

RESEARCH

Open Access



Risk prediction of heart failure in patients with ischemic heart disease using network analytics and stacking ensemble learning

Dejia Zhou¹, Hang Qiu^{1,2*}, Liya Wang² and Minghui Shen³

Abstract

Background Heart failure (HF) is a major complication following ischemic heart disease (IHD) and it adversely affects the outcome. Early prediction of HF risk in patients with IHD is beneficial for timely intervention and for reducing disease burden.

Methods Two cohorts, cases for patients first diagnosed with IHD and then with HF (N = 11,862) and control IHD patients without HF (N = 25,652), were established from the hospital discharge records in Sichuan, China during 2015–2019. Directed personal disease network (PDN) was constructed for each patient, and then these PDNs were merged to generate the baseline disease network (BDN) for the two cohorts, respectively, which identifies the health trajectories of patients and the complex progression patterns. The differences between the BDNs of the two cohort was represented as disease-specific network (DSN). Three novel network features were extracted from PDN and DSN to represent the similarity of disease patterns and specificity trends from IHD to HF. A stacking-based ensemble model DXLR was proposed to predict HF risk in IHD patients using the novel network features and basic demographic features (i.e., age and sex). The Shapley Addictive exPlanations method was applied to analyze the feature importance of the DXLR model.

Results Compared with the six traditional machine learning models, our DXLR model exhibited the highest AUC (0.934 ± 0.004), accuracy (0.857 ± 0.007), precision (0.723 ± 0.014), recall (0.892 ± 0.012) and F₁ score (0.798 ± 0.010). The feature importance showed that the novel network features ranked as the top three features, playing a notable role in predicting HF risk of IHD patient. The feature comparison experiment also indicated that our novel network features were superior to those proposed by the state-of-the-art study in improving the performance of the prediction model, with an increase in AUC by 19.9%, in accuracy by 18.7%, in precision by 30.7%, in recall by 37.4%, and in F₁ score by 33.7%.

Conclusions Our proposed approach that combines network analytics and ensemble learning effectively predicts HF risk in patients with IHD. This highlights the potential value of network-based machine learning in disease risk prediction field using administrative data.

*Correspondence:

Hang Qiu
qiuhan@uestc.edu.cn

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Keywords Administrative data, Ischemic heart disease, Heart failure, Comorbidity network, Network feature, Ensemble learning

Background

Ischemic heart disease (IHD) is one of the major underlying causes of heart failure (HF) [1–3] and is related to increase mortality [4, 5]. Abdissa et al. [6] investigated 306 IHD patients, 64.1% of whom developed HF with the number of females being about twice the number of males. In our previous study on the comorbidity patterns of IHD patients, HF occurred in 29.39% of the IHD patients, and the incidence of HF in IHD patients was eight times higher than that in patients without IHD [7]. HF, as a complex cardiovascular syndrome, causes frequent hospitalization, leads to low quality of life, and accounts for a large portion of cardiovascular disease (CVD) morbidity and mortality [8, 9]. Therefore, early prediction of HF risk in IHD patients may improve patients' outcomes, and reduce medical costs and mortality.

In recent years, as a sub-field of artificial intelligence, machine learning (ML) techniques attracted much attention in the medical domain [10–17], and have been increasingly employed for HF prediction [18–25]. For instance, Rammal et al. [26] integrated different types of data, including demographic data, chest X-ray images data, and clinical diagnostic and symptoms data, of 100 HF patients to construct random forest (RF) and logistic regression (LR) predictive models which both achieved an accuracy of 93%. Akbilgic et al. [27] developed a convolutional neural network (CNN) model that utilized electrocardiographic (ECG) data for predicting the risk of developing HF, which achieved an AUC of 75%. According to Chen et al. [28], they employed demographic data, diagnostic data, clinical test data and intraoperative monitoring data of patients to construct a model with deep pyramid CNN and extreme gradient boosting (XGBoost) method for forecasting the risk of HF after operation. Although the recent advances in ML techniques have significantly improved the prediction accuracy for HF, these works have two major issues. First, most ML models developed so far relied on multiple types of detailed medical data, such as cardiac image, laboratory examination and ECG data, and these models mainly aimed to interpret clinical data and assist clinicians for screening and diagnosis of HF. Few studies have attempted to establish ML-based risk prediction models for people who are likely to progress to HF (e.g., patients with IHD) when detailed diagnostic tests are unavailable. Second, these methods considered comorbidities, such as hypertension, diabetes and atrial fibrillation, as clinical risk factors for HF, but did not consider their complex relationships and progression patterns among comorbidities. Since HF is a

major complication following IHD, capturing the disease progression pathways can reveal the multimorbidity risk, thus increasing the accuracy of disease risk prediction.

Recently, the availability of large amounts of administrative data (e.g., hospital discharge records, HDR) and the development of network theory provide new opportunities to apply a predictive model for improving the disease risk assessment. The administrative data contains useful proxies for missing clinical predictors, e.g. diagnoses and procedures recorded during hospitalizations, and the network analysis offers effective approach to explore comorbidity patterns [29–31], and the temporal disease trajectories [32–34] of patients hidden in these data. For instance, a cross-sectional study [35] used electronic health records of 34,099 discharged patients and network analysis techniques and determined that the comorbidity networks of CVDs were highly centralized in prevalent diseases, such as cardiac arrhythmias, HF, chronic kidney disease, hypertension, and ischemic diseases. Using large-scale datasets and network science, Ong et al. [36] constructed a directed disease network to identify rare and novel disease patterns in pediatric pulmonary hypertension. Nevertheless, to the best of our knowledge, only very few studies have further combined network analytics with ML techniques to improve the healthcare system [37–40], especially for disease risk prediction. Khan et al. [41] constructed disease networks from 1.4 million admission records to predict the risk of type 2 diabetes. Their results showed that the measurements based on network theory ranked highest among the parameter estimation, LR, and decision tree (DT) models, with 82–87% prediction accuracy. Hossain et al. [37] developed risk prediction models using social network analysis on administrative datasets to determine the risk of type 2 diabetes in patients with CVD. They extracted three network-based features from the comorbidity network to indicate the comorbidity prevalence, transfer pattern, and cluster membership, and they constructed ML models with 79–88% accuracy. Using network analytics and administrative data, Uddin et al. [42] constructed five traditional ML models and two deep learning models to predict the number of chronic diseases. They concluded that the network analysis approach allowed them to better represent the relationship among patients' diseases.

