Kardan Journal of Engineering and Technology (KJET)

ISSN: 2706-7815 (Print and Online), Journal homepage: https://kardan.edu.af/KJET

Advancing Cardiovascular Disease Network Analysis: A Comparative Study of Correlation-Based Graph Construction and Link Prediction

Zabihullah Burhani
Abolfazl Dibaji

To cite this article: Z. Burhani and A. Dibaji, "Advancing cardiovascular disease network analysis: a comparative study of correlation-based graph construction and link prediction," *Kardan Journal of Engineering and Technology*, vol. 6, no. 1, pp. 42-57, Dec 2024
DOI: 10.31841/KJET.2024.38

To link to this article: http://dx.doi.org/10.31841/KJET.2024.38

© 2024 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license

Published online: 30 December 2024

Submit your article to this

Article

Advancing Cardiovascular Disease Network Analysis: A Comparative Study of Correlation-Based Graph Construction and Link Prediction

Kardan Journal of Engineering and Technology 6 (1) 42–57
©2024 Kardan University
Kardan Publications
Kabul, Afghanistan
DOI: 10.31841/KJET.2024.38
https://kardan.edu.af/journals/CurrentIssue.aspx?j=KJET

Received: 28 July 24
Revised: 04 Sep 24
Accepted: 09 Dec 24
Published: 30 Dec 24

Zabihullah Burhani
Abolfazl Dibaji

Abstract

Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide, and accurate predictive models are essential for improving prevention and management strategies. This study addresses the challenge of enhancing CVD risk prediction through correlation-based graph construction and weighted link prediction algorithms. Using Pearson and Spearman correlation methods, we transformed a comprehensive dataset containing 1025 patient records and 14 key features into graph structures. Correlation-based graph construction captures feature dependencies by representing variable relationships as edges in a network. To evaluate the effectiveness of the graph representations, we applied weighted link prediction algorithms, including Weighted Common Neighbors (WCN), Weighted Preferential Attachment (WPA), and Weighted Jaccard Coefficient (WJC). The Pearson correlation-based network demonstrated exceptional performance, with the WCN algorithm achieving an Area Under the Curve (AUC) of 99.80% and a Precision of 48.0%. In contrast, the Spearman correlation-based network showed robust results, with WJC achieving an AUC of 96.60% and Precision of 67.16%. The comparative analysis, conducted using Python in a Jupyter environment and employing libraries such as NetworkX and various statistical libraries, highlights the superior ability of correlation-based graphs to capture linear and non-linear relationships in CVD data. While promising, the study acknowledges limitations related to dataset size and computational complexity. Our findings suggest that correlation-based graph methods significantly enhance CVD prediction, offering a more personalized CVD prevention and management approach.

Keywords: *Cardiovascular Disease, Link Prediction, Network Analysis, Graph*

1. Introduction

Cardiovascular diseases (CVDs) represent a significant global health challenge, remaining the leading cause of mortality worldwide [1]. These diseases, which affect the heart and blood vessels, encompass a wide range of conditions, including coronary heart disease, cerebrovascular disease, and peripheral arterial disease. The World Health Organization reports that CVDs are responsible for an estimated 31% of all deaths globally, with a disproportionate impact on low- and middle-income countries.

The complex nature of CVDs, involving multiple risk factors and intricate physiological mechanisms, necessitates advanced analytical approaches for better understanding, prediction, and management. Traditional risk assessment tools, while valuable, often fall short of capturing the full complexity of cardiovascular health. This limitation has spurred research into novel methodologies that can provide more comprehensive insights into the interplay of various factors contributing to CVDs [2].

Recent advancements in data science and network analysis offer promising avenues for enhancing our understanding of CVDs. Graph theory, in particular, provides a powerful framework for representing and analyzing complex relationships within medical data. By representing patient data as nodes and their relationships as edges in a graph, it becomes possible to uncover hidden patterns and associations that may not be apparent through conventional statistical methods [29].

This study proposes a novel approach to cardiovascular disease analysis that leverages correlation-based graph construction and weighted link prediction algorithms. Our method aims to transform cardiovascular disease data into graph structures, enabling the application of advanced network analysis techniques. By employing different correlation measures and combining their results, we seek to capture a more nuanced representation of the relationships within the data.

The core of our approach involves constructing multiple graph representations of cardiovascular disease data and applying weighted link prediction algorithms to these graphs. This methodology allows us to explore the predictive power of various graph structures and prediction algorithms in the context of CVDs. We aim to identify which combinations of graph construction methods and link prediction algorithms yield the most insightful and accurate results.

By advancing the application of network analysis in cardiovascular diseases, this research aims to contribute to developing more sophisticated tools for disease prediction, risk assessment, and relationship prediction. The insights gained from this study could potentially inform the creation of more personalized and effective CVD prevention and management strategies.

