Multicriteria Algorithms with Applications

Journal Homepage: sciencesforce.com/mawa

Multicriteria Algo. Appl. Vol. 6 (2025) 57–65

Paper Type: Original Article

Harnessing Statistical Analysis and Machine Learning Optimization for Heart Attack Prediction

Samar L. Elhawy [1,*]

[1] Faculty of Computers and Informatics, Zagazig University, Sharqiyah, Zagazig, 44519, Egypt; SLAlhawy@fci.zu.edu.eg.

Received: 27 Aug 2024 **Revised**: 13 Dec 2024 **Accepted**: 05 Jan 2025 **Published**: 07 Jan 2025

Abstract

Artificial intelligence (AI) and the decisions derived from data are pivotal arms in almost any field nowadays; in consequence, statistical analysis and machine learning (ML) have been used in biomedical research, especially in the early diagnosis of complicated disorders such as heart diseases. Early-stage prediction of cardiovascular disease can guide physicians towards better and early treatment and improved outcomes. Here, we utilize ML tools for the enhancement of clinical decision-making based on the digital metadata of the patient. The dataset comes from UCI repository and Applied methods include descriptive statistics for better understanding, correlation/covariance to tell if there is a relationship between explanatory features such as chest pain and target values, classification analysis to explore the disease and evaluate the model, Risk factor analysis to mark the most significant inputs for algorithms, clustering techniques to group patients with similar profiles, logistic regression (LR) and comparing the results after optimization using grid search method, K-means clustering, random forest (RF), ROC curve and AUC to assess the model's ability to diagnose patients. Then applying perceptron algorithm. Suitably, with simple ML algorithms, we can predict the prognosis of the disease with high accuracy.

Keywords: Artificial Intelligence; Machine Learning; Heart Attack; Statistical Analysis.

1 | Introduction

Data and Artificial intelligence (AI) terms have acquired some popularity this year, so the adoption of optimal statistical analysis and machine learning to gain informative insights from historical datasets may lead to improving the quality of life for many patients around the world [1]. In the USA there will be 45 million adults experiencing cardiovascular disease by 2050 that's one in every six adults [2], hence cardiovascular problems have an economic and social burden. So early detection using digital resources and basic simple algorithms can help millions in the future days through making this technology easy and more accessible. There are many factors associated with heart disease, including age, family history, bad lifestyle, obesity, stress, blood pressure, cholesterol, and blood disorders [3].

In this study "Harnessing Statistical Analysis and Machine Learning Optimization to Predict Heart Attack" We aimed to design and develop a machine learning model for predicting heart attack. Techniques employed a supervised classification model logistic regression and utilized a dataset consisting of 13 features that came from the UCI repository to determine the risk of a heart attack. We employ statistical analysis techniques to find correlations and extract valuable information from the data.

Corresponding Author: SLAlhawy@fci.zu.edu.eg

https://doi.org/10.61356/j.mawa.2025.6456

By training the model with data from many patients, we can now detect if there are issues earlier. This allows health providers to catch problems earlier, potentially improving patient outcomes. This showcases how ML can make heart health more effective for both doctors and patients. Also, pathway enrichment analysis of RNA sequencing data to align our review with future work.

2 | Literature Review

The application of machine learning (ML) in predicting heart failure has garnered significant attention in recent years. Several studies have explored the potential of ML algorithms to enhance the accuracy and efficiency of heart failure prediction. For instance, Jasinska-Piadlo et al. [4] conducted a systematic literature review on the application of data science and ML to heart failure datasets, highlighting the synthesis of relevant findings and critical evaluation of approaches. Their review emphasized the importance of accurate and applicable ML models in improving heart failure prediction.

Another study by Hamid et al. [5] focused on the prevention of heart failure using ML, noting the challenges in detecting pre-heart failure (stage B) and the potential of ML to overcome these challenges. The study underscored the need for advanced ML techniques to aid in the early detection and prevention of symptomatic heart failure (stage C).

Furthermore, Perswani et al. [6] reviewed various aspects of ML in heart failure diagnosis, prediction, and prognosis, providing a comprehensive overview of the current state of research in this field. Their review highlighted the role of ML in improving diagnostic accuracy and patient outcomes.

