MACHINE LEARNING-BASED PREDICTION OF HEART FAILURE USING GENETIC DATA

Tuan Anh Vu, Dang Tran Le Anh, Minh Tuan Nguyen
Posts and Telecommunications Institute of Technology

Abstract **–** Heart failure is a major global concern affecting millions of people. The disease is characterized by high mortality and significant economic burden. Therefore, in this study, we propose a highly accurate, rapid and timely model for the diagnosis of preclinical heart failure based on genetic biomarkers. This model consists of a Random Forest classifier and 10 differentially expressed genes selected using the particle swarm optimization algorithm. Our results demonstrated its effectiveness, with accuracy, precision, specificity, recall, F1-score, and AUC achieving 91.92%, 94.07%, 91.78%, 92.09%, 93.04%, and 91.94%, respectively, on the GSE57345 dataset using 5-fold cross-validation. These findings indicate that, despite differences among patient groups, our model remains highly effective and can be applied for personalized disease prediction and precision medicine.

Keywords— Heart failure, machine learning, gene selection, gene expression omnibus, differentially expressed genes.

I. INTRODUCTION

Currently, heart failure affects approximately 64 million people worldwide [1]. It is a common condition associated with high mortality, reduced quality of life, and significant economic impact. In addition, its prevalence continues to increase due to an aging population and improved access to evidence-based treatments that influence disease progression [2]. Heart failure arises from various causes, making it a complex syndrome [3]. A better understanding of its underlying mechanisms is essential to optimize management and provide personalized treatment. Prediction of heart failure at a preclinical stage can significantly improve patient outcomes by allowing for early interventions and lifestyle changes. Conventional methods of diagnosing heart failure are mainly based on the clinical signs and symptoms, with echocardiography and chest X-ray. However, these tests are inaccurate in the intermediate and late stages of the disease and lack clinical specificity and sensitivity.

Approaches to characterizing heart failure patients include comprehensive, multimodal assessments ranging from electrocardiography and echocardiography to advanced imaging techniques such as cardiac magnetic resonance and nuclear imaging, with the recent addition of artificial intelligence (AI)-assisted diagnostic tools [4]. Genetic biomarkers are emerging as fundamental tools for both heart failure diagnosis and prognosis [5]. Biomarkers are biological molecules found primarily in blood, other body fluids, or tissues and typically include DNA, RNA, microRNA, epigenetic modifications, or antibodies. They have high sensitivity, specificity and positive diagnostic value for diseases. Heart failure is one of the diseases that has a complex pathophysiological process, involving many factors. The genotype is identified at a preclinical heart failure stage, which might be beneficial in delaying or preventing disease progression. Moreover, heart failure can be predicted using a single gene [6].

In recent years, the rapid development of high-throughput technologies and bioinformatics has allowed the simultaneous analysis of thousands of genes in different disease samples [7]. As a result, the use of potential biomarkers for diagnosis, prognosis, and personalized medical services has increased. Therefore, heart failure diagnosis has been conducted on biomarkers by numerous researchers. Among the earliest biomarkers used for detecting acute HF was B-type natriuretic peptide (BNP) [8]. In patients with chronic HF, Nt-proANP and Nt-proBNP exhibit higher plasma concentrations, greater stability, and improved diagnostic value [9]. Studies indicate that assessing both adiponectin and NT-proBNP together provides greater accuracy compared to NT-proBNP alone [10].

Furthermore, AI is a rapidly growing tool with active applications in the medical field [11]. With the continued exploration of the potential of artificial intelligence, AI-based clinical research will lead to a paradigm shift in medical practice, thereby significantly improving the survival rate of many diseases including cancer [12]. Indeed, Machine Learning (ML) is transforming healthcare by guiding individual and population health through a variety of computational benefits. It contributes to patient

Contact author: Minh Tuan Nguyen
Email: nmtuan@ptit.edu.vn
Manuscript received: 3/2025, revised: 4/2025, accepted: 5/2025.

observation, disease pattern analysis, diagnosis and prescription of drugs, patient-centered care delivery, clinical error reduction, predictive scoring, treatment decision making, and detection of sepsis and high-risk emergencies in patients.

In study [13], three ML algorithms: Least Absolute Shrinkage and Selection Operator (LASSO), RF, and Support Vector Machine Recursive Feature Elimination (SVM-RFE) were used to screen 14 genes related to heart failure aging. These genes were verified through various ML algorithms.