In the following sections, we will detail our methodology, present the results of our analysis, and discuss the implications of our findings. We will also address the limitations of our approach and suggest potential avenues for future research in this critical area of health informatics.

2. Related work

Recent years have witnessed significant advancements in cardiovascular disease (CVD) prediction and analysis, with researchers employing various computational methods to enhance accuracy and efficiency. This section critically examines recent studies in this field, focusing on machine-learning approaches, feature selection techniques, and network-based methods.

2.1 Machine Learning Approaches

Machine learning algorithms have demonstrated considerable promise in CVD prediction. Convolutional neural networks (CNNs) have shown high accuracy, as evidenced by Mehmood et al.'s (2021) CardioHelp method, which achieved an accuracy of 97% [3]. While this result is impressive, the study's reliance on a single dataset may limit its generalizability.

Decision tree and random forest algorithms have consistently performed well across multiple studies. Rahman et al. (2022) developed a web-based heart disease prediction system utilizing thirteen health parameters and eight algorithms, with decision tree and random forest algorithms achieving an accuracy of 99% [5]. Similarly, Khan et al. (2023) found the random forest algorithm to be the most accurate in their machine learning-based model for CVD prediction [6]. However, these studies' high accuracy rates warrant cautious interpretation, as they may indicate potential overfitting issues.

Ensemble methods have shown the potential to improve prediction accuracy. Dritsas and Trigka (2023) demonstrated that the Stacking ensemble model performed better after applying the Synthetic Minority Over-Sampling Technique (SMOTE) [9]. This approach addresses class imbalance issues common in medical datasets, but its effectiveness may vary depending on the specific dataset characteristics. The study by Alfaidi et al. (2022) explores using ML for CVD diagnosis. Seven algorithms were tested on a cardiovascular dataset, with Chi-square tests identifying key predictive features. The Multi-Layer Perceptron achieved the highest accuracy (87.23%), indicating the promise of AI in assisting early diagnosis of heart disease. Baghdadi et al. (2023) proposed using the Catboost model, which outperformed existing methods with an average accuracy of 90.9% [7]. While promising, this study's results highlight the ongoing challenge of achieving consistently high accuracy across diverse datasets and populations.

2.2 Feature Selection and Data Analysis

Identifying key features for CVD prediction has been the focus of several studies, aiming to improve model efficiency and interpretability. Guernaros-Nolasco et al. (2021) analyzed ten machine-learning algorithms to identify the most predictive features for CVDs [4]. However, their study did not provide a comprehensive ranking of feature importance across all algorithms, which could have offered valuable insights.

Mahmoud et al. (2021) developed a model combining genetic algorithms and recursive feature elimination for feature selection, demonstrating excellent accuracy and performance [8]. This hybrid approach shows promise in optimizing feature selection, but its computational complexity may limit its applicability in real-time clinical settings.

Tallin et al. (2022) identified eight key clinical features for heart disease diagnosis, including chest pain and the number of major blood vessels [14]. While this study provides a concise set of predictive features, it may oversimplify the complex nature of CVD by focusing on a limited number of factors.

Tanyildizi Kokkulonk (2023) used multiple linear regression for heart disease classification, achieving an accuracy of 88% and noting the minimal impact of age data on predictions [11]. This finding contradicts some previous studies and warrants further investigation into the role of age in CVD prediction across different populations.

2.3 Network-based and Graph Theory Approaches

Recent research has explored the potential of network-based methods and graph theory in CVD analysis, offering new perspectives on disease relationships and comorbidities.

Garcia del Valle et al. (2021) introduced a Metapath-based method for predicting disease comorbidities, which demonstrated higher accuracy compared to traditional methods by utilizing clinical data and heterogeneous networks [13]. This approach shows promise in capturing complex disease interactions, but its computational requirements may pose challenges for large-scale implementations.

Wang and Qiu (2019) proposed a framework for predicting multiple disease risks using directed disease networks and recommendation systems, validating their results with real data from two hospitals [15]. While innovative, this study's reliance on data from only two hospitals may limit its generalizability to diverse healthcare settings.

Lu and Edin (2022) developed a framework for predicting chronic diseases and comorbidities, finding the matrix completion of the aggregate graph to be the best-performing model [16]. This approach offers a novel way to handle missing data in medical records, but its effectiveness may vary depending on the sparsity of the input data.

Wang et al. (2020) used Deep Graph Convolutional Networks (GCNs) to analyze and predict comorbidities in health records, modelling conditions and features as graph nodes [19]. While this method shows potential in capturing complex relationships between diseases, its interpretability remains a challenge, which is crucial in clinical applications.

Dibaji and Sulaimany (2023) analyzed a graph-based model with 1190 samples, achieving an accuracy of 95% and suggested that examining advanced network features could lead to further improvements [2]. This study highlights the potential of graph-based approaches in CVD analysis and prediction, but the relatively small sample size may limit its statistical power.

