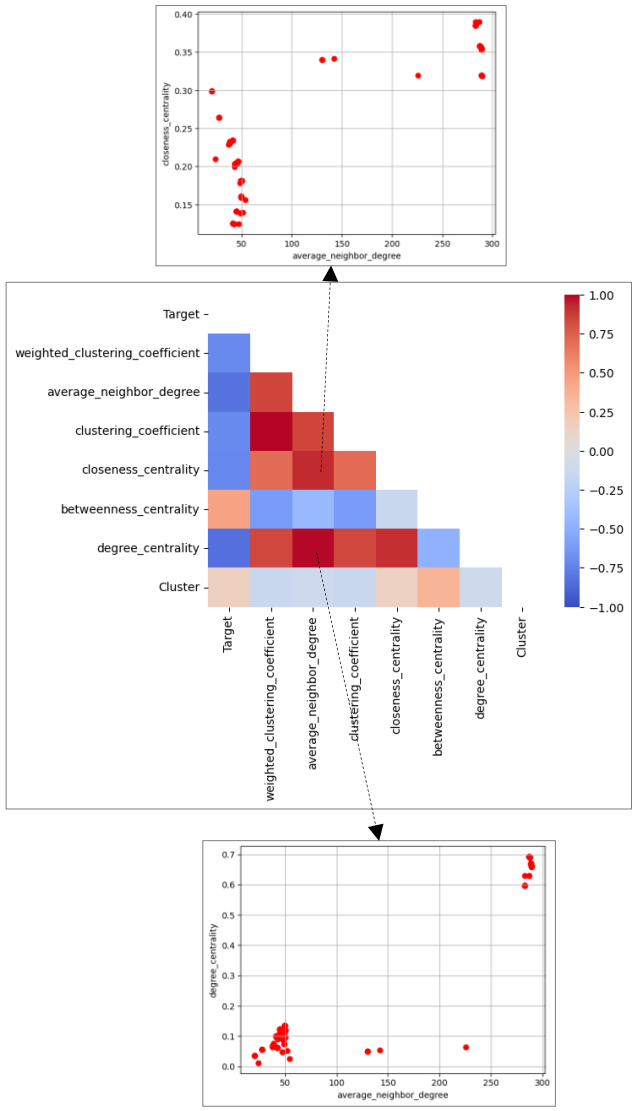

FIGURE 3. Correlation plot and scatter-plot of the 2 highest correlated relationships in the correlation-based network.

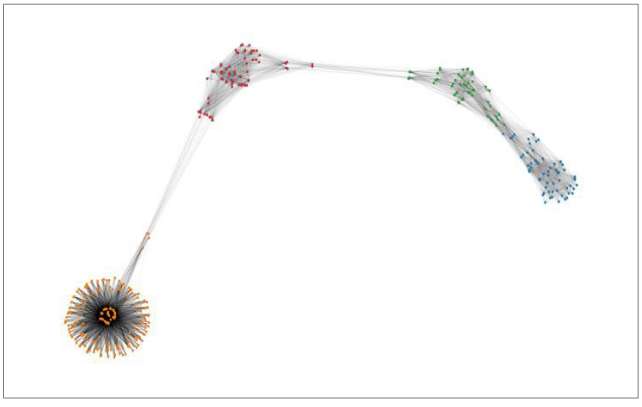

FIGURE 4. Network construction of similarity-based network: In this network each node represents a company and different node colors represent each company's cluster.

Table 8 and Fig. 6-13 show that the distribution of healthy and financially distressed companies within different clusters significantly differs. These findings suggest that analyzing a

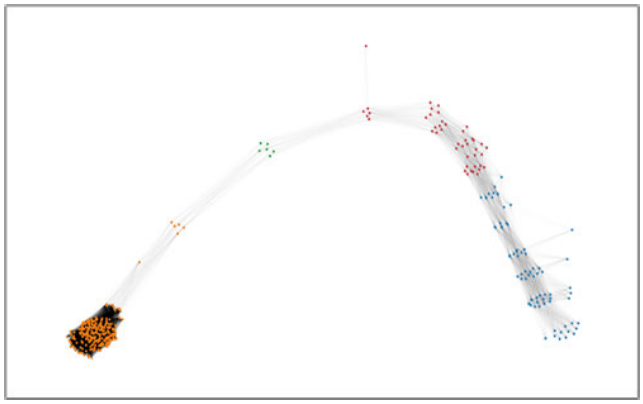

FIGURE 5. Network construction of correlation-based network: In this network, each node represents a company, and different node colors represent each company's cluster.

TABLE 7. Members of each cluster.

Cluster Label	Similarity-based Network		Correlation-based Network	
	Healthy	Distressed	Healthy	Distressed
Cluster 0	12	62	11	54
Cluster 1	273	25	216	9
Cluster 2	0	6	2	57
Cluster 3	1	43	57	16

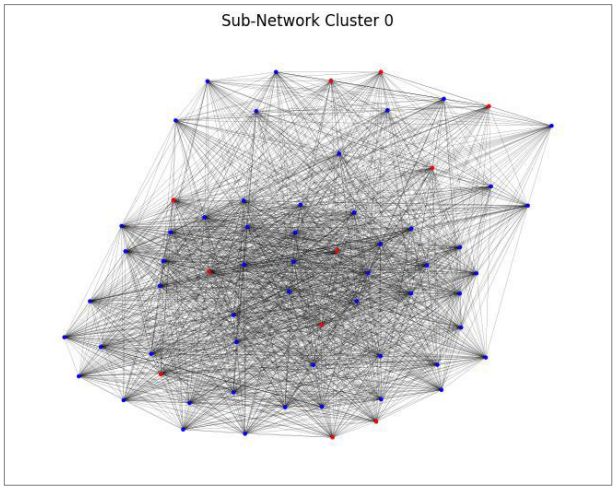

FIGURE 6. Subnetworks in similarity-based network: Cluster 0.

company's position within a network of other companies and identifying its cluster can provide significant insights into its overall health. Comparing the results from the two network formation approaches demonstrates that in both methods, the ratio of healthy companies to financially distressed companies varies within each cluster.