This study aims to identify IHD patients at high risk of HF. Inspired by the previous studies [37, 41], we propose a risk prediction approach using disease network analysis combined with ensemble learning technique based on routinely collected administrative data. It has two goals: (1) design novel network features to capture the

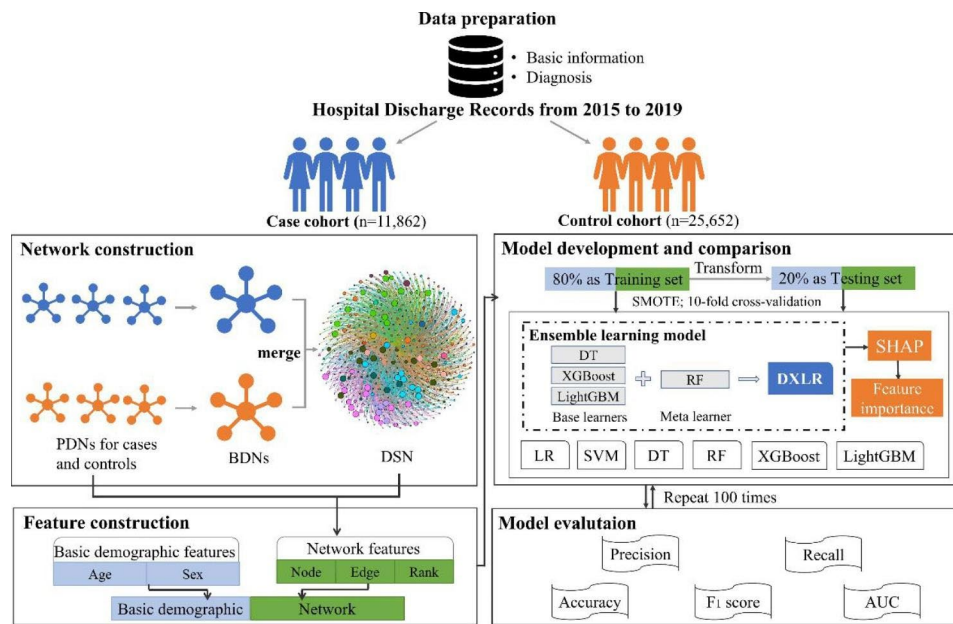


Fig. 1 The overview of the framework for predicting HF risk in IHD patients. PDNs: Personal Disease Networks. BDNs: Baseline Disease Networks. DSN: Disease-Specific Network. DXLR: the two-stage ensemble machine learning model our study proposed

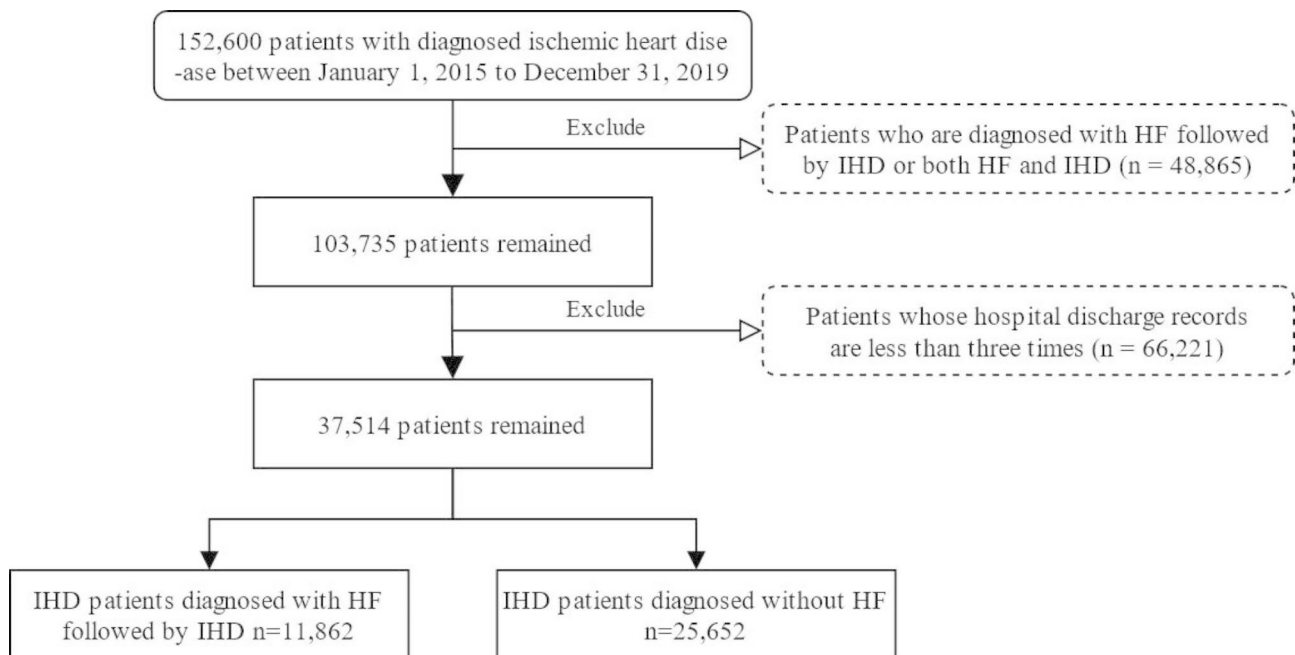


Fig. 2 The extraction process for two patient cohorts

specific complex progression pattern from IHD to HF; (2) develop a stacking-based ensemble model to predict HF risk for IHD patients using basic demographic information and network features.

Methods

Overview

Figure 1 shows the architecture of the framework for predicting HF risk in patients with IHD. First, two cohorts,

cases for patients first diagnosed with IHD and then with HF, and control IHD patients without HF, were established from the HDR dataset. Each patient’s HDR includes basic information, diagnosis history, admission and discharge time, etc. Then, three types of comorbidity networks, including personal disease network (PDN), baseline disease network (BDN), and disease-specific network (DSN), were constructed to identify patterns of patients’ diseases over time as well as the complex

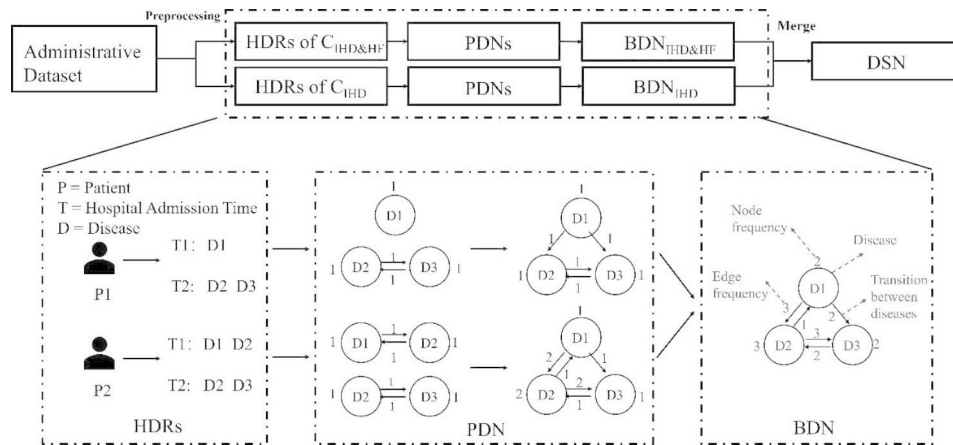


Fig. 3 The process of constructing networks. Nodes represent diseases and directed edges represent the sequential relationship between diseases. HDRs: Hospital Discharge Records. $C_{IHD\&HF}$: Cohort of patients with IHD and HF. C_{IHD} : Cohort of patients with IHD. PDNs: Personal Disease Networks (adapted from [37]). BDN: Baseline Disease Network. DSN: Disease-Specific Network

progression pattern from IHD to HF. Next, based on the DSN and PDN, three novel network features were generated to better characterize the progression patterns from IHD to HF. Moreover, to predict HF risk in IHD patients, while validating the effectiveness of the designed features, a stacking ensemble model and six traditional ML models were developed using different input features (basic demographic features and network features). Finally, the Shapley Addictive exPlanations (SHAP) [43] method was applied to analyze the feature importance of the proposed model.