2.4 Comparative Studies

Several studies have compared different methodologies, providing insights into the relative strengths of various approaches. Zariqat et al. (2016) compared five data mining classification methods and found that the decision tree with an accuracy of 0.99 performed best [20]. However, this study's focus on a single metric (accuracy) may not comprehensively evaluate model performance, especially in imbalanced datasets common in medical research.

A study on feature selection techniques and hybrid classifiers showed that the Random Forest Bagging method with Relief feature selection achieved a high accuracy of 99.09% [10]. While this result is impressive, the study did not extensively explore the trade-offs between model complexity and performance, which is crucial for practical implementation.

In conclusion, while these studies have made significant contributions to CVD prediction and analysis, there remains a need for more comprehensive approaches that can capture the complex relationships within cardiovascular health data. Many studies achieve high accuracy rates, but issues such as potential overfitting, limited generalizability, and challenges in interpretability persist. Our study aims to address these gaps by combining correlation-based graph construction with weighted link prediction algorithms, offering a novel perspective on CVD network analysis that balances accuracy, interpretability, and generalizability.

Table 1
Summary of Previous Research Conducted on Heart Diseases

Ref.	Year	Journal and Conference	Method	Evaluation Methods	Best Result
[6]	2023	Health & Social Care in the Community Article Information	DT, RF LR, NB & SVM	CM & ROC Curve	92.11%
[7]	2023	Journal of Big Data	Catboost & Gradiant Boosting Models	F1-Score & Average Accuracy Precision	93% & 90.94%
[2]	2023	13th International Conference on Computer and Knowledge Engineering (ICCKE 2023), Ferdowsi University of Mashhad, Iran	GBoost, NN, SVM & RF	Accuracy, AUC, Precision & Recall	98.7%
[9]	2023	Sensors	NB, LR, MLP, 3NN, RF, RotF, AdaBoostM1, Stacking, Bagging, Voting	Accuracy, Precision, Recall, AUC	98.2%
[11]	2023	Physical Science & Biophysics Journal	Classification & RT, HRFLM Classification, NB & MLR	Accuracy, Sensitivity, Specificity, Precision	88%
[5]	2022	Network Biology	KNN, XGBOOST, LR, SVM, AdaBOOST, DT, NB, RF	Accuracy & Precision	99% & 97%
[3]	2021	Arabian Journal for Science and Engineering	CNN	Precision, Recall, F1 & Accuracy	97.06%
[4]	2021	Mathematics	AdaBoost, CatBoost, DT, GBoosting, KNN, LGBM, LR, RF, SVM & XGBRF	Accuracy, Precision, Recall, F1 & ROC AUC	80.25%
[18]	2021	Webology	RBN, MIP, GAN, CNN, DBN, Linear GAN	Accuracy, F1, Recall	92.23%
[10]	2021	IEEE Access	DT, RF, KNN, AB, GB, DTBM, RFBM, KNNBM, ABBM & GBBM	Accuracy	99.05%
[17]	2020	European Journal of Heart Failure	Boosted DT	AUC	88%

3. Data and Methodology

Our study introduces a novel cardiovascular disease (CVD) analysis approach by combining correlation-based graph construction with weighted link prediction algorithms. This methodology captures complex relationships within CVD data and enhances predictive capabilities. The proposed approach consists of the following key steps:

3.1 Dataset Preparation

We utilized a heart disease dataset as the foundation for our analysis. This dataset contained various features related to cardiovascular health for multiple patients. The Cardiovascular Diseases (CVDs) datasets, collected since 1988, serve as a valuable resource for medical analysis and machine learning applications. These datasets, sourced from Cleveland, Hungary, Switzerland, and Long Beach V, encompass 76 attributes, though most research focuses on 14 key features. The primary aim is to predict CVD presence in patients using a binary "target" field (1 for presence, 0 for absence). Patient identifiers have been removed to ensure privacy. This widely used dataset has been instrumental in developing predictive models for heart disease, as evidenced by numerous studies summarized in the literature [21].