Continuing to examine the results obtained from the introduced algorithms in the previous section, we tested them in three scenarios. In the first scenario, only the features available in the initial dataset were utilized, and the features extracted from the network were not yet added. In this case,

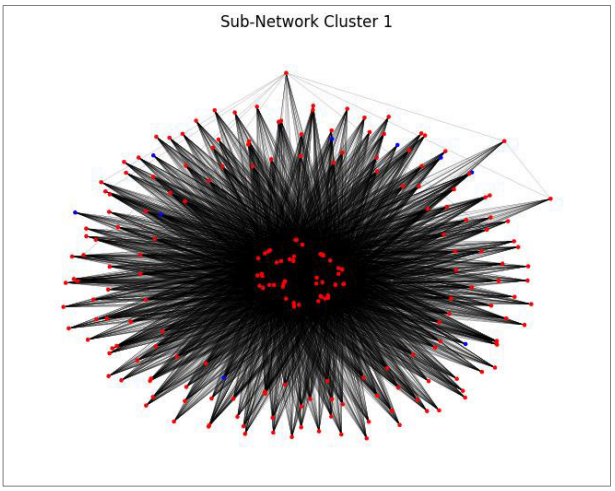

FIGURE 7. Subnetworks in similarity-based network: Cluster 1.

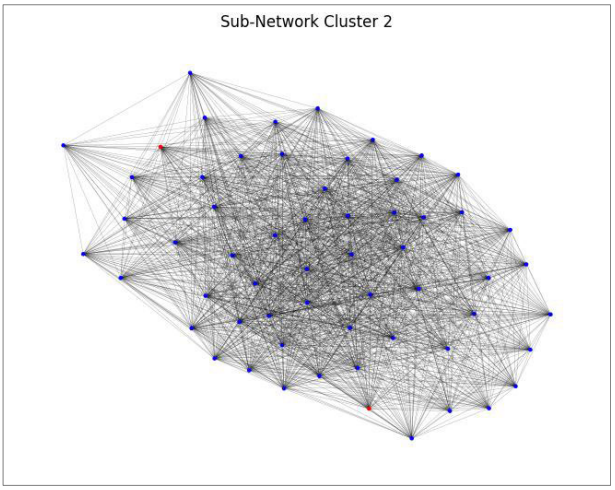

FIGURE 8. Subnetworks in similarity-based network: Cluster 2.

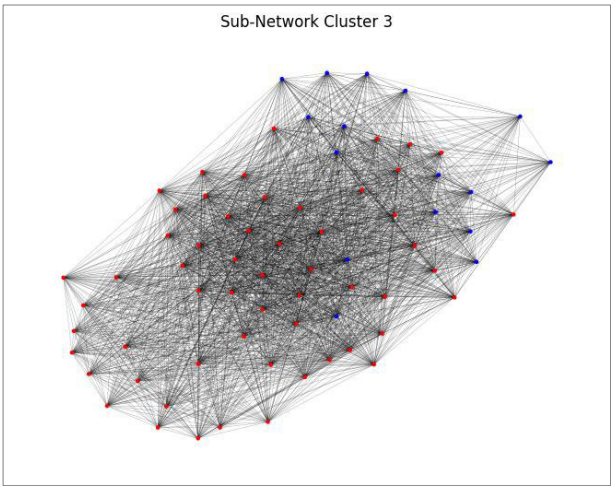

FIGURE 9. Subnetworks in similarity-based network: Cluster 3.

using the methods introduced in section III, we evaluated the prediction accuracy of each algorithm. Table 8 presents the results of this scenario.

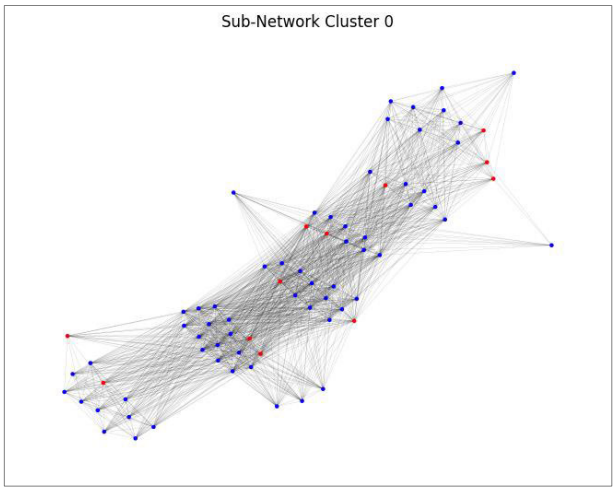

FIGURE 10. Subnetworks in correlation-based network: Cluster 0.

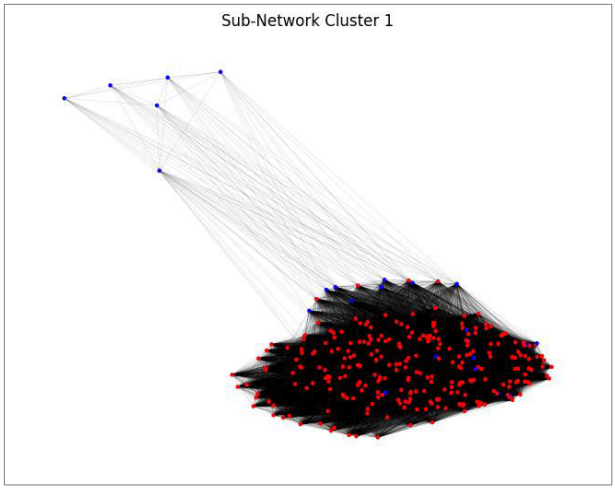

FIGURE 11. Subnetworks in correlation-based network: Cluster 1.

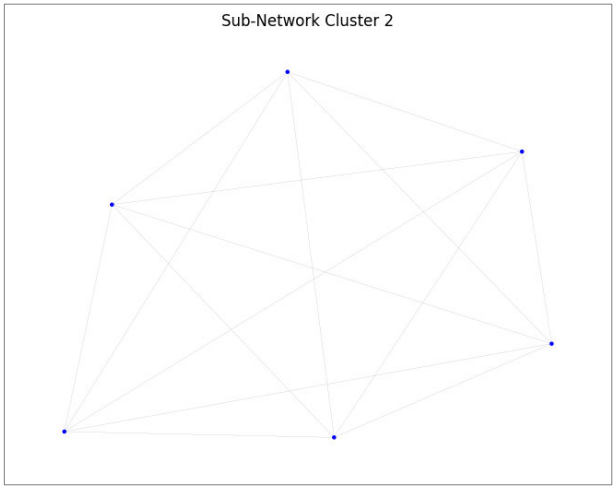

FIGURE 12. Subnetworks in correlation-based network: Cluster 2.