Data preparation

This is a large-scale, retrospective study based on anonymized HDRs collected from all the secondary and tertiary hospitals in Sichuan Province, China from January 1, 2015 to December 31, 2019. Each record contains de-identified codes, sex, age, admission and discharge times, and diagnosis information. Standard ICD-10 (International Classification of Diseases, 10th Revision) coding was used for all the disease diagnosis data. Furthermore, IHD and HF patients were identified by the first three digits of the ICD code (i.e., IHD: I20-I25 and HF: I50).

As shown in Fig. 2, two patient cohorts that met the selection criteria were selected from the database. After excluding data with missing values, invalid values, inconsistent data, and redundant data, HDR data for a total of 152,600 IHD patients were obtained. Patients firstly diagnosed with HF then followed by IHD or with both HF and IHD at the same time were excluded (n=48,865). To identify the evolution pattern of disease over time, a more restrictive inclusion criterion was utilized, excluding patients with fewer than three hospital admissions (n=66,221). Finally, a total of 37,514 IHD patients with mean ages of 70.2 ± 10.7 years were included in this study, of which 48.2% were male. Among them, 11,862 patients

who were first diagnosed with IHD and then with HF during the study period served as the case group, while the other 25,652 patients without a diagnosis of HF after IHD diagnosis were served as the control group.

This study was approved by the Ethics Committee of Health Information Center of Sichuan Province. The requirement to obtain informed consent was waived because of the secondary nature of the de-identified data in the retrospective study design.

Network construction

The construction process of the disease networks is shown in Fig. 3. Three types of directed comorbidity networks were constructed using 65% of the entire dataset and the remaining 35% was used for modeling the ML models.

First, PDN was constructed for each patient in the two cohorts to describe the health trajectory of a patient during subsequent admissions over time [32]. In the PDN, the nodes, denoted as $v_i (v_i \in V(PDN))$, represent the diseases, and the edges, denoted as $ase_i (e_i \in E(PDN))$, represent the sequential relationships among diseases. The node weight, denoted as $freq(v_i)$, indicates the prevalence of a patient’s disease in all admission events, and the edge weight, denoted as $freq(e_i)$, indicates the number of times two diseases occurred during the same or consecutive admissions. Therefore, all nodes in PDN can be represented by a one-dimensional vector called disease vector. Each item in the vector represents a node v_i and the corresponding value is the node weight $freq(v_i)$. Similarly, all edges in the PDN can be represented by a two-dimensional matrix called disease adjacency matrix. Each element represents edge e_i and the corresponding value is edge weight $freq(e_i)$.

Next, to obtain the disease progression patterns of the patients in different cohorts, two BDNs (i.e., BDN_{IHD} &

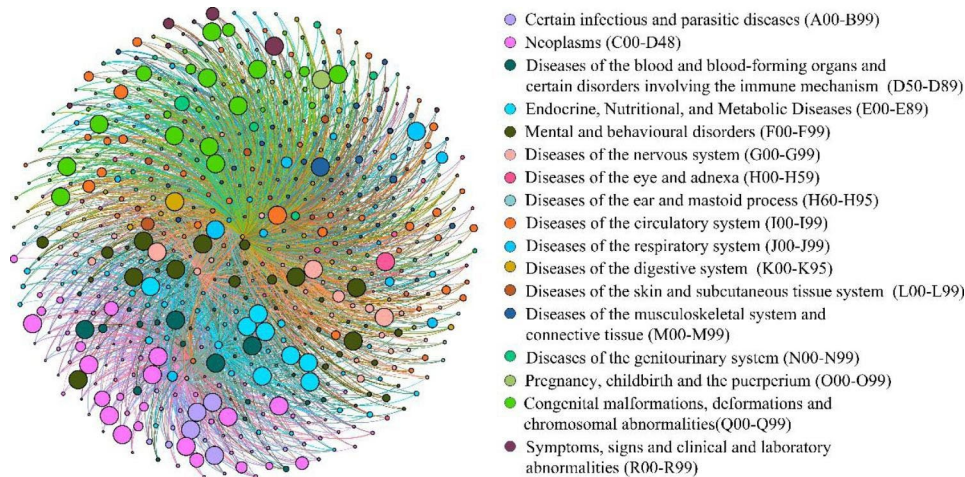


Fig. 4 Visualization of the Disease-Specific Network. Nodes represent diseases and node sizes represent disease prevalence. Directed edges represent the sequence of occurrence between diseases and the frequency of two diseases that occurred during the same or consecutive admissions. Edges weighted less than 0.5 are hidden for simplicity

Table 1 List of features considered in this study

Features	Descriptions	Number
Basic demographic information		2
Age	Patient's age	1
Sex	Male or female	1
Network		3
Node	Network node cosine similarity match score extracted from PDN and DSN	1
Edge	Network edge cosine similarity match score extracted from PDN and DSN	1
Rank	Network node centrality match score extracted from PDN and DSN	1

HF and BDN_{IHD}) were constructed by merging the corresponding PDNs from the two cohorts. The nodes and edges of the BDN and the corresponding weights were calculated by summing the nodes and edges of all the PDNs in the same cohort.

By considering the attribution theory [44], a final DSN was generated by combining $BDN_{IHD \& HF}$ and BDN_{IHD} . To better characterize the disease patterns in the control group, all the control patients were included, which led to an unequal number of patients in the two cohorts. Then considering the impact of the number of patients included in the cohort on the weight of the disease network, the network weight was modified to the relative frequency. The DSN affords more weight to the chronic comorbidities which are more prevalent in $BDN_{IHD \& HF}$ than in BDN_{IHD} . Moreover, it affords a low priority to the opposite conditions. The weight of the node and edge for DSN were calculated by determining their relative increments in $BDN_{IHD \& HF}$ compared to BDN_{IHD} . The final DSN exhibited the specific disease trajectory of patients in cases [37]. Figure 4 displays the disease network visualization of DSN.