Table 2
CVD Dataset Information

Index	Attribute Name	Attribute Information
[1]	age	Age of the patient in years.
[2]	sex	Represented as a binary number. 1 = male, 0 = female.
[3]	cp	Chest pain type. Values range from 1 to 4. Value 1: typical angina. Value 2: atypical angina. Value 3: non-anginal pain. Value 4: asymptomatic
[4]	trestbps	Resting blood pressure was measured in mm Hg upon admission to the hospital.
[5]	chol	Serum cholesterol of the patient was measured in mg/dl.
[6]	fbs	Fasting blood sugar of the patient. If greater than 120 mg/dl, the attribute value is 1 (true); else, the attribute value is 0 (false). Value 1 = true. Value 0 = false.
[7]	restecg	Resting electrocardiographic results for the patient. This attribute can take 3 integer values: 0, 1, or 2. Value 0: normal. Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV). Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.
[8]	thalach	Maximum heart rate achieved of the patient.
[9]	exang	Exercise-induced angina. Values can be 0 or 1. Value 1 = yes. Value 0 = no.
[10]	oldpeak	ST depression induced by exercise relative to rest.
[11]	slope	Measure of slope for peak exercise. Values can be 1, 2, or 3. Value 1: up-sloping. Value 2: flat. Value 3: downsloping.
[12]	ca	Number of major vessels (0-3) coloured by fluoroscopy. Attribute values can be 0 to 3.
[13]	thal	Represents the heart rate of the patient. It can take values 3, 6, or 7. Value 3 = normal. Value 6 = fixed defect. Value 7 = reversable defect.
[14]	target	Contains a numeric value between 0 and 4. Each value represents a heart disease or absence of all of them. Value 0: < 50% diameter narrowing. (Absence of heart disease). Value 1 to 4: > 50% diameter narrowing. (Presence of different heart diseases).

3.2 Correlation Analysis and Graph Construction

We employed two correlation methods to analyze relationships within the dataset:

- **Pearson Correlation:**

$$r = \frac{\Sigma\left((X_i - \overline{X})\,(Y_i - \overline{Y})\right)}{\sqrt{(X_i - \overline{X})^2}\ .\ \sqrt{(Y_i - \overline{Y})^2}} \qquad (1)$$

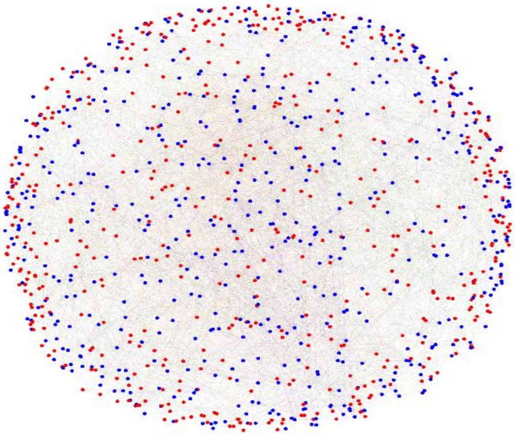

Fig. 1. Network created by Pearson Correlation

We calculated the Pearson correlation coefficient between pairs of samples in the dataset. This measure quantifies the linear relationship between variables, ranging from -1 to +1 [27].

- **Spearman Correlation:**

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (2)$$

Similarly, we computed the Spearman correlation coefficient, which captures monotonic relationships between variables, including non-linear associations [28].

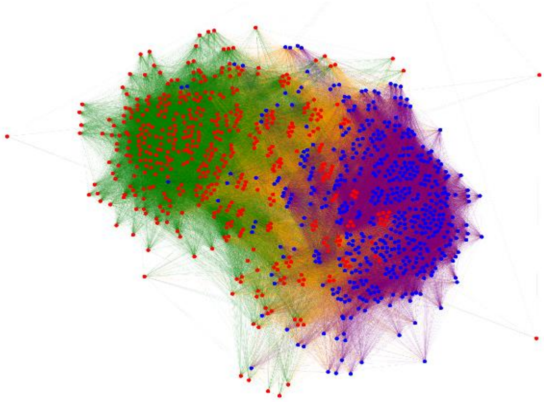

Fig. 2. Network created by Spearman correlation

3.3 Graph Generation

For each method, we constructed a graph where:

- Each line in the dataset corresponds to a node in the graph.
- Nodes are classified as either 0 (healthy individuals) or 1 (individuals with cardiovascular disease).
- Edges between nodes are established based on a dynamically determined threshold limit.
- Three types of edges can exist: 0-0 (between healthy individuals), 1-1 (between individuals with CVD), and 0-1 (between a healthy individual and one with CVD).

Threshold Determination Process:

- We start with a low threshold value and incrementally increase it.
- At each threshold level, we calculate the number of edges for each type (0-0, 1-1, and 0-1).
- We continue increasing the threshold until we reach a point where the number of 0-0 edges and 1-1 edges are greater than the number of 0-1 edges.
- This point is then set as the optimal threshold for edge creation in the graph.

Rationale for this approach:

1. **Dynamic Adaptation:** This method allows the threshold to adapt to the specific characteristics of each dataset and correlation method, ensuring optimal graph construction.
2. **Improved Intra-Group Correlation:** By ensuring that 0-0 and 1-1 edges outnumber 0-1 edges, we strengthen the connections within each group (healthy and CVD), which is crucial for meaningful network analysis.
3. **Enhanced Discriminative Power:** This approach helps create a network structure that distinguishes healthy individuals and those with CVD, potentially improving the predictive power of subsequent analyses.
4. **Reduction of Noise:** By minimizing the relative number of 0-1 edges, we reduce potential noise in the network, focusing on the most significant correlations within each group.
5. **Balancing Connectivity and Specificity:** This method strikes a balance between maintaining sufficient network connectivity and ensuring specificity in the relationships represented by the edges.