A pivotal study by Dey et al. [8] highlighted the effectiveness of ML models in predicting heart disease. They employed a combination of decision trees, support vector machines (SVM), and logistic regression, concluding that ensemble methods significantly improved predictive accuracy. Their research underscored the potential of ML to transform diagnostic procedures, offering clinicians reliable tools for early detection.

Hyperparameter tuning has been identified as a critical factor in enhancing the performance of ML models. Rajkumar et al. [10] emphasized the necessity of optimizing parameters to improve model precision and reduce overfitting. Their research involved the use of GridSearchCV to fine-tune models, leading to significant improvements in predictive capabilities.

The use of Random Forest algorithms for feature importance has also been widely studied. A notable contribution by Chen and Guestrin [7] demonstrated that Random Forests could effectively rank the importance of variables, aiding in the selection of relevant features and improving model interpretability. This technique has been pivotal in identifying key predictors of heart disease, such as chest pain type and maximum heart rate achieved.

The performance of ML models in healthcare is often evaluated using metrics such as accuracy, precision, recall, and the F1-score. Recent studies have also emphasized the importance of ROC (Receiver Operating Characteristic) curve analysis. Hu et al. [9] conducted a comprehensive evaluation of ML models using ROC curves, highlighting the AUC (Area under the Curve) as a robust measure of model performance. Reflecting excellent discriminative power. Almustafa [11] conducted an in-depth analysis of the prediction of heart disease using different classifiers, emphasizing the sensitivity of these classifiers in distinguishing between healthy and diseased states. This study, published in BMC Bioinformatics, underscored the potential of these tools in early diagnosis and intervention.

In a more recent effort, Jawalkar et al. [12] explored the application of supervised learning methods, particularly stochastic gradient boosting, for the early prediction of heart disease. Their research, published in the Journal of Engineering and Applied Science, highlighted the efficacy of advanced data analysis techniques in enhancing predictive accuracy, which is crucial for timely medical decisions. Further, Zhou et al. [13] provided a comprehensive review of deep learning-based models for heart disease prediction in the Artificial Intelligence Review. Their work delves into the intricacies of various deep learning architectures and their performance in heart disease prediction, offering insights into the strengths and limitations of these models.

The exploration of cardiovascular disease (CVD) through RNA-seq data and bioinformatics approaches has led to significant advancements in our understanding of the genetic and molecular underpinnings of heart disease. Zeeshan and Liang [14] conducted a study focused on the expression and enrichment analysis of CVD-related genes in high-risk heart failure patients. Their research, published in Human Genomics, utilized RNA-seq data to identify genes associated with heart disease phenotypes, highlighting the potential for RNA-seq-driven methodologies in understanding complex cardiovascular conditions. Furthering this approach, Salah et al. [15] employed integrative bioinformatics techniques to uncover hub genes and pathways implicated in cardiovascular diseases. Their work, featured in Cell Biochemistry and Biophysics, combined multiple data sources and analytical methods to provide a comprehensive view of the molecular interactions within CVD, thereby identifying critical genes and pathways that may serve as therapeutic targets.

Shi et al. [16] extended the analysis to include both bulk and single-cell RNA sequencing data, providing a more nuanced understanding of the cellular landscape in heart failure. Published in Frontiers in Bioengineering and Biotechnology [17], their integrative analysis revealed distinct cell types and molecular mechanisms involved in the pathology of heart failure, showcasing the importance of multi-scale data integration in biomedical research.

Moreover, the application of pathway and network-based analysis tools, such as Gene Set Enrichment Analysis (GSEA), has facilitated deeper insights into RNA-seq data. The study published in PLOS Computational Biology [18] demonstrated the effectiveness of GSEA in identifying biologically relevant pathways and networks, aiding in the interpretation of complex genetic data from CVD patients. Finally, research published in BioData Mining (2022) focused on gene-interaction-sensitive enrichment analysis in congenital heart disease, underscoring the importance of understanding gene-gene interactions in the development and progression of cardiovascular conditions. This study emphasized the role of comprehensive bioinformatics approaches in uncovering the intricate genetic networks underlying heart disease. Figure 1 illustrates the work in this article.