Feature construction

As shown in Table 1, two types of features were extracted, including basic demographic features and network features.

Basic demographic features

Sex and age are risk factors for HF in IHD patients [6]. As shown in Table 1, these two basic characteristics were selected as the basic demographic features for modeling.

Network features

Three network features, including node score, edge score, and rank-based score were generated from the network to evaluate the HF risk in IHD patients, and to provide support from the network science perspective for early HF risk prediction.

Node score Inspired by the node match score described in [41], a modified network feature, called the node score, was proposed based on the weighted disease vector similarity. To measuring the angular similarity between the disease vectors of the PDN and the DSN, the node score takes into account how closely the two vectors are positioned in the disease vector space. The closer the two vectors, the higher their similarity, and the higher the node score. The node score is a disease similarity-based metric that measures the relationship between PDN and DSN from the disease similarity perspective. A higher node score denotes that PDN has more similar diseases to DSN, specifically the diseases present in IHD patients with HF as compared to those without HF.

Mathematically, the node score for a patient (i.e., PDN) is defined as follows:

$$F_{node} = \frac{\sum_{v_i \in V(PDN), v_j \in V(DSN), v_i = v_j} freq(v_i) * freq(v_j)}{\sqrt{\sum_{v_i \in V(PDN)} freq(v_i)^2} * \sqrt{\sum_{v_j \in V(DSN)} freq(v_j)^2}} \quad (1)$$

where v_i is the vertex i (i.e., disease i) in the PDN, v_j is the vertex j (i.e., disease j) in the DSN, and $freq(v)$ is the prevalence of a patient's disease occurring in all admission events.

Edge score Considering the evolutionary relationship between diseases over time, edge score was proposed as a metric based on the similarity of weighted disease vectors. The edges in the three constructed networks represent the temporal sequential relationship among diseases. The edge score was calculated based on the similarity of the edge vectors of the PDN and the DSN, which allowed for the characterization of the differences in the disease evolution paths between the two networks. A higher edge score suggests that the patient's disease progression pattern is more similar to that of the DSN.

Mathematically, the edge score for a patient (i.e., PDN) is defined as follows:

$$F_{edge} = \frac{\sum_{e_i \in E(PDN), e_j \in E(DSN), e_i = e_j} freq(e_i) * freq(e_j)}{\sqrt{\sum_{e_i \in E(PDN)} freq(e_i)^2} * \sqrt{\sum_{e_j \in E(DSN)} freq(e_j)^2}} \quad (2)$$

where e_i is the edge i (i.e., disease pairs i) in the PDN, e_j is the edge j (i.e., disease pairs j) in the DSN, and $freq(e)$ is the number of times the two diseases occurred during the same or consecutive admissions.

Rank-based score The significance of disease in the PDN compared to the DSN was identified by applying the PageRank algorithm [45] to determine the disease importance of nodes within the DSN. The sum of the weighted disease importance of PDN nodes was calculated to obtain the rank-based score, which reflects the importance of diseases in the PDN with respect to the DSN. The score is based on the importance of nodes in the DSN's network structure, i.e., it takes into account the importance of the disease nodes in the disease network as well as the relationship between the different diseases. A higher rank-based score for a patient's PDN indicates the presence of more diseases in the PDN that are also found in the DSN. These diseases have a high prevalence in the PDN and are characterized by a high node importance in the DSN.

Mathematically, the rank-based score for a patient (i.e., PDN) is defined as follows:

$$F_{rank} = \frac{\sum_{v_i \in V(PDN), v_j \in V(DSN), v_i = v_j} freq(v_i) * pg(v_j)}{|V(PDN)|} \quad (3)$$

where v_i is the vertex i (i.e., disease i) in the PDN, v_j is the vertex j (i.e., disease j) in the DSN, $freq(v)$ is the prevalence of a patient's disease occurring in all the admission events and $pg(v)$ is the PageRank value for vertex in the DSN, $|V(PDN)|$ is the total number of nodes (diseases) in the PDN.

Ensemble learning model construction

Traditional ML methods are becoming increasingly popular in the field of disease prediction due to their excellent prediction abilities, while models generated from the same data among different ML algorithms have great heterogeneity [46]. Integrating various ML models might be a feasible way to produce a more powerful and robust model. This study developed a two-stage stacked ensemble learning model DXLR using network features and basic demographic features, which was comprised of three base learners (DT, XGBoost and Light Gradient Boosting Machine (LightGBM)) and a meta learner (RF).

As shown in Fig. 5, in the first stage, five-fold cross-validation was performed for each of the models to generate a training set for the meta classifier. Among these folds, the base classifiers were used on four-folds, leaving one-fold for validation. Each base classifier output a new feature of the training set by merging the five validation folds and generated a new testing feature by averaging the five prediction results. Moreover, in the second stage, the most important feature of each basic learner was merged as crucial features to form the new training and testing set. The crucial features consisted of the most critical features in each base learner that have high feature importance for predicting HF risk in IHD patients.

Model comparison and evaluation

To compared with DXLR model, six traditional ML classifiers, including LR [47], support vector machines (SVM) [48], DT [49], RF [50], XGBoost [51], and LightGBM [52], were developed. LR is a widely used classical linear model and has the advantage of fast convergence. As for SVM, the linear kernel was selected as the kernel function to improve the training efficiency of the model under millions of datasets. Before training the LR and SVM models, standard normalization was applied to the datasets. The DT, RF, XGBoost, and LightGBM models are tree-based ensemble models, which proved the non-linear fitting ability and better ideas for improving the prediction model performance and robustness. A grid search strategy was adopted to determine the best model parameters.

A series of evaluation metrics, such as precision, recall, accuracy, and F_1 score, were used to evaluate the model performance. Furthermore, the area under the receiver operating characteristic curve (AUC) was obtained to compare the discrimination of the different ML models.

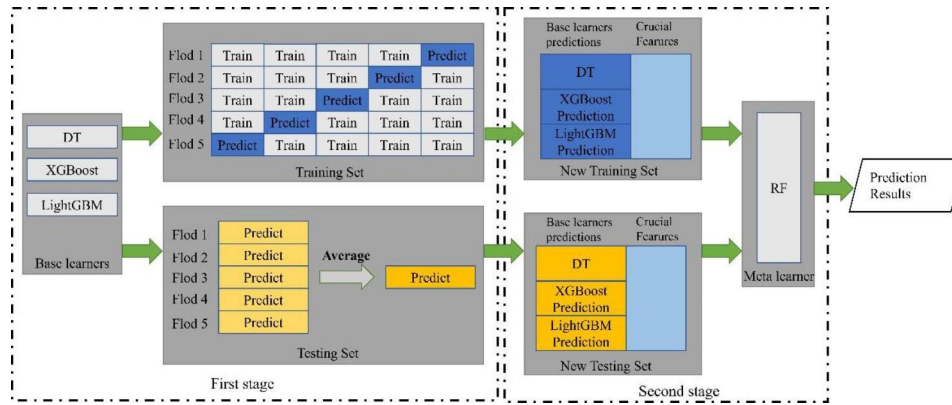


Fig. 5 The overall framework of DXLR.