In research [14], the authors used two ischemic heart failure datasets from the GEO database (GSE76701 and GSE21610) and identified four potential diagnostic candidate genes for ischemic heart failure using bioinformatics and machine learning algorithms, namely RNASE2, MFAP4, CHRDL1, and KCNN3. They constructed a nomogram and validated the diagnostic value of these genes on additional GEO datasets (GSE57338).

Study [15] proposed a novel diagnostic model that is capable of predicting worsening heart failure while providing easy interpretation of the results. They proposed a threshold-based binary classifier built on a mathematical model derived from the Genetic Programming (GP) approach. The results showed that the proposed GP-based classifier achieved an average score of 96% for all the considered evaluation metrics and fully supported the control measures of healthcare workers.

There is increasing evidence that aberrant gene expression is an important event in heart failure [16]. Differentially expressed genes refer to genes whose expression levels significantly increase or decrease between different conditions or groups. After comprehensive gene expression analyzes such as RNA-Seq or microarrays, differential expression analysis is performed to identify differentially expressed genes (DEGs).

In the study [17], genes were tested for differential expression using DESeq2, and the DEGs were analyzed for protein-protein interactions (PPIs) and associated ontological pathways using Metascape. As a result, seven genes were identified, which were involved in two possible mechanisms of pain in heart failure: immune/inflammatory processes and atherosclerotic processes. In research [18], principal component analysis and hierarchical clustering were tested for transcriptional differences between groups, the impact of comorbidities, and DEG with pathway enrichment between heart failure and donor controls.

Several studies have demonstrated the efficacy of combining DEG and ML to predict heart failure. However, the number of these studies is not much. In study [19], by using the DEGs between normal and heart failure samples in the Gene Expression Omnibus (GEO) database with circadian rhythm-related genes, differentially expressed circadian rhythm-related genes were obtained. The authors used Machine Learning (ML) to screen the feature genes, and diagnostic models were built based on these feature genes. The results demonstrated that the ML model-based diagnosis had higher accuracy and could perfectly distinguish HF patients from normal patients. In another study [20], the authors proposed an accurate heart failure diagnostic model using Random Forest (RF) and Artificial Neural Networks (ANN) based on DEGs in subject with or without heart failure. The performance of their proposed method was illustrated as the area under the curve (AUC) of the training and testing sets were 0.996 and 0.863, respectively.

The identification of key genes in heart failure patients remains a challenge, and the potential of combining DEG and ML has yet to be fully explored. Therefore, in this study, we propose an algorithm based on public genomic data. By using differential gene expression analysis, we then use the particle swarm optimization (PSO) method to select potential genes from this gene set. Finally, we use four machine learning models, namely Random Forest (RF), XGBoost, Logistic regression (LR), and Support Vector Machine (SVM). The ML models are optimized by grid search cross-validation, and then trained and tested on different datasets. Our main contributions include:

- Using the PSO method combined with RF to select the most relevant genes for diagnosing heart failure.
- Proposing a set of biomarker genes to recommend to physicians, helping physicians gain deeper insights into the mechanisms of heart failure.
- Developing a simple model combining DEG genes and ML models to improve the performance of gene-based diagnosis.

The rest of paper follows this organizational framework, with the methodology outlined in Section II. Sections III and IV present the simulation results and discussion, respectively, followed by the concluding remarks in Section V.

II. METHOD

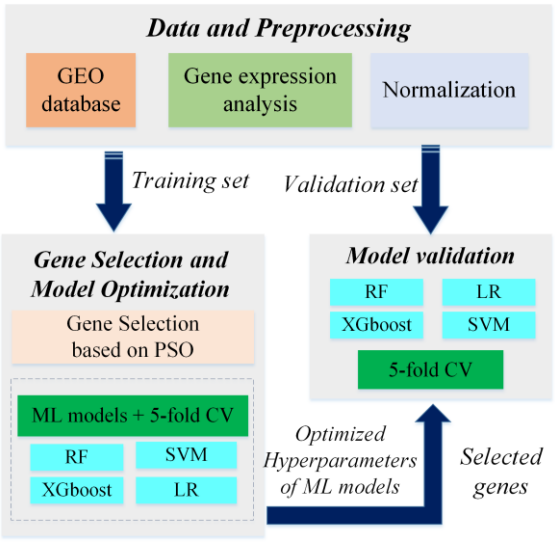

Figure 1: Flowchart of the proposed method

Our proposed method includes four main steps namely data preprocessing, gene selection, model optimization and

model testing, as shown in Figure 1. Firstly, in the data preprocessing stage, gene expression levels are computed to filter DEGs, and then normalized. Secondly, in step 2, Particle swarm optimization (PSO) method combined with a RF classifier is applied to find an optimal subset of DEGs. Thirdly, in the model optimization step, we use grid search with 5-fold cross-validation (CV) to optimize four ML models. Finally, in the model evaluation step, the optimal ML models are assessed on the validation set with 5-fold CV to evaluate their performance.