3.4 Combined Graph

$$combined\ weighted\ network\ =\ [(x, y, spearman\_corr\ +\ pearson\_corr)] \tag{3}$$

We developed a Network-based Correlation Integration (NCI) method to combine the Pearson and Spearman graphs, resulting in a third, unified graph. This combined graph aims to leverage the strengths of both correlation techniques.

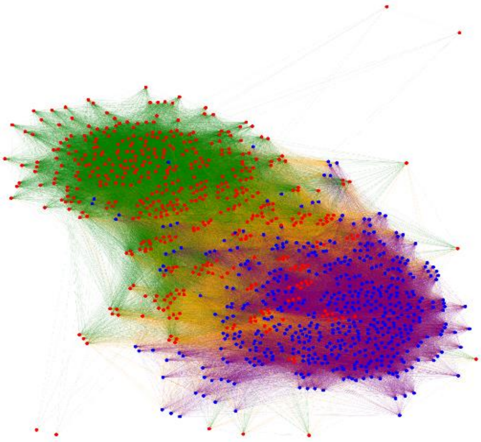

Fig. 3. The combined network created (by Spearman correlation and Pearson correlation)

Table 3
Graphs Characteristics

Network	Nodes	Edges	0-0	1-1	0-1	Threshold	Density
Pearson	1025	4286	1339	1745	1202	0.9997	0.00816
Spearman	1025	189551	63142	82920	43489	0.8	0. 362
Combined	1025	191063	62885	83444	44734	1.78	0. 364

Source: Created by the authors

3.5 Weighted Link Prediction

We applied several weighted link prediction algorithms to all three graphs (Pearson-based, Spearman-based, and combined). These algorithms included [22-25]:

- Weighted Common Neighbors (WCN)

$$\mathrm{WCN(x, y)}\ =\ \textstyle\sum_{\mathrm{z}\in|\Gamma(\mathrm{x})\cap\Gamma(\mathrm{y})|} \mathrm{W(x, z) + W(y, z)} \tag{3}$$

- Weighted Preferential Attachment (WPA)

$$WPA(x, y)\ =\ \textstyle\sum_{z\in|\Gamma(x)\cap\Gamma(y)|} W(x, z) * W(y, z) \tag{3}$$

- Weighted Jaccard Coefficient (WJC)

$$WJC(x, y)\ =\ \frac{\sum_{z\in|\Gamma(x)\cap\Gamma(y)|} W(x,z)+W(y,z)}{\sum_{z1\in|\Gamma(x)|} W(x,z1)+\sum_{z2\in|\Gamma(y)|} W(y,z2)} \tag{3}$$

3.6 Performance Evaluation

To assess the effectiveness of our approach, we calculated two key metrics for each graph and prediction algorithm combination [26]:

- Area Under the Curve (AUC): Measures the overall performance of the prediction model.

- Precision (Pre): Evaluates the accuracy of the predicted links.

This methodology allows us to compare the efficacy of different correlation-based graph constructions and weighted link prediction algorithms in the context of cardiovascular disease analysis. By applying these methods to patient data, we aim to uncover hidden patterns and relationships that could enhance our understanding and prediction of cardiovascular diseases.

4. Results and Discussion

4.1 Performance of Weighted Link Prediction Algorithms

Table 4.
Results of Evaluation Metrics on Pearson Network

Algorithms (WLP)	AUC	Precision
WCN	99.8%	48.0%
WPA	70.0%	18.6%
WJC	99.7%	73.6%

Source: Created by the authors

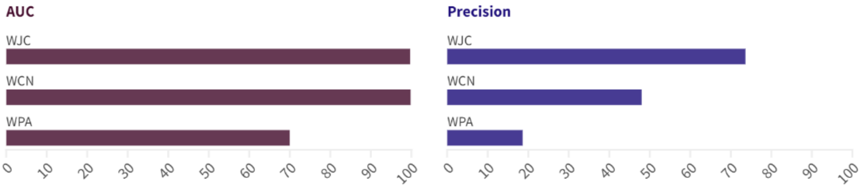

Fig. 4. Bar chart to compare AUC and Precision results obtained from link prediction in Pearson network

Table 5
Results of Evaluation Metrics on Spearman Network

Algorithms (WLP)	AUC	Precision
WCN	95.1%	59.74%
WPA	70.0%	25.35%
WJC	96.6%	67.16%