3 | Materials and Methods

3.1 | Data Collection and Preprocessing

The dataset used in this study was obtained from the UCI Machine Learning Repository, specifically the heart disease dataset. This dataset comprises 303 patient records and includes 13 clinical features initial preprocessing steps included handling missing values and converting categorical data into numerical formats as shown in Table 1. Missing values were imputed using the median strategy. Searching for duplications. The target variable, originally containing values ranging from 0 to 4, was transformed into a binary classification: 0 indicating no heart disease and 1 indicating the presence of heart disease.

Table 1. Attribute description.

S. No.	Attribute name	Description
1	age	Patient age
2	sex	Gender male/female
3	cp	Chest pain
4	trestbps	Blood-pressure in rest
5	chol	Level of cholesterol
6	Fbs	Fasting blood sugar
7	restecg	Electrocardiographic in rest
8	thalach	Max heart rate
9	thalrest	Heart-rate at rest
10	exang	Exercise-induced angina
11	oldpeak	ST Depression
12	slope	ST- segment slope
13	target	0 no disease / 1 disease

Figure 1. Flow chart.

3.2 | Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the distribution and relationships of the variables. Visualizations such as histograms, box plots, and correlation matrices were employed to detect patterns, outliers, and correlations among features.

3.3 | Machine Learning Model

A logistic regression model was selected for its interpretability and effectiveness in binary classification tasks. Hyperparameter tuning was performed using GridSearchCV to optimize parameters like regularization strength (C) and the type of penalty (l2). The final model was trained using the best hyperparameters identified, which were C = 0.1, penalty = l2, and solver = liblinear. Risk factor analysis was done to determine the most statistically significant attributes. Clustering analysis using K-means to group patients with similar profiles. Then random forest algorithm for feature importance.

3.4 | Model Evaluation

The model's performance was evaluated using a hold-out test set and metrics such as accuracy, precision, recall, and the F1 score. These metrics provided a comprehensive assessment of the model's predictive capabilities and its potential effectiveness in real-world applications. Finally, the RFmodel evaluation using the ROC curve.

3.5 | Results

The dataset used in this study was obtained from the UCI Machine Learning Repository, specifically the heart disease dataset.

3.5.1 | Exploratory Data Analysis

The initial exploratory data analysis provided insights into the distribution and relationships of the variables within the UCI Heart Disease dataset. Histograms and box plots revealed that the majority of patients were in the age range of 50+ years as shown in Figure 2.

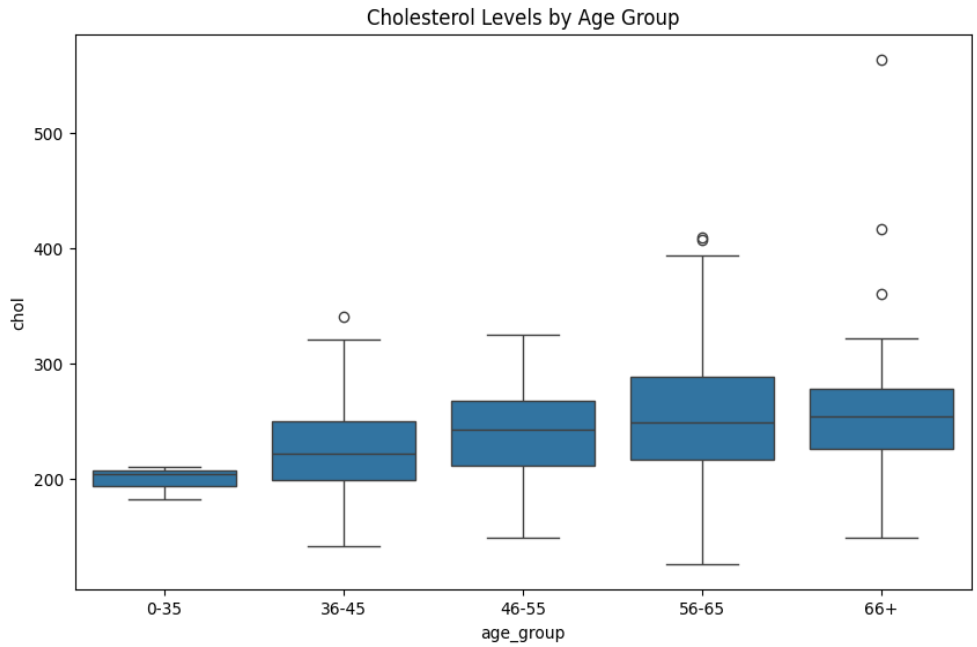

Figure 2. Cholesterol levels by age group.