In the first scenario, Random Forest algorithms stand out with the highest accuracy of 94.12%, demonstrating its effectiveness in correctly classifying instances. It also achieves

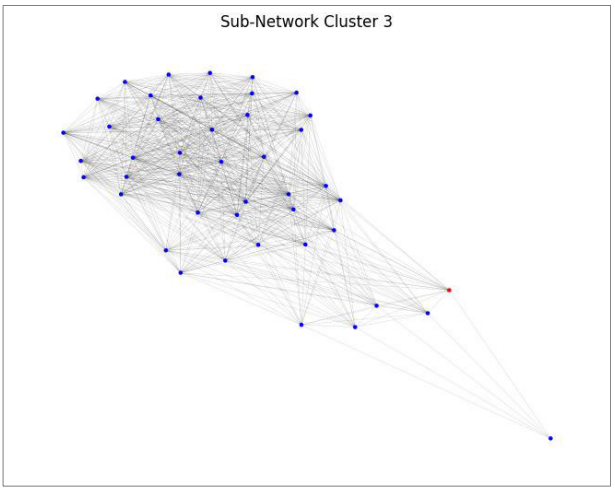

FIGURE 13. Subnetworks in correlation-based network: Cluster 3.

TABLE 8. Final results with initial features.

	Model	Accuracy	Precision	Recall	F1-Score
Initial Features	Logistic Regression	0.84	0.84	0.84	0.88
	KNN	0.83	0.81	0.86	0.88
	SVM	0.86	0.87	0.84	0.89
	Decision Tree	0.87	0.84	0.90	0.91
	Random Forest	0.92	0.94	0.91	0.94

TABLE 9. Final results with initial features and similarity network features.

	Model	Accuracy	Precision	Recall	F1-Score
Initial and Similarity Network Features	Logistic Regression	0.93	0.95	0.91	0.96
	KNN	0.95	0.99	0.91	0.98
	SVM	0.88	0.90	0.86	0.94
	Decision Tree	0.84	0.90	0.79	0.92
	Random Forest	0.95	0.95	0.95	0.98

notable precision, recall, and F1-score values, indicating a well-balanced performance in precision and recall. Decision Tree follows closely with an accuracy of 90.59% and intense precision. In comparison, Support Vector Machine (SVM) exhibits a well-balanced performance with a high accuracy of 89.41% and competitive precision and recall scores. Logistic Regression and KNN both achieve an accuracy of 88.24%, with Logistic Regression displaying balanced precision and recall. At the same time, KNN demonstrates slightly higher precision at the cost of a marginally lower recall. Overall, these results highlight the efficacy of Random Forest as the leading model in this experimental setup.

In the second scenario, when we add network-centric feature from the similarity-based network, KNN emerges as the top-performing model with an outstanding F1-Score of 0.98, showcasing a remarkable balance between precision and recall. It achieves a high accuracy of 95% and an impressive precision of 0.99, indicating its ability to make accurate positive predictions. Random Forest follows closely with an F1-Score of 0.98 and excels in both precision and recall, showcasing its robustness in handling the complexity of the dataset. Logistic Regression also performs well, achieving a high F1-Score of 0.96 and maintaining a good balance between precision and recall. SVM and Decision Tree exhibit slightly lower performance. Still, all models contribute to a comprehensive understanding of financial distress prediction, emphasizing the significance of incorporating initial and network-derived features.

In the last scenario, when we add network-centric features from correlation-based network, among the models, Random Forest stands out with the highest F1-Score of 0.94, reflecting its robust performance in achieving a balance between precision and recall. It also attains a noteworthy accuracy of 92%, indicating its proficiency in accurate predictions. Logistic Regression follows closely with a commendable F1-Score of 0.93, demonstrating a well-balanced trade-off between precision and recall. SVM and Decision Tree both exhibit strong performance with F1-Scores of 0.91, showcasing their effectiveness in financial distress prediction. While KNN demonstrates a slightly lower F1-Score, all models collectively contribute to a comprehensive understanding of financial distress prediction, emphasizing the significance of incorporating initial and correlation network features for enhanced model performance.

Fig. 14, 15 and 16 represent the ROC curve ratio for each model in the explained scenario. The ROC curves illustrate the trade-off between the sensitivity (the rate of correctly identifying positive cases) and the 1-specificity (the rate of incorrectly identifying negative cases) at various classification thresholds. The Logistic Regression algorithm excels in the first scenario with the initial features, indicating its strength in distinguishing between healthy and distressed companies. In the second scenario, achieving an AUC between 92% and 98% is commendable when we add features from the similarity-based network. Here again, Logistic Regression emerges as the top performer in this scenario Random Forest shows an excellent performance too. In the third scenario, when initial features and features from the correlation network are utilized, Logistic Regression stands out with an impressive AUC of 98%, signifying its effectiveness in accurately identifying companies at risk of financial distress of companies.

To further substantiate the effectiveness of our proposed model, we conducted comprehensive comparative experiments, aiming to provide robust evidence of its superiority. In addition to the statistical analyses of relations within the extracted procedure, we employed K-fold cross-validation (with accuracy as scoring metrics) across three distinct scenarios. This approach validates the model's performance and ensures its generalizability across diverse datasets and conditions.

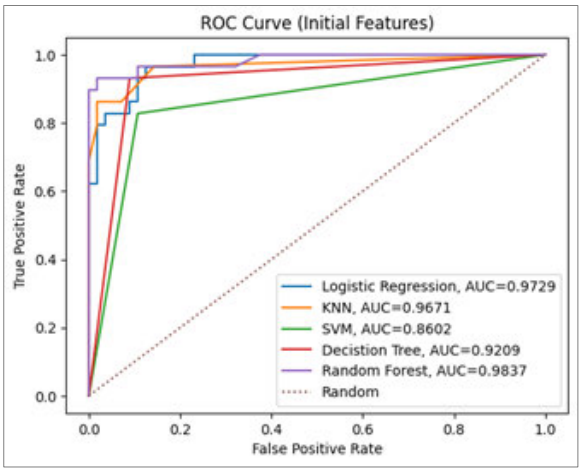

FIGURE 14. ROC Curve for the dataset with initial features.

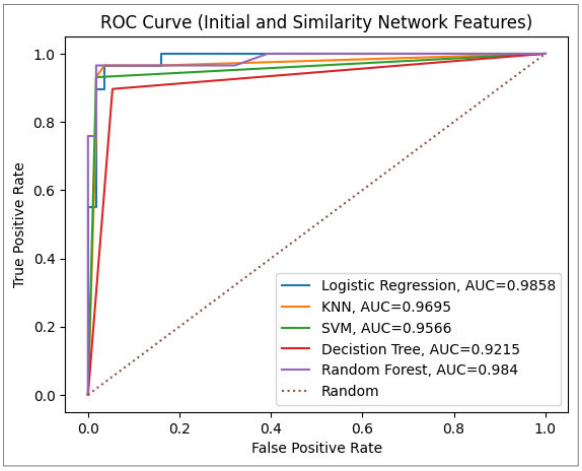

FIGURE 15. ROC Curve for the dataset with initial and similarity-based network features.