To avoid data deviation caused by dataset partition, the dataset was randomly split into a training set (80%) and a testing set (20%) 100 times. The results were represented in the form of mean±standard deviation. The six traditional ML models and the DXLR model were trained and validated on the training set through 10-fold cross-validation on each randomly divided sample set. Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) [53] was used to avoid data imbalance.

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

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

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{6}$$

$$F_1score = \frac{2*TP}{2*TP + FP + FN} \tag{7}$$

where *TP* denotes true positive, *FP* represents false positive, *TN* indicates true negative, *FN* denotes false negative.

To better understand the impact of different features on the results, the SHAP methods was applied to further extend and enhance the prediction results given by the DXLR model.

Results

Descriptive statistics

This study included 37,514 patients with IHD from January 1, 2015, to December 31, 2019. The basic characteristics of all IHD patients are shown in Table 2 and are classified by the presence or absence of subsequent HF diagnosis. Overall, 11,862 (31.6%) patients who were first diagnosed with IHD and then with HF during their hospitalization served as the case group, while the remaining

Table 2 Basic characteristics of the IHD patients

Characteristics	Total (n=37,514)	Patients with HF (n=11,862)	Patients without HF (n=25,652)	P-value
Age(years)	70.21 ± 10.74	74.28 ± 9.55	68.33 ± 10.74	< 0.001
Sex				< 0.001
Male	18,067(48.16%)	5,870(49.48%)	12,197(47.55%)	
Female	19,447(51.84%)	5,992(50.51%)	13,455(52.45%)	

25,652 (68.4%) patients who were not diagnosed with HF served as the control group. The average age of the total patient group was 70.21±10.74 years and 48.16% of the patients were male. The average age of the case group was 5.95 years higher than that of the control group (74.28±9.55 vs. 68.33±10.74; *P*-value<0.001).

Comparison of models

Table 3 displays the performance of the six traditional ML models and the DXLR model with both network features and basic demographic features considered. The results show that the proposed DXLR model exhibited higher performance and better stability on all evaluation metrics compared with other models. Although, XGBoost was the best performing model among the six traditional ML models, DXLR showed significantly improvement (*P*-value<0.0001) in all metrics. As a result, the DXLR model was selected as the representative classifier for subsequent experiments.

Performance comparison of network features

The contributions of the three network features were further analyzed. As is shown in Table 4, the DXLR model with the rank-based score removed (i.e., using basic demographic features, node score, and edge score as input features) had the largest degradation in AUC performance. Meanwhile, the DXLR model with the node feature removed (AUC=0.927) was marginally better than model with the edge feature removed (AUC=0.923).

Table 3 Comparison of the performance in six traditional models and the proposed DXLR model

Models	Metric [Mean ± SD]				
	Precision	Recall	Accuracy	F ₁ score	AUC
LR	0.538±0.016	0.649±0.016	0.712±0.008	0.589±0.013	0.766±0.009
SVM	0.540±0.016	0.644±0.015	0.713±0.008	0.587±0.013	0.765±0.009
DT	0.647±0.015	0.827±0.020	0.802±0.007	0.726±0.011	0.879±0.006
RF	0.681±0.015	0.833±0.013	0.823±0.007	0.749±0.011	0.905±0.005
XGBoost	0.714±0.014	0.878±0.012	0.850±0.006	0.788±0.010	0.928±0.005
LightGBM	0.674±0.015	0.836±0.011	0.820±0.007	0.746±0.011	0.901±0.005
DXLR	0.723±0.014	0.892±0.012	0.857±0.007	0.798±0.010	0.934±0.004
P-value ^a	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

^a: t-test for the DXLR model and the best performing traditional model (XGBoost); The bold font is the best performing model

Table 4 Performance comparison of the DXLR model with removal of network feature separately

Method	Metric [Mean ± SD]				
	Precision	Recall	Accuracy	F ₁ score	AUC
DXLR	0.723±0.014	0.892±0.012	0.857±0.007	0.798±0.010	0.934±0.004
–Node	0.699±0.014	0.905±0.010	0.846±0.006	0.788±0.009	0.927±0.004
–Edge	0.681±0.013	0.907±0.012	0.836±0.006	0.778±0.009	0.923±0.004
–Rank	0.492±0.015	0.663±0.019	0.676±0.010	0.565±0.012	0.737±0.010

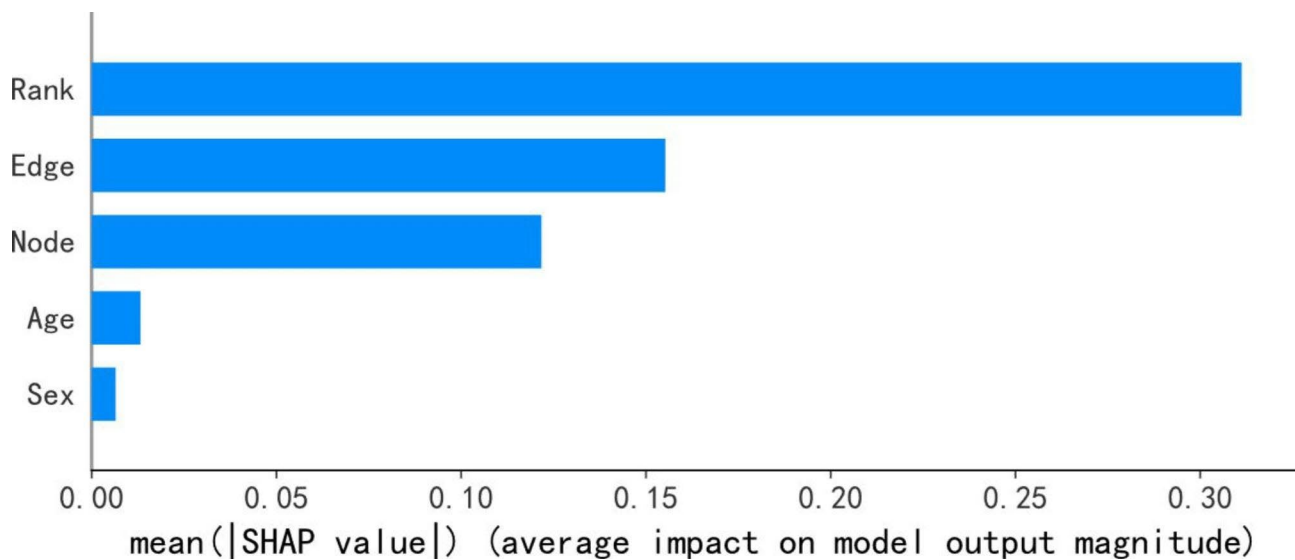


Fig. 6 SHAP summary plot of the DXLR model. Average absolute impact of features on the final model output magnitude ordered by decreasing feature importance. Rank: the rank-based score. Edge: the edge score. Node: the node score

Feature importance

To visually explain the importance of different characteristics in the classification of the two groups of patients considered, SHAP was applied to illustrate how these features affect the performance of the DXLR model. Figure 6 shows all the features evaluated by the average absolute SHAP values. The feature ranking (y-axis) indicates the importance of the prediction model and the SHAP value (x-axis) is a uniform index reflecting the impact of a particular feature in the model. Overall, the absolute SHAP values of network features were considerably higher than those of the basic demographic features. The highest was

the rank-based score (0.311), followed by the edge score (0.155) and the node score (0.122).