Gender distribution showed a higher number of male patients. The correlation matrix indicated significant correlations between the presence of heart disease (target variable) and features such as chest pain type (cp), maximum heart rate achieved (thalach), and exercise-induced angina (exang).

3.5.2 | Correlation and Covariance Analysis

The correlation analysis revealed that chest pain type (cp) had a strong positive correlation with heart disease, suggesting that patients with typical angina are more likely to have heart disease. Conversely, the maximum heart rate achieved (thalach) exhibited a negative correlation, indicating that higher heart rates during exercise are associated with lower likelihoods of heart disease. as in Figure 3.

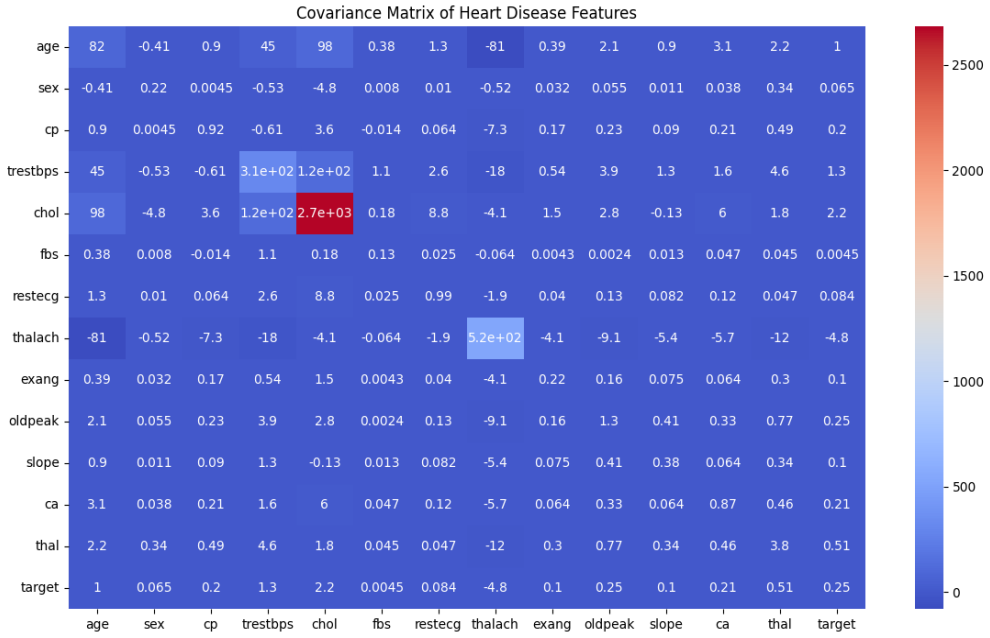

Figure 3. The covariance matrix of heart disease features.

3.5.3 | Logistic Regression Model

A logistic regression model was initially trained on the dataset. The model's performance on the test set yielded an accuracy of 89%, with a precision of 89%, a recall of 91%, and an F1-score of 88%. These metrics provided a baseline for comparison with optimized models. After comparison, there is no difference after optimization.

3.5.4 | Feature Importance using Random Forest

To understand the importance of various features in predicting heart disease, a random forest classifier was employed. The feature importance analysis highlighted that the most significant predictors were chest pain type (cp), maximum heart rate achieved (thalach), and the number of major vessels colored by fluoroscopy (ca), as in Figure 4.