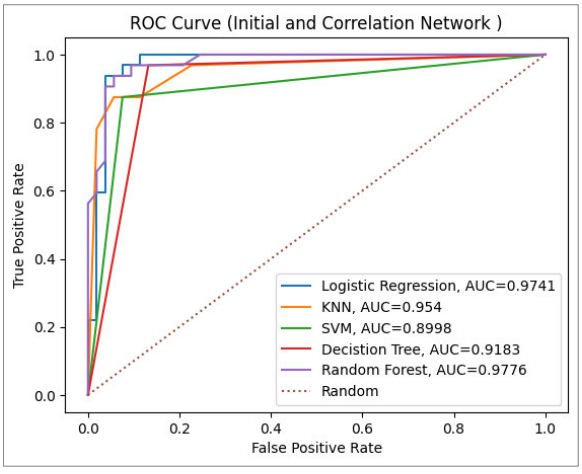

FIGURE 16. ROC Curve for the dataset with initial and correlation-based network features.

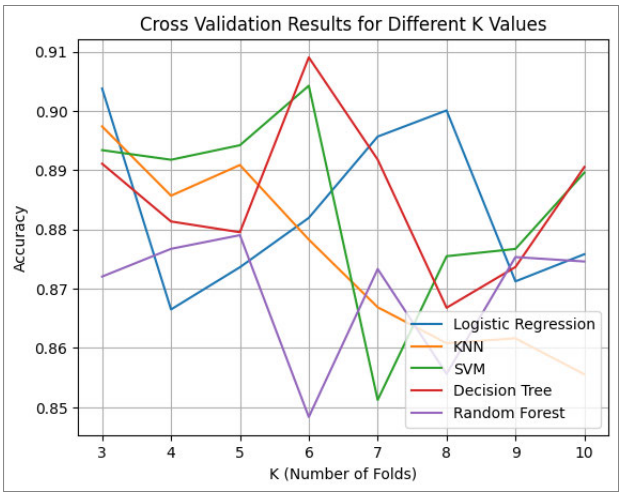

FIGURE 17. K-fold cross validation: first scenario (initial features).

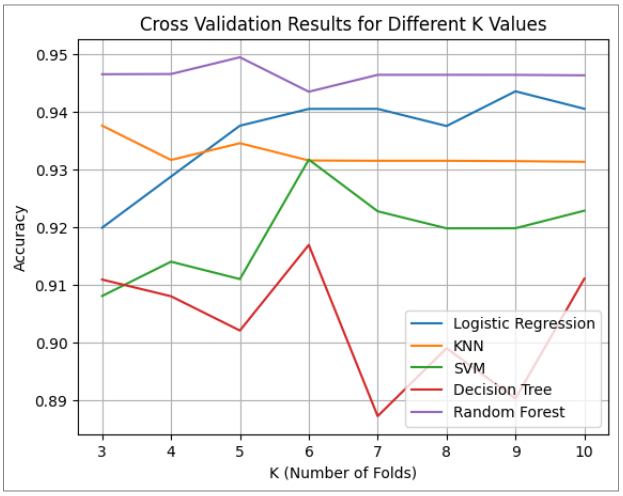

FIGURE 18. K-fold cross validation: second scenario (initial and similarity network features).

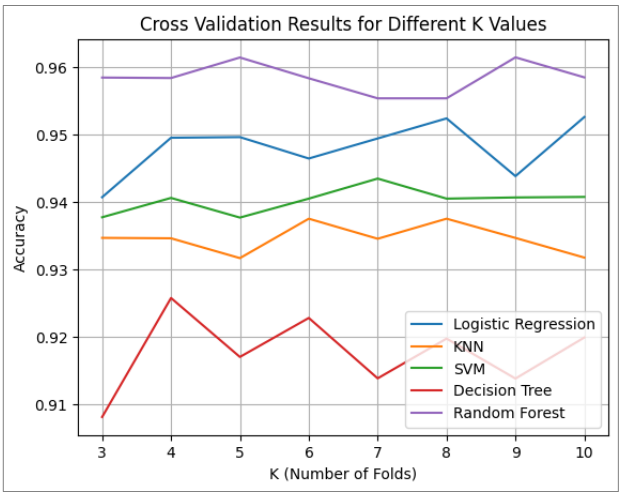

FIGURE 19. K-fold cross validation: third scenario (initial and correlation network features).

TABLE 10. Final results with initial features and correlation network features.

	Model	Accuracy	Precision	Recall	F1-Score
Initial and Correlation Network Features	Logistic Regression	0.91	0.94	0.88	0.93
	KNN	0.86	0.88	0.85	0.89
	SVM	0.88	0.88	0.88	0.91
	Decision Tree	0.89	0.97	0.82	0.91
	Random Forest	0.92	0.94	0.91	0.94

TABLE 11. The average results in three scenarios.

Features	F1-score	Recall	Precision	Accuracy
Initial Features	**0.86**	**0.86**	**0.87**	**0.90**
Initial and Similarity Network	**0.91**	**0.94**	**0.89**	**0.96**
Initial and Correlation Network	**0.89**	**0.92**	**0.87**	**0.92**

Fig. 17-19 illustrate the result of K-fold cross validation of three scenarios. In the first scenario, employing only the initial features resulted in varying accuracy levels ranging from 58% to 91%, demonstrating some inconsistency. However, in the second scenario, where similarity-based network features were added, the accuracy substantially improved, reaching a more reliable range of 89% to 95%. Notably, including features derived from the correlation-based network in the last scenario further enhanced model performance, with accuracy ranging from 91% to 96%. This underscores the significant impact of incorporating network features on the predictive capabilities of the models, with successive improvements observed as additional network-derived features are integrated.

In continuation, we calculated the average accuracies obtained from the five algorithms to demonstrate that adding features extracted from the network contributes to improved predictions. Table 11 presents this analysis's results, indicating that including network-derived features has increased the average accuracy across all metrics. Additionally, features extracted from the similarity network have shown a more significant impact on enhancing prediction accuracy. This emphasizes the substantial role in incorporating network-centric features in financial distress prediction, reinforcing that these features provide valuable information for a more effective and accurate prediction model.

In addition, Fig. 20 illustrates the average performance matrices obtained in each scenario. As evident in the figure, the average performance has increased in all the utilized metrics with the addition of network-based features. Features derived from the similarity-based network significantly impact accuracy improvement, as the chart highlights.