Sensitivity analysis of the predictive model

To analyze the generalization performance of the DXLR model, the performance of our best classifier (with basic demographic features and network features) was compared on different subsets of patients. Table 5 lists the comparison results. The AUC and accuracy values of the predictive model for females were slightly higher (about 0.01) than those for males. The DXLR model showed no statistically significant difference in precision, recall, and F₁ score values for different sex stratifications

Table 5 Performance in stratified subgroups

Subgroup	Metrics [Mean \pm SD]				
	Precision	Recall	Accuracy	F ₁ score	AUC
Sex					
Male	0.716 \pm 0.020	0.881 \pm 0.017	0.848 \pm 0.010	0.790 \pm 0.015	0.926 \pm 0.007
Female	0.722 \pm 0.018	0.875 \pm 0.019	0.857 \pm 0.009	0.791 \pm 0.014	0.932 \pm 0.006
P-value ^a	0.0327	0.0277	< 0.0001	0.5533	< 0.0001
Age groups					
18–44	0.798 \pm 0.185	0.901 \pm 0.143	0.960 \pm 0.030	0.832 \pm 0.148	0.975 \pm 0.044
45–59	0.685 \pm 0.053	0.801 \pm 0.060	0.919 \pm 0.014	0.736 \pm 0.042	0.953 \pm 0.013
60–69	0.658 \pm 0.034	0.809 \pm 0.029	0.860 \pm 0.013	0.725 \pm 0.026	0.924 \pm 0.010
70–79	0.725 \pm 0.022	0.887 \pm 0.018	0.838 \pm 0.012	0.798 \pm 0.015	0.921 \pm 0.008
80+	0.757 \pm 0.021	0.928 \pm 0.017	0.816 \pm 0.014	0.834 \pm 0.015	0.894 \pm 0.013
P-value ^b	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

^a: t-test; ^b: one-way analysis of variance.

Table 6 Comparison of the classifier and features proposed in this study with the features proposed in previous study

Classifier	Network Features	Metrics [Mean \pm SD]				
		Precision	Recall	Accuracy	F ₁ score	AUC
DXLR model	our study	0.723 \pm 0.014	0.892 \pm 0.012	0.857 \pm 0.007	0.798 \pm 0.010	0.934 \pm 0.004
	[37]	0.553 \pm 0.017	0.649 \pm 0.021	0.722 \pm 0.009	0.597 \pm 0.013	0.779 \pm 0.010
Random Forest	our study	0.681 \pm 0.015	0.833 \pm 0.013	0.823 \pm 0.007	0.749 \pm 0.011	0.905 \pm 0.005
	[37]	0.532 \pm 0.016	0.704 \pm 0.016	0.710 \pm 0.009	0.606 \pm 0.013	0.781 \pm 0.010

(P -value $>$ 0.0001). In terms of the age groups, all the performance metrics exhibited statistically significant differences. The DXLR model showed the highest accuracy, precision, and AUC values for the 18–44 age group among the five age groups (0.798 \pm 0.185, 0.960 \pm 0.030, and 0.975 \pm 0.044, respectively). While for the 80+ age group, the DXLR model had the highest recall and F₁ score values and the lowest accuracy and AUC values.

Comparison with state-of-the-art study

Our proposed network features for predicting HF risk for IHD patients were also compared with those in a state-of-the-art study (three network features designed by Hossain [37]) under different classifiers. All the features were generated based on our dataset, and basic demographic features were sex and age. The DXLR model and RF (the best model in [37]) were selected as the models for feature comparison and verification. Table 6 verifies the superior performance of our features in all the evaluation metrics for both classifiers. The DXLR model using the network features proposed in this study outperformed the model using network features in [37], with an increase in precision by 30.7%, in recall by 37.4%, in accuracy by 18.7%, in F₁ score by 33.7% and in AUC by 19.9%, respectively. Compared with the RF model using network features in [37], the precision, recall, accuracy, F₁ score and AUC of the RF model using our network features increased by 28.0%, 18.3%, 15.9%, 23.6%, and 15.9%, respectively.

Discussion

This study proposes an approach to predict high-risk groups for HF among IHD patients using routinely collected administrative data. By integrating network analytics with ensemble learning, our approach is able to extract disease patterns hidden in administrative data and identify patients at high risk that may benefit from screening and a preventive strategy. This method could be used in other regions where large administrative datasets can be linked at the individual person level to help health authorities identify high-risk groups and formulate targeted policies to better guide individuals, thus reducing the risk of illness.

In this study, the performance of the DXLR model was compared with six traditional models, and the results showed that the DXLR model outperformed all the other models. Among the traditional models, tree-based models performed relatively better compared to LR and SVM, with XGBoost exhibiting the best performance. However, the proposed DXLR model performed significantly better (P -value $<$ 0.0001) than the XGBoost model in all metrics, achieving a precision of 0.723, a recall of 0.892, an accuracy of 0.857, an F₁ score of 0.798 and an AUC of 0.934. This performance improvement highlights the effectiveness of our DXLR model, which combines the strengths of multiple models to achieve higher prediction accuracy.

Although there were some variances in performances across subgroups, such as a slightly lower AUC in male patients and decreasing AUC with increasing age, our proposed DXLR model still demonstrated a strong and

stable predictive ability. These differences in performance could be due to their differences in comorbidity burden and in complex comorbidity relationships in male and older IHD patients [7]. Nevertheless, our model provides a promising tool for identifying high risk groups for HF in diverse IHD patient populations.

The comparison of the contributions of the three network features revealed that the removal of the rank-based score led to the most significant drop in AUC (0.197), followed by a 0.011 and 0.007 decrease in AUC when the edge score and node score were removed, respectively. Furthermore, SHAP was applied to validate the contribution of the three network features to the prediction results of the DXLR model. The rank-based score yielded the highest SHAP value score, with about twice as important as the edge and the node score. These results suggest that the rank-based score is the most critical network feature for predicting HF risk in IHD patients. One possible explanation for this finding is that the rank-based score captures the relative importance of diseases within the DSN network. A higher number of specific diseases in a patient's PDN that are important for the disease progression patterns from IHD to HF in the DSN, may increase the likelihood of developing HF. Additionally, the edge score reflect the similarity of disease progression trajectories between patients, while the node score capture the overall disease burden of patients. The combination of these network features provides a more comprehensive characterization of the risk of HF in IHD patients and helps to identify high risk patients that could benefit from early screening and prevention strategies.