Source: Created by the authors

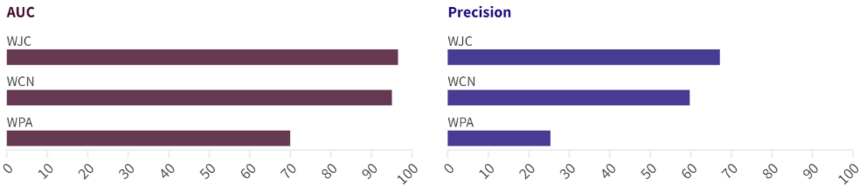

Fig. 5. Bar chart to compare AUC and Precision results obtained from link prediction in Spearman network

Table 6
Results of Evaluation Metrics on Combined Network

Algorithms (WLP)	AUC	Precision
WCN	94.7%	56.69%
WPA	70.2%	25.41%
WJC	96.2%	65.22%

Source: Created by the authors

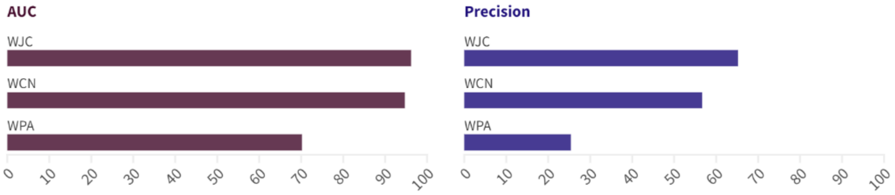

Fig. 6. Bar chart to compare AUC and Precision results obtained from link prediction in Combined network

The results of our study demonstrate the effectiveness of correlation-based network analysis and weighted link prediction algorithms in cardiovascular disease analysis. Tables 4, 5, and 6 present each algorithm's performance metrics (AUC and Precision) across the three networks: Pearson, Spearman, and Combined.

- **Pearson Network Results:**

  The Pearson correlation-based network showed strong performance across all algorithms (Table 4). The Weighted Common Neighbors (WCN) algorithm achieved the highest AUC of 99.80%, while the Weighted Jaccard Coefficient (WJC) demonstrated the best Precision at 73.60%. These results suggest that the Pearson network effectively captured linear relationships in the CVD data.

- **Spearman Network Results:**

  The Spearman correlation-based network also showed robust performance (Table 5). The WJC algorithm performed best on this network, with an AUC of 96.60% and Precision of 67.16%. The strong performance of the Spearman network indicates its ability to capture non-linear relationships in the data.

- **Combined Network Results:**

  The combined network, integrating both Pearson and Spearman correlations, showed comparable performance to the individual networks (Table 6). The WJC algorithm again performed best, with an AUC of 96.20% and Precision of 65.22%. These results suggest that the combined approach offers a balanced representation of the data's linear and non-linear relationships.

4.2 Comparative Analysis

Our network-based approach, particularly using the Pearson and Spearman networks, demonstrated competitive performance compared to existing methods in the literature. The high AUC values (>95%) across all networks indicate excellent discriminative power in predicting CVD links (Figure 7).

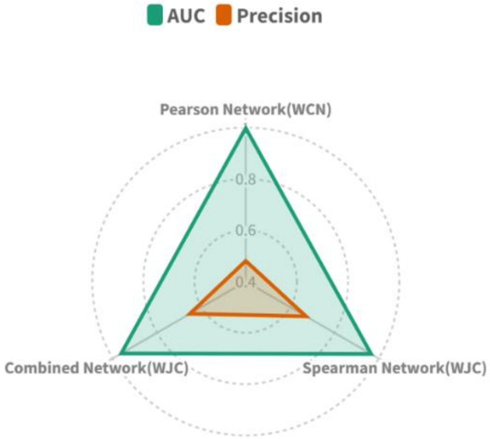

Fig. 7. Comparison of the highest results of weighted link prediction in each network

4.3 Network Characteristics and Performance

The performance differences between networks can be attributed to their structural properties. The Pearson network's high performance may be due to its ability to capture strong linear correlations in the CVD data. The Spearman network's robust performance suggests it effectively represents non-linear relationships.

4.4 Advantages of the Network-Based Approach

Our model offers several advantages over traditional methods:

- **Feature Dependency:** The model accounts for complex interactions between features.

- **Non-linear Relationships:** The approach can capture both linear and non-linear relationships within the CVD data.

- **High Accuracy:** The method achieved high AUC values across all networks, indicating strong predictive power.

4.5 Limitations and Future Directions

Despite the promising results, our study has limitations that warrant further investigation:

- **Dataset Size:** The sample size of 1,025 cases may limit generalizability. Future studies should validate these methods on larger, more diverse datasets.

- **Computational Complexity:** Optimizing these algorithms for practical implementation in clinical settings is an important area for future research.

- **Edge Threshold:** Further investigation into optimal threshold selection for network construction is needed.