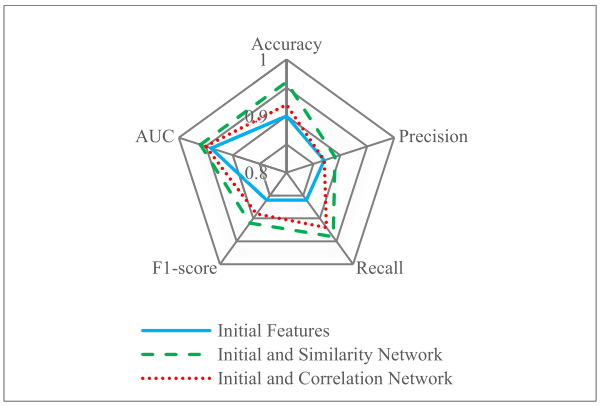

FIGURE 20. Comparison of average results in 3 scenarios.

TABLE 12. Network-based features correlation with target.

Feature	Network Construction Method	
	Similarity	Correlation
Betweenness Centrality	0.042	0.460
Cluster	-0.055	0.154
Page Rank Centrality	-0.045	-0.150
Weighted Clustering Coefficient	0.019	-0.711
Clustering Coefficient	0.019	-0.711
Closeness Centrality	-0.645	-0.721
Average Neighbor Degree	-0.533	-0.819
Degree Centrality	-0.211	-0.833

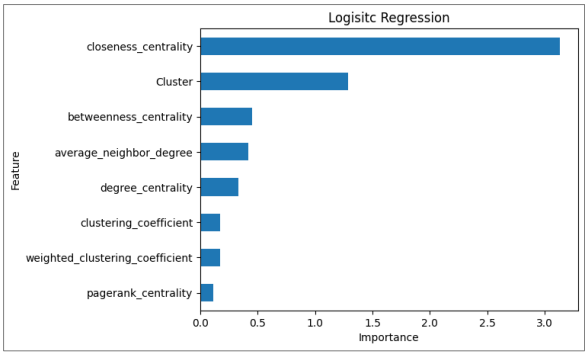

FIGURE 21. Feature importance: Logistic regression.

For further analysis, the correlation coefficients of the network-based features with the target variable have been calculated and reported in Table 12. Analyzing the results in this table contributes to a better understanding of how these features function in predicting financial distress.

The correlations vary across features and between the two networks. Notably, specific centrality measures, like Closeness Centrality, Average Neighbor Degree, and Degree Centrality, exhibit strong negative correlations in both networks, suggesting potential significance in predicting the target variable. Conversely, features like Cluster and PageRank Centrality show weaker correlations and the direction of correlation can differ between the networks. When incorporating these network-based features into predictive models, careful consideration and potential feature selection may be needed.

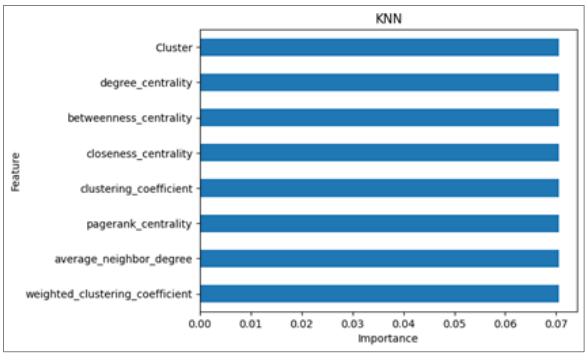

FIGURE 22. Feature importance: KNN.

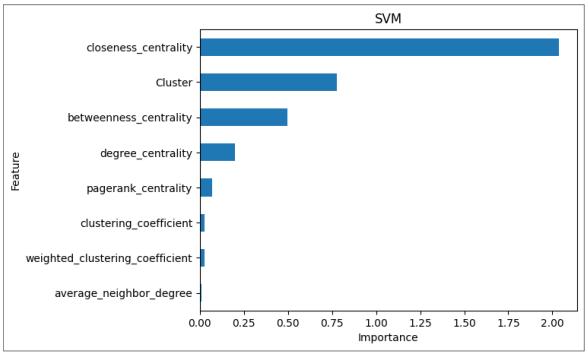

FIGURE 23. Feature importance: SVM.

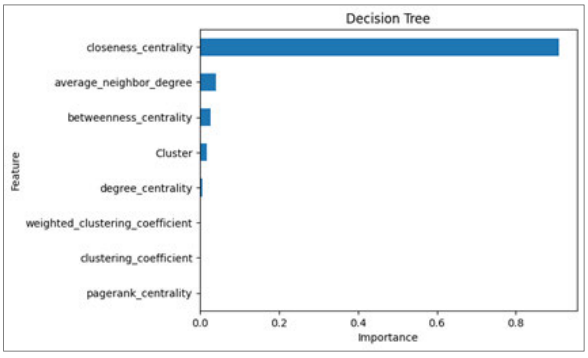

FIGURE 24. Feature importance: Decision tree.

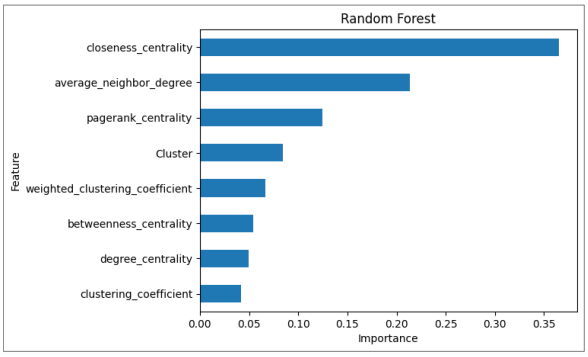

FIGURE 25. Feature importance: Random forest.

TABLE 13. Dataset variables.