The three network features proposed by our study exhibit better prediction performance than the network features designed in a state-of-the-art study [37] under the same model and basic demographic features. As shown in Table 6, compared with the features designed by Hossain et al. [37], our network features captured the complex comorbid and progressive relationship between IHD and HF, improving the predictive metrics of the DXLR model by 0.135–0.243. Compared with the best-performing RF model in [37], the precision and F_1 score improved by 0.149 and 0.143, respectively. These network features were used to measure the propensity of patients to progress from IHD to HF with node score and edge score characterized the disease propensity and disease pair progression propensity of IHD patients by measuring the similarity of disease vectors and disease pair vectors of PDN and DSN, respectively. The rank-based score portrays the disease-weighted propensity of IHD patients by measuring the similarity of the weighted disease vectors of PDN and DSN. Therefore, models using our proposed network features performed better than those in previous work.

Our study has several limitations. First, the model this study proposed is only suitable for early risk prediction, not for clinical auxiliary diagnosis. Second, our network features were proposed based on the association of disease pairs or the progressive relationship of disease pairs. The prediction accuracy could be further improved by considering the patient-to-patient similarity inherent in the administrative dataset [38, 54]. In addition, graph neural networks (GNNs) are increasingly popular for learning network-based tasks [55, 56]. Future research is recommended to incorporate GNN-based algorithms to better utilize the network and improve its predictive performance.

Conclusions

This study proposed an approach to predict risk of HF in patients with IHD by integrating network analytics with ensemble learning. Experimental results showed our proposed DXLR model outperformed the other traditional ML models. Further experiments also demonstrated our proposed network features exhibited better performance on the same data and model compared with the features created by the state-of-the-art study. These results highlight the potential value of network-based ML in disease risk prediction field using administrative data.

Abbreviations

IHD	Ischemic heart disease
HF	Heart failure
CVD	Cardiovascular disease
PDN	Personal disease network
BDN	Baseline disease network
DSN	Disease-specific network
SHAP	Shapley additive explanations
XGBoost	Extreme gradient boosting
DT	Decision tree
LightGBM	Light gradient boosting machine
RF	Random forest
LR	Logistic regression
SVM	Support vector machine

Acknowledgements

Not applicable.

Author Contribution

HQ conceived and designed the study. DZ performed the experiments, analyzed the data and wrote the first draft of the manuscript. LW and MS participated in data collection and performed the statistical analyses. All authors contributed to the interpretation of the results. HQ and LW critically revised the manuscript. All authors have read and approved the final manuscript.

Funding

This research was supported by the National Natural Science Foundation of China (No. 71661167005), and the Key Research and Development Program of Sichuan Province (No. 2019YFS0271), which provide financial support in the design of study, analysis of data and writing the manuscript.

Data Availability

The data that support the findings of this study are available from Health Information Center of Sichuan Province but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the

corresponding author upon reasonable request and with permission of the Health Information Center of Sichuan Province.

Declarations

Ethics approval and consent to participate

This study was conducted according to the ethical guidelines of the Helsinki Declaration and was approved by the Ethics Committee of Health Information Center of Sichuan Province. The Ethics Committee exempted informed consent because of the retrospective nature of this research. Prior to the analysis, patients' data were anonymized and de-identified.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹School of Computer Science and Engineering, University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu, Sichuan 611731, P.R. China

²Big Data Research Center, University of Electronic Science and Technology of China, Chengdu, China