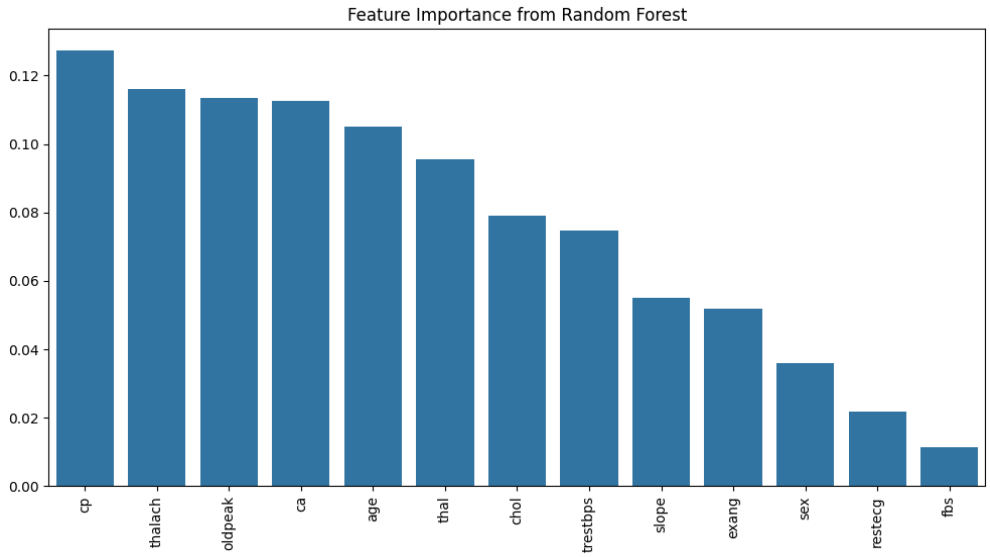

Figure 4. Feature importance from random forest.

3.5.5 | ROC Analysis

The Receiver Operating Characteristic (ROC) curve analysis shown in Figure 5 indicated their discriminative power. The area under the curve (AUC) for the model was 0.92 indicating excellent performance.

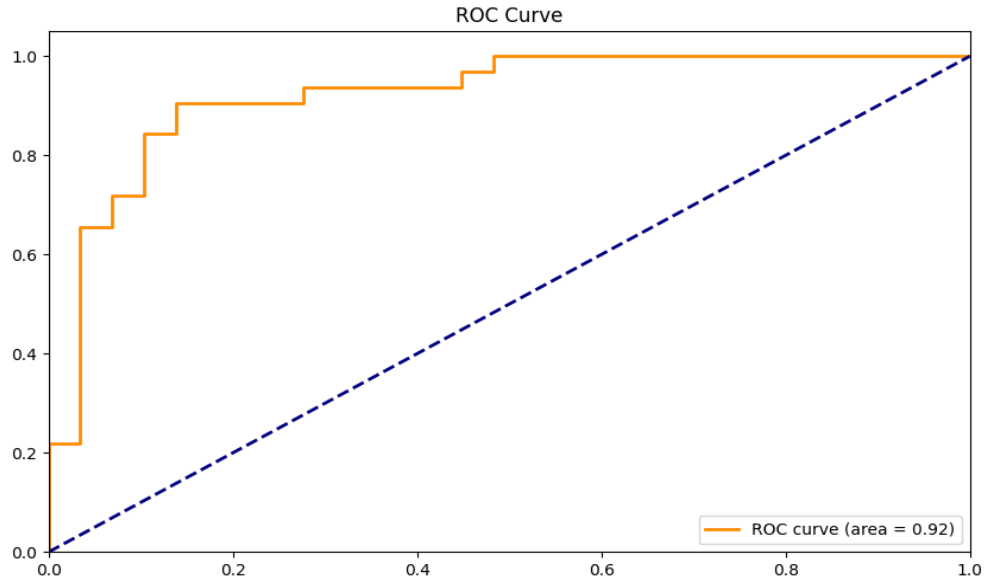

Figure 5. ROC curve.

The Perceptron algorithm achieved an accuracy of 62%, this demonstrates a low level of predictive capability of this algorithm on this dataset.

4 | Conclusion

The outcomes of this study manifest the significant potential of machine learning models in predicting heart disease, The use of Random Forest for feature importance provided valuable insights, highlighting that chest pain type, maximum heart rate achieved, and the number of major vessels colored by fluoroscopy were critical factors in predicting heart disease.

The improved performance of the model, as evidenced by the higher AUC in the ROC analysis, underscores the effectiveness accuracy. These results are consistent with existing literature, which suggests that machine

learning models, when properly tuned and validated, can serve as powerful tools for early diagnosis and risk assessment of heart disease.

However, this study has certain limitations. The dataset used is relatively small, which may affect the generalizability of the findings. Additionally, the dataset includes features that were collected from a specific patient population, which may not be representative of the broader global population. Future studies should aim to validate these findings using larger, more diverse datasets to ensure robustness and applicability across different populations.