Category	Variable
Earnings and Per Share Analysis	Cash Flow Per Share
	Net Value Per Share
	Operating Profit Per Share
	Per Share Net profit before tax
	Persistent EPS in the Last Four Seasons
	Revenue Per Share
	Revenue per person
Efficiency and Turnover Ratios	Accounts Receivable Turnover
	Average Collection Days
	Cash Turnover Rate
	Inventory and accounts receivable/Net value
	Inventory Turnover Rate (times)
	Inventory/Working Capital
	Net Worth Turnover Rate (times)
	Quick Asset Turnover Rate
	Total Asset Turnover
	Working capital Turnover Rate
Financial Structure and Leverage	Borrowing dependency
	Continuous interest rate
	Debt ratio
	Degree of Financial Leverage (DFL)
	Equity to Liability
	Equity to Long-term Liability
	Interest-bearing debt interest rate
	Liability to Equity
	Long-term fund suitability ratio
	Long-term Liability to Current Assets
	Net worth/Assets
	Total assets to GNP price
	Total debt/Total net worth
Growth and Sustainability	After-tax Net Profit Growth Rate
	Cash Reinvestment
	Continuous Net Profit Growth Rate
	Net Value Growth Rate
	Operating Profit Growth Rate
	Realized Sales Gross Profit Growth Rate
	Regular Net Profit Growth Rate
	Total Asset Growth Rate
	Total Asset Return Growth Rate Ratio
Industry and Non-Operating Factors	Allocation rate per person
	Non-industry income and expenditure/revenue
	Geographic Region
Liquidity and Solvency Ratios	Cash/Current Liability
	Cash/Total Assets
	CFO to Assets
	Current Liability to Assets
	Current Liability to Current Assets
	Current Liability to Equity
	Current Ratio
	Inventory/Current Liability
	Quick Assets/Current Liability
	Quick Ratio
	Working Capital to Total Assets
	Working Capital/Equity
Operational Efficiency and Expenses	Cash flow rate
	Cash Flow to Total Assets
	Operating Expense Rate
	Operating Funds to Liability
	Research and development expense rate
	Gross Profit to Sales
Profitability Ratios	Net Income to Total Assets
	Net profit before tax/Paid-in capital
	Operating Gross Margin
	Operating Profit Rate
	Operating profit/Paid-in capital
	Realized Sales Gross Margin
	ROA before interest and % after tax
	ROA before interest and depreciation before interest
	ROA before interest and depreciation after tax
	Operating profit per person
Risk and Coverage Ratios	Cash Flow to Equity
	Cash Flow to Liability
	Cash Flow to Sales
	Contingent liabilities/Net worth
	Continuous interest rate (after tax)
	Interest Coverage Ratio (Interest expense to EBIT)
	Interest Expense Ratio
	Net Income to Stockholder's Equity
	No-credit Interval
	Pre-tax net Interest Rate
	Retained Earnings to Total Assets
	Tax rate
	Total income/Total expense

Due to the strong performance of features related to the similarity-based network, we evaluated the importance of each extracted feature by removing the initial dataset features and exclusively using features derived from the similarity-based network for training the models. The significance of each feature in predicting each model was computed. Fig. 21-25 visualize the results of this analysis, showcasing the importance of each network-derived feature in the prediction process. These figures underscore the significance of these features in influencing the models and emphasizes their role in enhancing the predictive accuracy of financial distress. As evident in the figure, Closeness Centrality is the most crucial feature across all models in predicting financial distress, except for the algorithm KNN, where all features hold equal importance. This observation underscores the consistent significance of the Closeness Centrality in contributing to the predictive capabilities of the models, highlighting its crucial role in the financial distress prediction task.

VI. CONCLUSION

In conclusion, this research introduces a pioneering approach for financial distress prediction, combining network analysis and machine learning methodologies. The study reveals the intricate relationships and interdependencies among financial entities by leveraging company similarity and correlation networks. The derived network features significantly enhance the predictive accuracy of machine learning models, emphasizing the crucial role of features extracted from the similarity

network, exemplified by the Closeness Centrality feature. The proposed model demonstrates impressive predictive performance and offers a comprehensive understanding of the complex dynamics within financial markets. The outcomes highlight the efficacy of network-based strategies in refining financial distress prediction models and provide valuable insights for decision-makers. Future research directions may explore the integration of additional network metrics and including dynamic data further to bolster the resilience and applicability of the proposed model.

APPENDIX

See Table 13.