³Health Information Center of Sichuan Province, Chengdu, China

Received: 23 November 2022 / Accepted: 15 May 2023

References

- Klein L, Gheorghiadu M. Coronary artery disease and prevention of heart failure. *Med Clin North Am*. 2004;88:1209–35.
- Vedin O, Lam CSP, Koh AS, Benson L, Teng THK, Tay WT, et al. Significance of ischemic heart disease in patients with heart failure and preserved, Midrange, and reduced ejection fraction. *Circulation: Heart Failure*. 2017;10:e003875.
- Lund LH, Mancini D. Heart failure in women. *Med Clin North Am*. 2004;88:1321–45.
- Badar AA, Perez-Moreno AC, Jhund PS, Wong CM, Hawkins NM, Cleland JGF, et al. Relationship between angina pectoris and outcomes in patients with heart failure and reduced ejection fraction: an analysis of the controlled rosuvastatin multinational trial in Heart failure (CORONA). *Eur Heart J*. 2014;35:3426–33.
- Tromp J, Ouwerkerk W, Cleland JGF, Angermann CE, Dahlstrom U, Tiew-Hwa Teng K, et al. Global differences in Burden and Treatment of Ischemic Heart Disease in Acute Heart failure. *JACC: Heart Failure*. 2021;9:349–59.
- Abdissa SG. Predictors of incident heart failure in a cohort of patients with ischemic heart disease. *Pan Afr Med J*. 2020;35:1–12.
- Zhou D, Wang L, Ding S, Shen M, Qiu H. Phenotypic Disease Network Analysis to identify comorbidity patterns in hospitalized patients with ischemic heart Disease using large-scale Administrative Data. 2022;:17.
- Levy D, Kenchaiah S, Larson MG, Benjamin EJ, Kupka MJ, Ho KKL, et al. Long-term trends in the incidence of and survival with heart failure. *N Engl J Med*. 2002;347:1397–402.
- Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics-2020 update: a Report from the American Heart Association. *Circulation*. 2020;141:e139–596.
- Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Comput Struct Biotechnol J*. 2016;15:26–47.
- Lorenzoni G, Sabato SS, Lanera C, Bottigliengo D, Minto C, Ocagli H, et al. Comparison of machine learning techniques for prediction of hospitalization in heart failure patients. *J Clin Med*. 2019;8:1298.
- Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep*. 2020;10:16057.
- Alizadehsani R, Roshanzamir M, Abdar M, Beykikhoshk A, Khosravi A, Nahavandi S, et al. Hybrid genetic-discretized algorithm to handle data uncertainty in diagnosing stenosis of coronary arteries. *Expert Syst*. 2022;39:e12573.
- Khozeimeh F, Sharifrazi D, Izadi NH, Joloudari JH, Shoeibi A, Alizadehsani R, et al. RF-CNN-F: random forest with convolutional neural network features for coronary artery disease diagnosis based on cardiac magnetic resonance. *Sci Rep*. 2022;12:11178.
- Zhang Z, Qiu H, Li W, Chen Y. A stacking-based model for predicting 30-day all-cause hospital readmissions of patients with acute myocardial infarction. *BMC Med Inform Decis Mak*. 2020;20:335.
- Duan C, Deng H, Xiao S, Xie J, Li H, Zhao X, et al. Accelerate gas diffusion-weighted MRI for lung morphometry with deep learning. *Eur Radiol*. 2022;32:702–13.
- Jin K, Huang X, Zhou J, Li Y, Yan Y, Sun Y, et al. FIVES: a Fundus Image dataset for Artificial Intelligence based Vessel Segmentation. *Sci Data*. 2022;9:475.
- Javeed A, Khan SU, Ali L, Ali S, Imrana Y, Rahman A. Machine learning-based Automated Diagnostic Systems developed for heart failure prediction using different types of data modalities: a systematic review and future directions. *Comput Math Methods Med*. 2022;2022:9288452.
- Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning: the multi-ethnic study of atherosclerosis. *Circ Res*. 2017;121:1092–101.
- Yang G, Ren Y, Pan Q, Ning G, Gong S, Cai G et al. A heart failure diagnosis model based on support vector machine. In: 2010 3rd International Conference on Biomedical Engineering and Informatics. 2010. p. 1105–8.
- Ali MM, Paul BK, Ahmed K, Bui FM, Quinn JMW, Moni MA. Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. *Comput Biol Med*. 2021;136:104672.
- Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. 2019;40:1975–86.
- Guo C-Y, Wu M-Y, Cheng H-M. The Comprehensive Machine Learning Analytics for Heart failure. *Int J Environ Res Public Health*. 2021;18:4943.
- Alotaibi FS. Implementation of machine learning model to Predict Heart failure disease. *Int J Adv Comput Sci Appl (IJACSA)*. 2019;10.
- Menshaw A, Hassan MM, Allheeb N, Fortino G. A Hybrid Generic Framework for Heart Problem diagnosis based on a machine learning paradigm. *Sens (Basel)*. 2023;23:1392.
- Rammal HF. Z. A. Heart Failure Prediction Models using Big Data Techniques. *ijacsa*. 2018;9.
- Akbilgic O, Butler L, Karabayir I, Chang PP, Kitzman DW, Alonso A, et al. ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure. *Eur Heart J Digit Health*. 2021;2:626–34.
- Chen Y, Qin X, Zhang L, Yi B. A Novel Method of Heart failure prediction based on DPCNNXGBOOST Model. *cmc*. 2020;65:495–510.
- Qiu H, Wang L, Zeng X, Pan J. Comorbidity patterns in depression: a disease network analysis using regional hospital discharge records. *J Affect Disord*. 2022;296:418–27.
- Wang L, Qiu H, Luo L, Zhou L. Age- and sex-specific differences in multi-morbidity patterns and temporal Trends on assessing Hospital Discharge Records in Southwest China: Network-Based study. *J Med Internet Res*. 2022;24:e27146.
- Divo MJ, Casanova C, Marin JM, Pinto-Plata VM, de-Torres JP, Zulueta JJ, et al. COPD comorbidities network. *Eur Respir J*. 2015;46:640–50.
- Khan A, Uddin S, Srinivasan U. Comorbidity network for chronic disease: a novel approach to understand type 2 diabetes progression. *Int J Med Informatics*. 2018;115:1–9.
- Derevitskii I, Funkner A, Metsker O, Kovalchuk S. Graph-Based Predictive Modelling of Chronic Disease Development: Type 2 DM Case Study. *pHealth* 2019. 2019;:150–5.
- Ko K, Lee CW, Nam S, Ahn SV, Ho J, Ban CY et al. Epidemiological characterization of a Directed and Weighted Disease Network using data from a cohort of one million patients: Network Analysis. *JOURNAL OF MEDICAL INTERNET RESEARCH*:12.
- Cruz-Ávila HA, Vallejo M, Martínez-García M, Hernández-Lemus E. Comorbidity Networks in Cardiovascular Diseases. *Front Physiol*. 2020;11.
- Ong M-S, Mullen MP, Austin ED, Szolovits P, Natter MD, Geva A, et al. Learning a comorbidity-driven taxonomy of Pediatric Pulmonary Hypertension. *Circ Res*. 2017;121:341–53.
- Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Syst Appl*. 2021;164:113918.
- Hu Z, Qiu H, Wang L, Shen M. Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission. *BMC Med Inform Decis Mak*. 2022;22:62.

39. Wang K, Tian J, Zheng C, Yang H, Ren J, Liu Y, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput Biol Med.* 2021;137:104813.
40. Yang P, Qiu H, Wang L, Zhou L. Early prediction of high-cost inpatients with ischemic heart disease using network analytics and machine learning. *Expert Syst Appl.* 2022;210:118541.
41. Khan A, Uddin S, Srinivasan U. Chronic disease prediction using administrative data and graph theory: the case of type 2 diabetes. *Expert Syst Appl.* 2019;136:230–41.
42. Uddin S. Comorbidity and multimorbidity prediction of major chronic diseases using machine learning and network analytics. *Expert Syst Appl.* 2022;10.
43. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. 2017.
44. Hossain ME, Uddin S, Khan A, Moni MA. A Framework to understand the progression of Cardiovascular Disease for type 2 diabetes Mellitus Patients using a Network Approach. *IJERPH.* 2020;17:596.
45. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. 1999. <http://ilpubs.stanford.edu:8090/422/>. Accessed 2 Sep 2022.
46. Taha K, Ross HJ, Peikari M, Mueller B, Fan C-PS, Crowdy E, et al. An ensemble-based approach to the development of clinical prediction models for future-onset heart failure and coronary artery disease using machine learning. *J Am Coll Cardiol.* 2020;75 11 Supplement1:2046–6.
47. Logistic regression and artificial. Neural network classification models: a methodology review. *J Biomed Inform.* 2002;35:352–9.
48. Chen P-H, Lin C-J, Schölkopf B. A tutorial on v-support vector machines. *Appl Stoch Models Bus Ind.* 2005;21:111–36.
49. Rokach L, Maimon O. Decision Trees. In: *The Data Mining and Knowledge Discovery Handbook.* 2005. p. 165–92.
50. Pal M. Random forest classifier for remote sensing classification. *Int J Remote Sens.* 2005;26:217–22.
51. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94.
52. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems.* Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149–57.
53. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Int Res.* 2002;16:321–57.
54. Lu H, Uddin S, Hajati F, Moni MA, Khushi M. A patient network-based machine learning model for disease prediction: the case of type 2 diabetes mellitus. *Appl Intell.* 2022;52:2411–22.
55. Lu H, Uddin S. OPEN a weighted patient network-based framework for predicting chronic diseases using graph neural networks. *Sci Rep.* 13.
56. Sun Z, Yin H, Chen H, Chen T, Cui L, Yang F. Disease Prediction via Graph neural networks. *IEEE J Biomed Health Inform.* 2020;1–1.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.