A. Data

Genetic datasets, including heart failure patients and healthy controls are collected from three datasets, namely GSE5406 (210 subjects) [21], GSE3586 (25 subjects) [22], and GSE57345 (319 subjects) [23] at the GEO database (https://www.ncbi.nlm.nih.gov/geo/). GSE5406 includes 16 cases of non- heart failure, 86 cases of dilated cardiomyopathy, and 108 cases of ischemic heart disease. GSE3586 includes 15 subjects without heart failure and 13 subjects with dilated cardiomyopathy. GSE57345 includes 139 subjects without heart failure, 96 subjects with ischemic heart disease, and 84 subjects with dilated cardiomyopathy.

For training, we use 238 subjects of two datasets (GSE5406 and GSE3586), including 31 non-heart failure cases and 207 heart failure cases. For validation, we use GSE57345 dataset, which includes 139 without heart failure cases and 180 with heart failure cases.

B. Gene preprocessing

Initially, background correction, normalization, and log2 transformation were applied to the three heart failure raw datasets using R (version 4.1.2). For genes identified by multiple probes, the average value was calculated to determine their expression. After merging the datasets, the Bioconductor "SVA" R package was used to remove batch effects. Finally, genes with |log Fold Change (FC)| > 0.6, false discovery rate (FDR) <0.05, and an adjusted p-value < 0.05 were considered differentially expressed genes (DEGs) using the Limma package.

C. Gene selection

PSO [24] is used for gene selection. It is a heuristic optimization algorithm inspired by the social behavior of birds flocking or fish schooling. PSO is used to identify the most relevant subset of genes that enhances the performance of a predictive model. A subset of genes is represented by each particle in the swarm, which explores the solution space by leveraging both its own experience and that of neighboring particles. Through iterative position updates and performance evaluations of the gene subsets, the algorithm efficiently explores the complex search space to identify an optimal gene set.

PSO is implemented with a swarm size of 30 particles, a maximum of 50 iterations, and an inertia weight of 0.9, with both the cognitive (c1) and social (c2) acceleration factors set to 1.5. The fitness function is based on the cross-validated AUC score of RF model.

D. Machine learning models

Four machine learning models, namely Random Forest (RF), XGBoost, Logistic Regression (LR), and Support Vector Machine (SVM), are trained using two gene datasets (GSE5406 and GSE3586), and then tested on GSE57345 dataset. To improve performance in diagnosing heart failure disease, these models are optimized using grid search with 5-fold CV to find optimal models. The ML models are described as follows:

Random forest [25]: RF is an ensemble learning method that generates an ensemble of decision trees, each trained on a random dataset. It is known for its robustness and ability to handle complex and noisy datasets. In our study, RF aggregates the results from multiple trees to classify gene expression data and identify patterns associated with heart failure.

XGBoost [26]: XGBoost is a gradient boosting algorithm that focuses on optimizing predictive performance through sequential model building. It creates a series of weak learners and improves them iteratively by correcting the errors of previous models.

Logistic regression [27]: LR is a statistical method used for binary classification. It models the probability that a given input belongs to a specific class using a logistic function

Support Vector Machine [28]: SVM is a supervised learning algorithm used for classification tasks. It finds an optimal hyperplane that separates data points of different classes. It works well with high-dimensional data.