REFERENCES

[1] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *J. Finance*, vol. 23, no. 4, p. 589, Sep. 1968.
[2] E. B. Deakin, "A discriminant analysis of predictors of business failure," *J. Accounting Res.*, vol. 10, no. 1, pp. 167–179, 1972.
[3] D. Kuizinienė, T. Krilavičius, R. Damaševičius, and R. Maskeličnas, "Systematic review of financial distress identification using artificial intelligence methods," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2138124.
[4] Y. M. Mensah, "An examination of the stationarity of multivariate bankruptcy prediction models: A methodological study," *J. Accounting Res.*, vol. 22, no. 1, pp. 380–395, 1984.
[5] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Syst. Appl.*, vol. 83, pp. 405–417, Oct. 2017.
[6] F. Mai, S. Tian, C. Lee, and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," *Eur. J. Oper. Res.*, vol. 274, no. 2, pp. 743–758, Apr. 2019.
[7] A. I. Dimitras, S. H. Zanakis, and C. Zopounidis, "A survey of business failures with an emphasis on prediction methods and industrial applications," *Eur. J. Oper. Res.*, vol. 90, no. 3, pp. 487–513, May 1996.
[8] S. Udin, M. A. Khan, and A. Y. Javid, "The effects of ownership structure on likelihood of financial distress: An empirical evidence," *Corporate Governance, Int. J. Bus. Soc.*, vol. 17, no. 4, pp. 589–612, Aug. 2017.
[9] H. Waqas and R. Md-Rus, "Predicting financial distress: Applicability of O-score model for Pakistani firms," *Bus. Econ. Horizons*, vol. 14, no. 2, pp. 389–401, 2018.
[10] M. Farooq, A. Noor, and S. F. Qureshi, "The impact of corporate social responsibility on financial distress: Empirical evidence," *Social Responsibility J.*, vol. 18, no. 5, pp. 1050–1067, Jun. 2022.
[11] S. Sehgal, R. K. Mishra, F. Deisting, and R. Vashisht, "On the determinants and prediction of corporate financial distress in India," *Managerial Finance*, vol. 47, no. 10, pp. 1428–1447, Oct. 2021.
[12] S. Ashraf, E. G. S. Félix, and Z. Serrasqueiro, "Do traditional financial distress prediction models predict the early warning signs of financial distress?" *J. Risk Financial Manage.*, vol. 12, no. 2, p. 55, Apr. 2019.
[13] R. Lumbantobing, "The effect of financial ratios on the possibility of financial distress in selected manufacturing companies which listed in Indonesia stock exchange," in *Proc. 6th Annu. Int. Conf. Manage. Res. (AICMaR)*. Atlantis Press, 2020, pp. 60–63.
[14] L. A. Lucky and A. O. Michael, "Leverage and corporate financial distress in nigeria: A panel data analysis," *Asian Finance Banking Rev.*, vol. 3, no. 2, pp. 26–38, Aug. 2019.
[15] Y. Nugrahanti, T. Sutrisno, A. Rahman, and E. Mardiati, "Does diversification strategy reduce the level of financial distress? (evidence from Indonesia)," in *Proc. 1st Sampoerna Univ.-AFBE Int. Conf. (SU-AFBE)*, Jakarta, Indonesia, Dec. 2019, pp. 6–7.
[16] M. S. Veronica, I. Ida, and V. T. Winata, "Using cash flow ratios to establish a manufacturing bankruptcy prediction model," *Jurnal Manajemen Indonesia*, vol. 20, no. 2, pp. 114–121, Aug. 2020.
[17] S. Elviani, R. Simbolon, Z. Riana, F. Khairani, S. P. Dewi, and F. Fauzi, "The accuracy of the Altman, Ohlson, Springate and Zmejewski models in bankruptcy predicting trade sector companies in Indonesia," *Budapest Int. Res. Critics Inst. (BIRCI-Journal)*, vol. 3, no. 1, pp. 334–347, Feb. 2020.
[18] A. Siekelova, T. Kliestik, and P. Adamko, "Predictive ability of chosen bankruptcy models: A case study of Slovak Republic," *Econ. Culture*, vol. 15, no. 1, pp. 105–114, Jun. 2018.
[19] J. Samuel Baixauli and A. Módica-Milo, "The bias of unhealthy SMEs in bankruptcy prediction models," *J. Small Bus. Enterprise Develop.*, vol. 17, no. 1, pp. 60–77, Feb. 2010.
[20] Z.-D. Shen and S. Chen, "Financial distress prediction: A hybrid tracking model approach," *Asian J. Econ., Bus. Accounting*, pp. 185–192, Dec. 2022.
[21] S. Zeng, Y. Li, W. Yang, and Y. Li, "A financial distress prediction model based on sparse algorithm and support vector machine," *Math. Problems Eng.*, vol. 2020, pp. 1–11, Nov. 2020.
[22] N. W. D. Ayuni, N. N. Lasmini, and A. A. Putrawan, "Support vector machine (SVM) as financial distress model prediction in property and real estate companies," in *Proc. Int. Conf. Appl. Sci. Technol. Social Sci. (iCAST-SS)*. Atlantis Press, 2022, pp. 397–402.
[23] S. Chen and Z.-D. Shen, "Financial distress prediction using hybrid machine learning techniques," *Asian J. Econ., Bus. Accounting*, vol. 16, no. 2, pp. 1–12, Jul. 2020.
[24] S. Wu, H. Zhang, Y. Tian, and L. Shi, "Financial distress warning: An evaluation system including ecological efficiency," *Discrete Dyn. Nature Soc.*, vol. 2021, pp. 1–9, Jul. 2021.
[25] M. Elhoseny, N. Metawa, G. Sztano, and I. M. El-hasnony, "Deep learning-based model for financial distress prediction," *Ann. Oper. Res.*, pp. 1–23, May 2022.
[26] J. Huang, H. Wang, and G. Kochenberger, "Distressed Chinese firm prediction with discretized data," *Manage. Decis.*, vol. 55, no. 5, pp. 786–807, Jun. 2017.
[27] L. Ruan and H. Liu, "Financial distress prediction using GA-BP neural network model," *Int. J. Econ. Finance*, vol. 13, no. 3, p. 1, Feb. 2021.
[28] W. Xinli, "Genetic neural network model of forecasting financial distress of listed companies," in *Proc. Int. Conf. Inf. Manage., Innov. Manage. Ind. Eng.*, vol. 1, Nov. 2011, pp. 487–490.
[29] E. Gregova, K. Valaskova, P. Adamko, M. Tumpach, and J. Jaros, "Predicting financial distress of Slovak enterprises: Comparison of selected traditional and learning algorithms methods," *Sustainability*, vol. 12, no. 10, p. 3954, May 2020.
[30] Y. Wu, M. Mahfouz, D. Magazzeni, and M. Veloso, "Towards robust representation of limit orders books for deep learning models," 2021, *arXiv:2110.05479*.
[31] L. A. Africa, "Bankometer models for predicting financial distress in banking industry," *Jurnal Keuangan dan Perbankan*, vol. 22, no. 2, pp. 373–379, Jun. 2018.
[32] L. A. Africa, "Financial distress prediction using RGEC model on foreign exchange banks and non-foreign exchange banks," *J. Accounting Strategic Finance*, vol. 2, no. 1, pp. 48–55, Jun. 2019.
[33] B. Feng, W. Xue, B. Xue, and Z. Liu, "Every corporation owns its image: Corporate credit ratings via convolutional neural networks," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 1578–1583.
[34] H. Xu, G. Fan, and Y. Song, "Novel key indicators selection method of financial fraud prediction model based on machine learning hybrid mode," *Mobile Inf. Syst.*, vol. 2022, pp. 1–12, Mar. 2022.
[35] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Using deep learning to detect price change indications in financial markets," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 2511–2515.
[36] S. Hota, S. K. Jena, B. K. Gupta, and D. Mishra, "An empirical comparative analysis of NAV forecasting using machine learning techniques," in *Proc. Intell. Cloud Comput. (ICICC)*, vol. 2. Springer, 2021, pp. 565–572.
[37] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. London, U.K.: Chapman & Hall, 2019.
[38] G. Kou, X. Chao, Y. Peng, F. E. Alsaadi, and E. H. Viedma, "Machine learning methods for systemic risk analysis in financial sectors," Tech. Rep., 2019.
[39] G. W. M. Zulma, I. Wahyudi, and S. Sutyowati, "Analisis rasio keuangan untuk memprediksi kondisi financial distress (studi empiris pada perusahaan property dan real estate Yang terdaftar di BEI tahun 2018–2020)," *Neraca Keuangan : Jurnal Ilmiah Akuntansi dan Keuangan*, vol. 17, no. 1, pp. 97–104, Apr. 2022.
[40] M. V. Geraci and J.-Y. Gnabo, "Measuring interconnectedness between financial institutions with Bayesian time-varying vector autoregressions," *J. Financial Quant. Anal.*, vol. 53, no. 3, pp. 1371–1390, Jun. 2018.