- **Clinical Application:** Addressing challenges such as incomplete data and the need for continuous model updates is crucial for real-world implementation.

- **Algorithm Limitations:** Due to computational constraints, we were unable to test the WAA and WRA algorithms. Future studies with access to high-performance computing resources should also evaluate these algorithms.

Our study demonstrates the potential of correlation-based network analysis and weighted link prediction algorithms in cardiovascular disease link prediction. The strong performance across network types suggests that this approach could significantly contribute to CVD diagnosis and treatment advances. However, further research is necessary to address the limitations and validate these findings across diverse datasets and clinical settings.

Conclusion

Our research into correlation-based graph construction and weighted link prediction algorithms for cardiovascular disease analysis has yielded promising results, potentially revolutionizing the field of CVD prediction and management. The exceptional performance of the Pearson correlation-based network, particularly when coupled with the Weighted Common Neighbor algorithm, demonstrates the power of network analysis in capturing complex relationships within medical data.

This approach offers several advantages over traditional methods, including the ability to account for feature dependencies and non-linear relationships, resulting in remarkably high accuracy. The success of our model in predicting links within the CVD dataset suggests that it could be a valuable tool for identifying hidden patterns and associations in patient data, potentially leading to earlier and more accurate diagnoses.

However, it is crucial to acknowledge the limitations of our study, including the relatively small dataset size and computational complexity of graph-based methods. These challenges present opportunities for future research, particularly in optimizing algorithms for real-time clinical applications and validating results across larger, more diverse datasets.

As we look to the future, the potential applications of this methodology extend beyond cardiovascular diseases. This network-based approach could be adapted to analyze other complex medical conditions, contributing to the broader field of precision medicine. By continuing to refine and expand upon this work, we can move closer to developing more sophisticated, personalized approaches to disease prediction, prevention, and treatment.

In conclusion, while further research and validation are necessary, our study represents a significant step forward in the application of advanced data analysis techniques to cardiovascular health. As we continue to bridge the gap between data science and medical research, we open new avenues for improving patient outcomes and advancing our understanding of complex diseases.

Data Availability

The dataset utilized in this article is accessible on the Kaggle platform and is freely available to the public. It can be accessed via the following link: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," Accessed: Jul. 10, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] A. Dibaji and S. Sulaimany, "Improving machine learning classification of heart disease using the graph-based techniques," in *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Islamic Republic of Iran: IEEE, 2023, pp. 474–479. doi: 10.1109/ICCKE60553.2023.10326233.

[3] A. Mehmood, M. Iqbal, Z. Mehmood, A. Irtaza, M. Nawaz, T. Nazir, and M. Masood, "Prediction of heart disease using deep convolutional neural networks," *Arab. J. Sci. Eng.*., vol. 46, no. 4, pp. 3409–3422, Apr. 2021, doi: 10.1007/s13369-020-05105-1.

[4] L. R. Guarneros-Nolasco, N. A. Cruz-Ramos, G. Alor-Hernández, L. Rodríguez-Mazahua, and J. L. Sánchez-Cervantes, "Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms," *Math.*., vol. 9, no. 20, p. 2537, Oct. 2021, doi: 10.3390/math9202537.

[5] M. Mahbubur Rahman, M. R. Rana, M. Nur-A-Alam, M. S. I. Khan, and K. M. M. Uddin, "A web-based heart disease prediction system using machine learning algorithms," *Netw. Biol.*., vol. 12, no. 2, pp. 64–80, 2022. [Online]. Available: http://www.iaees.org/publications/journals/nb/online-version.asp

[6] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," *Health Soc. Care Community.*, 2023. Accessed: Jun. 2024. [Online]. doi: 10.1155/2023/1406060.

[7] N. A. Baghdadi, S. M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis," *J. Big Data*, vol. 10, no. 1, p. 144, Sep. 2023, doi: 10.1186/s40537-023-00817-1.

[8] P. Rani, R. Kumar, A. Jain, and S. K. Chawla, "A hybrid approach for feature selection based on genetic algorithm and recursive feature elimination," *Int. J. Inf. Syst. Model. Des..*, vol. 12, no. 2, pp. 1–22, 2021, doi: 10.4018/IJISMD.2021040102.

[9] E. Dritsas and M. Trigka, "Efficient data-driven machine learning models for cardiovascular diseases risk prediction," *Sen..*, vol. 23, no. 3, p. 1161, Jan. 2023, doi: 10.3390/s23031161.

[10] P. Ghosh et al., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[11] T. K. H, "Prediction of heart disease using machine learning with data mining," *Phys. Sci. Biophys.*, vol. 7, no. 1, pp. 1–6, Jan. 2023, doi: 10.23880/psbj-16000228.