E. Performance metrics

We use six measures, including accuracy (Acc), precision (Pre), Specificity (Sp), Recall, F1-score and area under the curve (AUC), for the performance estimations of four ML models. Acc measures the proportion of correctly classified cases out of all subjects. Pre reflects the model's effectiveness in identifying only relevant instances among those retrieved. While Sp measures the proportion of actual negative cases correctly identified by the model, indicating its ability to correctly classify non-heart failure cases. Recall, also known as sensitivity, represents the proportion of actual positive cases correctly, reflecting its ability to detect heart failure cases. The F1-score is the harmonic mean of precision and recall, balancing both metrics to provide a single performance measure. AUC of the receiver operating characteristic curve quantifies the model's ability to distinguish between heart failure and non-heart failure cases.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Pre = \frac{TP}{TP + FP} \quad (2)$$

$$Sp = \frac{TN}{FP + TN} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1\text{-}score = 2\frac{Pre \times Recall}{Pre + Recall} \quad (5)$$

Where TP (True Positive) represents the test result that correctly identifies patients with heart failure; TN (True Negative) represents the number of correctly identified subjects without heart failure disease; FP and FN represent cases where the test incorrectly predicts heart failure and non-heart failure, respectively.

III. SIMULATION RESULTS

A. Identification of DEGs

After using an R library package to analyze differential expression with filtered parameters of |logFC| > 0.6, FDR < 0.05, and p-value < 0.05, the differential expression analysis results of the GEO dataset reveals 153 DEGs, of which 81 genes are down-regulated and 72 genes are up-regulated, as shown in Figure 2.

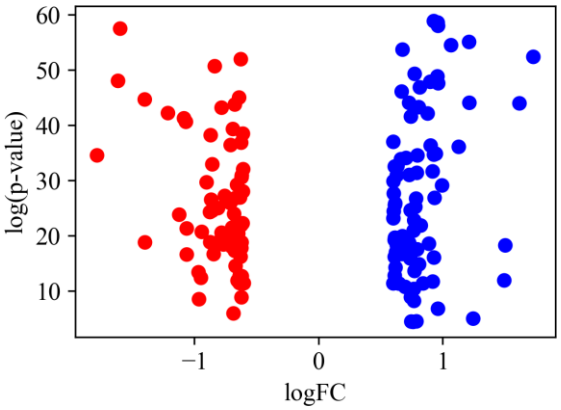

Figure 2: The scatter plot of log(p-value) and logFC, where red color represents up-regulated and blue color represent down-regulated

B. Gene selection

We implement the PSO algorithm in combination with the RF model to identify key genes from a set of 153 DEGs. Through this process, 10 optimal genes are selected. These genes are ECM2, ASPN, PTN, SFRP4, FCN3, TEAD4, NPTX2, LAD1, ALOX5AP, and RNASE2.

C. Model optimization

By applying grid search with 5-fold cross-validation, we obtain four optimal ML models including RF, SVM, XGboost and LR. The optimal hyperparameters for the RF model included *n_estimators=78*, *max_depth=50*, and *min_samples_split= 2*, *min_samples_leaf=4*. For the SVM, the best configuration used an RBF kernel with *C=1.0* and *gamma=0.1*. The XGBoost model achieved optimal performance with *n_estimators=100*, *learning_rate=0.01*, *max_depth=6*, and *subsample=0.8*, improving both stability and generalization. The LR model performed best with *C=2.0, max_iter=100,* and *penalty='l2'*.

D. Model validation

To evaluate the performance of ML models on new data, we further validated them on the GSE57345 dataset with 5-fold CV, with results shown in Table 1. The RF model demonstrated the best performance. Therefore, we choose RF as the model for diagnosing heart failure.

Table 1: Performance of ML models on the validation set

	RF	**XGBoost**
Acc	**0.9192± 0.02**	0.8818± 0.04
Pre	**0.9407± 0.02**	0.8980± 0.04
Sp	**0.9178± 0.03**	0.8621± 0.06
Recall	**0.9209± 0.02**	0.8975± 0.03
F1-score	**0.9304± 0.01**	0.8798± 0.04
AUC	**0.9194± 0.01**	0.8974± 0.03
	LR	**SVM**
Acc	0.8632± 0.03	0.8725± 0.03
Pre	0.8872± 0.05	0.8980± 0.04
Sp	0.8467± 0.06	0.8597± 0.06
Recall	0.8767± 0.02	0.8827± 0.01
F1-score	0.8815± 0.04	0.8900± 0.03
AUC	0.8617± 0.04	0.8712± 0.03

IV. DISCUSSION

Genetic biomarkers are particularly promising in identifying preclinical stages of HF and providing personalized treatment options. In our research, we propose an approach that includes differential expression analysis, followed by the use of the PSO algorithm to identify potential genes for predicting heart failure. Four ML models are trained and tested to determine the best-performing model.