[41] R. Nekhili, "Systemic risk and interconnectedness in Gulf cooperation council banking systems," *Banks Bank Syst.*, vol. 15, no. 1, pp. 158–166, Mar. 2020.
[42] S. Karyatun, Molina, Elwisam, and K. Wiweka, "Textile and garment sector financial distress and its prediction: A systematic Indonesian literature review," *J. Econ., Manage. Trade*, vol. 28, no. 7, pp. 7–18, Jun. 2022.
[43] T. A. Peltonen, A. Constantin, and P. Sarlin, "Network linkages to predict bank distress," *SSRN J.*, Apr. 2015.
[44] A. Hakim, A. N. M. Salman, Y. Ashari, and K. Syuhada, "Modifying (M)CoVaR and constructing tail risk networks through analytic higher-order moments: Evidence from the global forex markets," *PLoS ONE*, vol. 17, no. 11, Nov. 2022, Art. no. e0277756.
[45] S. Dwirianto, R. Linda, and N. Suryadi, "The effect of financial distress and profitability on auditor change in consumption goods sector companies listed on the Indonesia stock exchange in 2018–2021," *Int. J. Econ. Develop. Res. (IJEDR)*, vol. 4, no. 1, pp. 1–9, Jan. 2023.
[46] F. Aleskerov, I. Andrievskaya, and E. Permjakova, *Key Borrowers Detected by the Intensities of Their Short-Range Interactions*. Springer, 2016.
[47] J.-B. Hasse, "Systemic risk: A network approach," *Empirical Econ.*, vol. 63, no. 1, pp. 313–344, Jul. 2022.
[48] J. Nin, B. Salbanya, P. Fleurquin, E. Tomás, A. Arenas, and J. J. Ramasco, "Modeling financial distress propagation on customer–supplier networks," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 31, no. 5, May 2021.
[49] N. Maximillian and F. Septina, "The effect of profitability, liquidity, and solvency on financial distress of textile and garment companies in Indonesia," *Jurnal Ecodemica, Jurnal Ekonomi Manajemen dan Bisnis*, vol. 6, no. 2, pp. 150–161, Aug. 2022.
[50] M. Mashudi, R. Himmati, I. F. R. Ardillah, and C. Sarasmitha, "Financial distress prediction in infrastructure, utilities, and transportation sector companies 2015–2020," *Jurnal Keuangan dan Perbankan*, vol. 25, no. 3, pp. 656–670, Aug. 2021.
[51] H. Sasongko, A. F. Ilmiyono, A. Tiaranti, and U. Pakuan, "Financial ratios and financial distress in retail trade sector companies," *Jurnal Ilmiah Akuntansi Fakultas Ekonomi*, vol. 7, no. 1, pp. 63–72, Jun. 2021.
[52] M. A. Murtadha, M. Arfan, and M. Saputra, "Factors influencing financial distress and its impact on company values of the sub-sectors firms in Indonesian," *J. Accounting Res., Org. Econ.*, vol. 1, no. 2, pp. 191–204, Dec. 2018.
[53] I. Van Lelyveld and F. R. Liedorp, "Interbank contagion in the Dutch banking sector," Tech. Rep., 2004.
[54] M. Eboli, "Linearities, non-linearities and phase transitions in loss diffusion processes in financial networks," *Nonlinear Phenomena Complex Syst.*, vol. 23, no. 2, pp. 207–211, Jul. 2020.
[55] L. Zhang, Q. Yu, B. Zhou, Y. Zhang, and Z. Hu, "Incorporating feature interactions and contrastive learning for credit prediction," *IEEE Access*, vol. 11, pp. 111944–111955, 2023.
[56] Z. Shi, L. Hong, and C. Zhao, "Discovery and analysis of control associations in shareholding networks based on the voting game approach," *IEEE Access*, vol. 11, pp. 14089–14104, 2023.
[57] T. Chen, J. He, and X. Li, "An evolving network model of credit risk contagion in the financial market," *Technological Econ. Develop. Economy*, vol. 23, no. 1, pp. 22–37, Feb. 2016.
[58] T. Chen, B. Xiao, and H. Liu, "Credit risk contagion in an evolving network model integrating spillover effects and behavioral interventions," *Complexity*, vol. 2018, pp. 1–16, Mar. 2018.
[59] K. Abduraimova and P. Nahai-Williamson, "Solvency distress contagion risk: Network structure, bank heterogeneity and systemic resilience," Tech. Rep., 2021.
[60] Y. Cao, D. Wu, and L. Li, "Debt risk analysis of non-financial corporates using two-tier networks," *Ind. Manage. Data Syst.*, vol. 120, no. 7, pp. 1287–1307, Feb. 2020.
[61] X. Zhang, L. D. Valdez, H. E. Stanley, and L. A. Braunstein, "Modeling risk contagion in the venture capital market: A multilayer network approach," *Complexity*, vol. 2019, pp. 1–11, Dec. 2019.
[62] L. Lu, K. Fang, C. M. Liu, and C. Sun, "The spatial network contagion of environmental risks among countries along the belt and road initiative," *Frontiers Environ. Sci.*, vol. 9, Aug. 2021, Art. no. 721408.
[63] J. L. Bellovary, D. E. Giacomino, and M. D. Akers, "A review of bankruptcy prediction studies: 1930 to present," *J. Financial Educ.*, pp. 1–42, Dec. 1930.
[64] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017.
[65] J. Zhang and Y. Luo, "Degree centrality, betweenness centrality, and closeness centrality in social network," in *Proc. 2nd Int. Conf. Modelling, Simulation Appl. Math. (MSAM)*. Atlantis Press, 2017, pp. 300–303.
[66] L. Newman and A. Dale, "Homophily and agency: Creating effective sustainable development networks," *Environ., Develop. Sustainability*, vol. 9, no. 1, pp. 79–90, Feb. 2007.
[67] F. Grando, D. Noble, and L. C. Lamb, "An analysis of centrality measures for complex and social networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.
[68] B. Liu and B. Liu, "Social network analysis," in *Web Data Mining: Exploring Hyperlinks, Contents, Usage Data*, 2011, pp. 269–309.
[69] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, no. 1, pp. 3–24, 2007.
[70] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *Proc. 4th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2013, pp. 1–7.
[71] H. Dalianis and H. Dalianis, "Evaluation metrics and evaluation," in *Clinical Text Mining: Secondary Use of Electronic Patient Records*, 2018, pp. 45–53.
[72] Ž. Đ. Vujovic, "Classification model evaluation metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 599–606, 2021.

SABA TAHERI KADKHODA received the B.S. degree in industrial engineering from the Khajeh Nasir Toosi University of Technology, Tehran, Iran, and the M.S. degree in information technology from Iran University of Science and Technology. She was a Data Analyst in different companies. Her research interests include data mining and machine learning.

BABAK AMIRI received the Ph.D. degree in information technology from The University of Sydney, in 2014. He is currently an Assistant Professor with Iran University of Science and Technology. He has published extensively in highly regarded scientific journals and has contributed to the advancement of knowledge in complex systems, machine learning, and data science. His research interests include artificial intelligence, big data analytics, and social computing, with a particular focus on the application of complex network theories to various business and engineering problems.