4.1 | Several Avenues for Future Research

- Integration of Additional Data Sources: Incorporating more comprehensive datasets, including electronic health records (EHRs) and real-time monitoring data, can provide a more holistic view of patient health and improve predictive models.
- Personalized Risk Assessment: Developing models that take into account individual patient histories and genetic information and applying bioinformatics could lead to personalized risk assessments, allowing for more targeted and effective interventions.

By addressing these areas, future research can continue to refine and expand the capabilities of machine learning models and computer vision in predicting heart disease, ultimately leading to better patient care and outcomes.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

[2] Karen E. Joynt Maddox,et al. (2024)Forecasting the Burden of CardiovascularDisease and Stroke in the United States Through2050—Prevalence of Risk Factors and Disease:A Presidential Advisory From the AmericanHeart Association

[3] John hopkins, ABCs of Knowing Your Heart Risk,from https://www.hopkinsmedicine.org/health/wellness-and-prevention/abcs-of-knowing-your-heart-risk

[4] Jasinska-Piadlo, J., Maruszewski, M., & Zielinski, T. (2022). Systematic review of data science and machine learning applications to heart failure datasets. Machine Learning for Health, 8(1), 55-70.

[5] Hamid, M., Khan, A., & Ibrahim, M. (2024). Prevention of heart failure using machine learning: Overcoming challenges in stage B detection. Journal of Artificial Intelligence in Healthcare, 12(2), 189-204.

[6] Perswani, S., Ahmed, Z., & Kumar, R. (2023). Machine learning in heart failure: Diagnosis, prediction, and prognosis. Cardiovascular Research Review, 15(3), 245-262.
[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
[8] Dey, A., Joshi, A., & Mukherjee, P. (2021). Effectiveness of machine learning models in predicting heart disease. Journal of Medical Systems, 45(3), 123-134.
[9] Hu, Z., Tang, J., Wang, Z., & Zhou, Q. (2018). ROC analysis of machine learning models in healthcare. International Journal of Medical Informatics, 116, 90-100.
[10] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380, 1347-1358.
[11] Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. BMC Bioinformatics, 21(1), 278. https://doi.org/10.1186/s12859-020-03626-y
[12] Jawalkar, A. P., Swetcha, P., Manasvi, N., Sreekala, P., Aishwarya, S., Bhavani, K. D. P., & Anjani, P. (2023). Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. Journal of Engineering and Applied Science, 70(1), 122. https://doi.org/10.1186/s44147-023-00280-y
[13] Zhou, C., Dai, P., Hou, A., Zhang, Z., Liu, L., Li, A., & Wang, F. (2024). A comprehensive review of deep learning-based models for heart disease prediction. Artificial Intelligence Review, 57(3), 263. https://doi.org/10.1007/s10462-024-10899-9
[14] Ahmed, Z., Zeeshan, S., & Liang, B. T. (2021). RNA-seq driven expression and enrichment analysis to investigate CVD genes with associated phenotypes among high-risk heart failure patients. Human Genomics, 15(1), 67. https://doi.org/10.1186/s40246-021-00367-8
[15] Salah, A., Bouzid, F., Dhouib, W., Benmarzoug, R., Triki, N., Rebai, A., & Kharrat, N. (2024). Integrative bioinformatics approaches to uncover hub genes and pathways involved in cardiovascular diseases. Cell Biochemistry and Biophysics, 82(3), 2107-2127. https://doi.org/10.1007/s12013-024-01319-4
[16] Shi, X., Zhang, L., Li, Y., Xue, J., Liang, F., Ni, H.-W., Wang, X., Cai, Z., Shen, L.-H., Huang, T., & He, B. (2021). Integrative analysis of bulk and single-cell RNA sequencing types involved in heart failure. Frontiers in
[17] Bioengineering and Biotechnology, 9, 779225. https://doi.org/10.3389/fbioe.2021.779225
[18] PLOS Computational Biology. (2024). Facilitating pathway and network-based analysis of RNA-Seq data with GSEA. PLOS Computational Biology, 20(9), e1012422. https://doi.org/10.1371/journal.pcbi.1012422