DEG analysis is an important related event in heart failure. Indeed, the analysis of gene expression data is beneficial for predicting heart failure patients [13]. This type of data offers a wealth of information that can be utilized to identify significant biomarkers and genetic pathways. Gene expression profiles are typically high-dimensional, with tens of thousands of genes and high correlations between them. DEG analysis tools are useful for identifying biologically significant genes with rich information. By using this method, we obtain 153 DEGs with 72 up-regulated genes and 81 down-regulated genes. The up-regulated 72 genes are primarily involved in muscle system processes, extracellular matrix organization, extracellular structure organization, muscle contraction, and cell-substrate adhesion. 81 down-regulated genes are primarily associated with the positive regulation of vascular development, angiogenesis, neutrophil activation, L-amino acid transport, and neutrophil-mediated immunity [15].

Although DEGs also provide a lot of useful information, there are still many redundant genes, so we continue to select genes to find the most potential genes for predicting heart failure. In order to explore the key DEGs in heart

failure, we use PSO algorithm with RF model to obtain 10 potential DEGs are selected, namely ECM2 (Extracellular Matrix Protein 2), ASPN (Asporin), PTN (Pleiotrophin), SFRP4 (Secreted Frizzled-Related Protein 4), FCN3 (Ficolin-3), TEAD4 (TEA Domain Transcription Factor 4), NPTX2 (Neuronal Pentraxin 2), LAD1 (Ladinin 1), ALOX5AP (Arachidonate 5-Lipoxygenase Activating Protein), and RNASE2 (Ribonuclease A Family Member 2). The significance of these genes is shown in Table

Table 2: Explanation of the effects of the 10 genes.

Gene names	Explanation
ECM2 [20]	It is associated with immune processes
ASPN [29]	ASPN expression is induced in response to cardiac pressure overload or ischemia-reperfusion.
PTN [30]	It derived from cardiac fibroblasts may play potential role in pressure overload-induced hypertrophic cardiomyopathy through activating the PTN-SDC4 pathway in cardiac fibroblasts and macrophages.
SFRP4 [31]	The expression of SFRP4 in ventricular myocardium correlates with apoptosis related gene expression.
FCN3 [32]	It is related to chronic heart failure
TEAD4 [33]	TEAD is a transcription factor involved in regulating gene expression related to cell proliferation and apoptosis. It is associated with coronary artery disease risk.
NPTX2 [34]	It is a Protein Coding gene. Diseases associated with NPTX2 include Narcolepsy and Diabetes Insipidus, Neurohypophyseal.
LAD1 [35]	Diseases associated with LAD1 include Epidermolysis Bullosa Acquisita and Cicatricial Pemphigoid
ALOX5AP [36]	ALOX5AP gene variants and risk of coronary artery disease
RNASE2 [37]	It is a possible trigger of acute-on-chronic inflammation leading to mRNA vaccine-associated cardiac complication

We fit the 10 key genes on four ML models. The result show that RF model achieve the best average accuracy and AUC with 91.92% and 91.94% on validation set, respectively. The AUC values are greater than 75% and the model also demonstrates high sensitivity and specificity, indicating that our diagnostic model is accurate, reliable, and unaffected by alterations in the cohort group. Notably, the reproducibility of our findings is corroborated by their consistency across separate datasets. Therefore, 10 key genes can be used as signature genes for predicting of heart failure. This discovery paves the way for further exploration of crucial mechanisms in heart failure. Indeed, in a previous study [30] it was demonstrated that normal turnover of the extracellular matrix (ECM) is regulated by the balance between matrix metalloproteinases (MMPs) and their tissue inhibitors (TIMPs). This balance is altered in heart failure. MMPs, TIMPs, and ECM degradation products have been investigated as potential diagnostic and prognostic biomarkers for heart failure [38].