[12] A. Alfaidi, R. Aljuhani, B. Alshehri, H. Alwadei, and S. Sabbeh, "Machine learning-assisted cardiovascular diseases diagnosis," *Int. J. Adv. Comput. Sci. Appl..*, vol. 13, no. 2, 2022, doi: 10.14569/IJACSA.2022.0130216.

[13] E. P. G. Del Valle, L. P. Santamaria, G. L. Garcia, M. Zanin, E. M. Ruiz, and A. Rodriguez-Gonzalez, "A meta-path-based prediction method for disease comorbidities," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, Aveiro, Portugal: IEEE, Jun. 2021, pp. 219–224, doi: 10.1109/CBMS52027.2021.00022.

[14] I. A. Talin, M. H. Abid, Md. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Finding the influential clinical traits that impact on the diagnosis of heart disease using statistical and machine-learning techniques," *Sci. Rep..*, vol. 12, Article no. 20199, 2022, doi: 10.1038/s41598-022-24633-4.

[15] T. Wang, R. G. Qiu, M. Yu, and R. Zhang, "Directed disease networks to facilitate multiple-disease risk assessment modeling," *Decis. Support Syst..*, vol. 129, p. 113171, Feb. 2020, doi: 10.1016/j.dss.2019.113171.

[16] H. Lu and S. Uddin, "A disease network-based recommender system framework for predictive risk modelling of chronic diseases and their comorbidities," *Appl. Intell..*, vol. 52, no. 9, pp. 10330–10340, Jul. 2022, doi: 10.1007/s10489-021-02963-6.

[17] E. D. Adler *et al.*, "Improving risk prediction in heart failure using machine learning," *Eur. J. Heart Fail.*, vol. 22, no. 1, pp. 139–147, Jan. 2020, doi: 10.1002/ejhf.1628.

[18] M. A. Kumar, K. C. Purohit, A. Singh, and S. Bhatt, "Heart disease prediction model using deep learning algorithms," *Webology*, vol. 18, no. 4, 2021, doi: 10.29121/WEB/V18I4/106.

[19] R. Wang, M.-C. Chang, and M. Radigan, "Modeling latent comorbidity for health risk prediction using graph convolutional network," in *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS-33)*, Association for the Advancement of Artificial Intelligence (AAAI), 2020.

[20] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," *arXiv:1704.02799*, Apr. 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1704.02799

[21] Kaggle, "Heart disease dataset," Accessed: Jul. 10, 2024. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[22] B. Moradabadi and M. R. Meybodi, "Link prediction in weighted social networks using learning automata," *Eng. Appl. Artif. Intell..*, vol. 70, pp. 16–24, 2018, doi: 10.1016/j.engappai.2017.12.006.

[23] B. Liu, S. Xu, T. Li, J. Xiao, and X.-K. Xu, “Quantifying the effects of topology and weight for link prediction in weighted complex networks,” *Entropy*, vol. 20, no. 5, Art. no. 5, May 2018, doi: 10.3390/e20050363.

[24] A. Kumar, S. S. Singh, K. Singh, and B. Biswas, “Link prediction techniques, applications, and performance: A survey,” *Physica A Stat. Mech. Appl..*, vol. 553, p. 124289, Sep. 2020, doi: 10.1016/j.physa.2020.124289.

[25] M. Liu, Y. Wang, J. Chen, and Y. Zhang, “Link prediction model for weighted networks based on evidence theory and the influence of common neighbours,” *Complexity*, vol. 2022, p. e9151340, Mar. 2022, doi: 10.1155/2022/9151340.

[26] A. Kumar, S. S. Singh, and S. Mishra, “Empirical analysis of unsupervised link prediction algorithms in weighted networks,” in *oft Comput..*, vol. 627, Lecture Notes in Networks and Systems, Singapore: Springer Nature Singapore, 2023, pp. 173–183, doi: 10.1007/978-981-19-9858-4_15.

[27] R. R. Wilcox, “Modern insights about Pearson’s correlation and least squares regression,” *Int. J. Select. Assess..*, vol. 9, no. 1–2, pp. 195–205, 2001, doi: 10.1111/1468-2389.00172.

[28] P. Schober, C. Boer, and L. A. Schwarte, “Correlation coefficients: appropriate use and interpretation,” *PubMed*, Jul. 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29481436/.

[29] A. Dibaji and S. Sulaimany, “Community detection to improve machine learning based heart disease prediction,” in *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, IEEE, Feb. 2024, pp. 1–6, doi: 10.1109/AISP61396.2024.10475216.

About the Authors

Mr. Zabihullah Burhani, Lecturer*, Department of Computer Science, Takhar University, Takhar, Afghanistan.* <*zabihullah.burhani@tu.edu.af*>; < zabihullahburhani@gmail.com>
Mr. Abolfazl Dibaji, Phd Student, *Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran.* <*abolfazl.dibaji1379@gmail.com*>.