The comparison between our proposed method and existing methods is presented in Table 3. The result shows that the performance of our method is better than the performance in the study [20]. Although the method presented in [20] achieved higher accuracy than our results. However, Pre, Recall and F1-score have lower results. This results indicate that their proposed method did not correctly identify many cases of patients with heart failure. In additions, our model also achieved a higher AUC, reflecting its ability to discriminate between heart failure and non-heart failure cases. Besides, our results used 5-fold CV method, demonstrating the robustness and generalizability of our model across different data partitions. The authors in [20] proposed a heart failure model consisting of 16 characteristic genes (ECM2, LUM, ISLR, ASPN, PTN, SFRP4, GLT8D2, FRZB, FCN3, TEAD4, NPTX2, LAD1, ALOX5AP, RNASE2, IL1RL1, CD163) was constructed using machine learning and artificial intelligence. Although they performed enrichment analyses and explored immune cell infiltration, their approach did not involve a structured optimization of model parameters. While we propose RF model using 10 potential DEGs are selected, namely ECM2, ASPN, PTN, SFRP4, FCN3, TEAD4, NPTX2, LAD1, ALOX5AP, and RNASE2 to diagnose heart failure. Our method consists of for main stages: gene expression analysis, gene selection using PSO, model optimization by grid search with 5-fold CV, and evaluation performance of model on the validation set. This multi-stage is designed to ensure optimal gene subset selection and robust model performance.

Table 3: Comparison of our proposed method with existing work on the GSE57345 dataset

	Acc	Pre	Recall	F1-score	AUC
Our	0.9192	0.9407	0.9209	0.9304	0.9194
[20]	0.995	0.893	0.826	0.842	0.863

V. CONCLUSIONS

Early diagnosis and treatment of patients are very important to reduce the morbidity and mortality of heart failure patients. Therefore, in this study we propose an accurate, efficient and reliable model for heart failure prediction. It consists of an RF model with 10 biomarker genes. Despite the differences between patient groups, our model is still effective and can be used for personalized disease prediction and precision medicine. Our results demonstrated this with Acc, Pre, Sp, Recall, F1-score, AUC of 91.92%, 94.07%, 91.78%, 92.09%, 93.04%, and 91.94% on validation, respectively.

REFERENCES

[1] James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., ... & Briggs, A. M. (2018). Global, regional, and national incidence, prevalence, and years

lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789-1858.

[2] Savarese, G., Becher, P. M., Lund, L. H., Seferovic, P., Rosano, G. M., & Coats, A. J. (2022). Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovascular research*, 118(17), 3272-3287.

[3] Moreno-Sánchez, P. A. (2023). Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *Frontiers in cardiovascular medicine*, *10*, 1219586.

[4] Figueiral, M., Paldino, A., Fazzini, L., & Pereira, N. L. (2024). Genetic Biomarkers in Heart Failure: From Gene Panels to Polygenic Risk Scores. *Current heart failure reports*, 1-16.

[5] Shrivastava, A., Haase, T., Zeller, T., & Schulte, C. (2020). Biomarkers for heart failure prognosis: proteins, genetic scores and non-coding RNAs. *Frontiers in cardiovascular medicine*, *7*, 601364.

[6] Zhang, K., Qin, X., Wen, P., Wu, Y., & Zhuang, J. (2021). Systematic analysis of molecular mechanisms of heart failure through the pathway and network-based approach. *Life Sciences*, *265*, 118830.

[7] Yu, Y. D., Xue, Y. T., & Li, Y. (2023). Identification and verification of feature biomarkers associated in heart failure by bioinformatics analysis. *Scientific Reports*, *13*(1), 3488.

[8] Rogers, R. K., Stehlik, J., Stoddard, G. J., Greene, T., Collins, S. P., Peacock, W. F., ... & Michaels, A. D. (2009). Adjusting for clinical covariates improves the ability of B-type natriuretic peptide to distinguish cardiac from non-cardiac dyspnoea: a sub-study of HEARD-IT. *European journal of heart failure*, *11*(11), 1043-1049.

[9] Roberts, E., Ludman, A. J., Dworzynski, K., Al-Mohammad, A., Cowie, M. R., McMurray, J. J., & Mant, J. (2015). The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. *Bmj*, *350*.

[10] Dai, Z., Zhang, Y., Ye, H., Zhang, G., Jin, H., Chen, Z., ... & Zhang, Z. (2018). Adiponectin is valuable in the diagnosis of acute heart failure with renal insufficiency. *Experimental and therapeutic medicine*, *16*(3), 2725-2734.

[11] Yasmin, F., Shah, S. M. I., Naeem, A., Shujauddin, S. M., Jabeen, A., Kazmi, S., ... & Lak, H. M. (2021). Artificial intelligence in the diagnosis and detection of heart failure: the past, present, and future. *Reviews in cardiovascular medicine*, *22*(4), 1095-1113.

[12] Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer letters*, *471*, 61-71.

[13] Yu, Y., Wang, L., Hou, W., Xue, Y., Liu, X., & Li, Y. (2024). Identification and validation of aging-related genes in heart failure based on multiple machine learning algorithms. *Frontiers in Immunology*, *15*, 1367235.

[14] Yu, Y., Liu, X., Xue, Y., & Li, Y. (2023). Identification of immune-related genes for the diagnosis of ischemic heart failure based on bioinformatics. *Iscience*, *26*(11).

[15] Visco, V., Robustelli, A., Loria, F., Rispoli, A., Palmieri, F., Bramanti, A., ... & D'Angelo, G. (2024). An explainable model for predicting Worsening Heart Failure based on genetic programming. *Computers in Biology and Medicine*, *182*, 109110.

[16] Berulava, T., Buchholz, E., Elerdashvili, V., Pena, T., Islam, M. R., Lbik, D., ... & Toischer, K. (2020). Changes in m6A RNA methylation contribute to heart failure progression by modulating translation. *European journal of heart failure*, *22*(1), 54-66.

[17] Smith, A. B., Jung, M., Pressler, S. J., Mocci, E., & Dorsey, S. G. (2023). Differential gene expression among patients with heart failure experiencing pain. *Nursing research*, *72*(3), 175-184.

[18] Hahn, V. S., Knutsdottir, H., Luo, X., Bedi, K., Margulies, K. B., Haldar, S. M., ... & Sharma, K. (2021). Myocardial gene expression signatures in human heart failure with preserved ejection fraction. *Circulation*, *143*(2), 120-134.

[19] Wang, X., Rao, J., Zhang, L., Liu, X., & Zhang, Y. (2024). Identification of circadian rhythm-related gene classification patterns and immune infiltration analysis in heart failure based on machine learning. *Heliyon*, *10*(6).

[20] Chen, Y., Xue, J., Yan, X., Fang, D. G., Li, F., Tian, X., ... & Feng, Z. (2023). Identification of crucial genes related to heart failure based on GEO database. *BMC Cardiovascular Disorders*, *23*(1), 376.

[21] Hannenhalli, S., Putt, M. E., Gilmore, J. M., Wang, J., Parmacek, M. S., Epstein, J. A., ... & Cappola, T. P. (2006). Transcriptional genomics associates FOX transcription factors with human heart failure. *Circulation*, *114*(12), 1269-1276.

[22] Barth, A. S., Kuner, R., Buness, A., Ruschhaupt, M., Merk, S., Zwermann, L., ... & Sültmann, H. (2006). Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *Journal of the American College of Cardiology*, *48*(8), 1610-1617.

[23] Liu, Y., Morley, M., Brandimarto, J., Hannenhalli, S., Hu, Y., Ashley, E. A., ... & Li, M. (2015). RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics*, *105*(2), 83-89.

[24] Thakkar, H. K., Shukla, H., & Sahoo, P. K. (2022). Metaheuristics in classification, clustering, and frequent pattern mining. In *Cognitive big data intelligence with a metaheuristic approach* (pp. 21-70). Academic Press.

[25] Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In *International conference on information computing and applications* (pp. 246-252). Berlin, Heidelberg: Springer Berlin Heidelberg.

[26] Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Frontiers in genetics*, *10*, 1077.

[27] Bisong, E., & Bisong, E. (2019). Logistic regression. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, 243-250.

[28] Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207-235.

[29] Huang, C., Sharma, A., Thakur, R., Rai, D., Katiki, M., de Freitas Germano, J., ... & Piplani, H. (2022). Asporin, an extracellular matrix protein, is a beneficial regulator of cardiac remodeling. *Matrix Biology*, *110*, 40-59.

[30] Sheng, K., Ran, Y., Feng, X., Wang, Y., Zhou, S., Guan, Y., ... & Guo, X. (2025). PTN secreted by cardiac fibroblasts promotes myocardial fibrosis and inflammation of pressure overload-induced hypertrophic cardiomyopathy through the PTN-SDC4 pathway. *Life Sciences*, *363*, 123356.

[31] SFRP4 Gene - Secreted Frizzled Related Protein 4. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SFRP4

[32] Li, D., Lin, H., & Li, L. (2020). Multiple feature selection strategies identified novel cardiac gene expression signature for heart failure. *Frontiers in Physiology*, *11*, 604241.

[33] Almontashiri, N. A., Antoine, D., Zhou, X., Vilmundarson, R. O., Zhang, S. X., Hao, K. N., ... & Stewart, A. F. (2015). 9p21. 3 Coronary Artery Disease Risk Variants Disrupt

TEAD Transcription Factor–Dependent Transforming Growth Factor β Regulation of p16 Expression in Human Aortic Smooth Muscle Cells. *Circulation*, *132*(21), 1969-1978.
[34] NPTX2 Gene - Neuronal Pentraxin 2. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NPTX2
[35] LAD1 Gene - Ladinin 1https://www.genecards.org/cgi-bin/carddisp.pl?gene=LAD1
[36] Girelli, D., Martinelli, N., Trabetti, E., Olivieri, O., Cavallari, U., Malerba, G., ... & Corrocher, R. (2007). ALOX5AP gene variants and risk of coronary artery disease: an angiography-based study. *European journal of human genetics*, *15*(9), 959-966.
[37] Ong, E. Z., Koh, C. W., Tng, D. J., Ooi, J. S., Yee, J. X., Chew, V. S., ... & Ooi, E. E. (2023). RNase2 is a possible trigger of acute-on-chronic inflammation leading to mRNA vaccine-associated cardiac complication. *Med*, *4*(6), 353-360.
[38] Castiglione, V., Aimo, A., Vergaro, G., Saccaro, L., Passino, C., & Emdin, M. (2022). Biomarkers for the diagnosis and management of heart failure. *Heart failure reviews*, 1-19.

DỰ ĐOÁN SUY TIM DỰA TRÊN HỌC MÁY SỬ DỤNG DỮ LIỆU DI GENE

Tóm tắt: Hiện nay suy tim là một vấn đề đáng quan tâm lớn trên toàn cầu, nó ảnh hưởng đến hàng triệu người. Bệnh này có đặc điểm là tỷ lệ tử vong cao và gánh nặng kinh tế đáng kể. Do đó, trong nghiên cứu này, chúng tôi đề xuất một mô hình có độ chính xác cao, nhanh chóng và kịp thời để chẩn đoán suy tim tiền lâm sàng dựa trên các dấu ấn sinh học di truyền. Mô hình này bao gồm bộ phân loại Rừng ngẫu nhiên (RF) và 10 gen biểu hiện khác biệt được chọn bằng thuật toán tối ưu hóa bầy hạt (PSO). Kết quả của chúng tôi đã chứng minh tính hiệu quả của nó, với độ chính xác (Acc), độ chính xác (Pre), độ đặc hiệu (Sp), độ thu hồi, điểm F1 và AUC đạt lần lượt là 91.92%, 94.07%, 91.78%, 92.09%, 93.04%, và 91.94% trên tập dữ liệu GSE57345 với xác thực chéo 5 lần. Những phát hiện này chỉ ra rằng, mặc dù có sự khác biệt giữa các nhóm bệnh nhân, mô hình của chúng tôi vẫn rất hiệu quả và có thể được áp dụng để dự đoán bệnh cá nhân hóa và y học chính xác.

Từ khóa: Suy tim, học máy, chọn lọc gen, biểu hiện gen tổng hợp, gen biểu hiện khác biệt.

Tuan Anh Vu received the degree of Engineer in Information Technology and the M.S degree in Information Systems from the Post & Telecommunications Institute of Technology (PTIT), Hanoi, Viet Nam, in 2016 and 2018. His research interests include machine learning, deep learning, optimization, and bigdata. Email: vtanh@ptit.edu.vn

Le Anh Dang Tran received the B.S. degree in Electronics Telecommunication Engineering, and the M.S degree in Information Systems from Post and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam, in 2014 and 2019, respectively. He is currently a lecturer at the Department of Data Engineering, Faculty of Telecommunications1, PTIT. His research interests include network security, the Internet of Things, machine learning, deep learning, and brain-computer interface. Email: anhdtl@ptit.edu.vn

Minh Tuan Nguyen received the B.S. degree from the Post & Telecommunications Institute of Technology, Hanoi, Vietnam, in 2004, the M.S. degree from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2008, both in electronics and telecommunications engineering, and the Ph.D. degree at the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2018. He is with Posts and Telecommunications Institute of Technology. His research interests include network security, internet of things, biomedical signal processing, gene analysis, sentiment analysis, brain computer interface, machine learning, deep learning, optimization, and biomedical application design.