University of Houston Clear Lake

# REVIEW PAPER

***Title***: IDENTIFICATION OF BIOMARKERS IN HEART FAILURES USING BIOINFORMATICS AND MACHINE LEARNING.

***Prepared for BIOT 6838***


**Submitted by**:

***Last Name, First Name:*** Aadherlla Shashikanth Reddy

***Student id:*** 2008092

***E-mail:*** aadherllas7217@uhcl.edu


***Date:*** 11/30/2022

# Identification of Biomarkers in Heart Failures using Bioinformatics and Machine Learning.

**Shashikanth Reddy Aadherlla,**

M.S. in Biotechnology and Bioinformatics, University of Houston - Clear Lake, Texas, USA.

## Abstract:

***Background**:* Heart failure (HF), a diverse clinical illness that affects millions of individuals worldwide, is one of the most prevalent diseases today. HF is caused by heart overload and damage, and many healthcare professionals must perform an early diagnosis of the condition in order to protect their patients from the sickness and preserve lives. ***Goal:*** The purpose of this review article was to investigate active therapeutic compounds as well as to screen and validate potential biomarkers, such as hub genes implicated in developing HF. In addition, a novel machine learning (ML) strategy is being used in finding the specific target gene. ***Methods:*** In this review paper, firstly we reviewed several methods from different authors like differentially expresses gene (DEG) analysis, Screening of Hub Genes from the PPI Network, Gene set enrichment analysis (GSEA), Receiver operating characteristic (ROC) curve analysis, RT-PCR analysis and then we also tried to elucidate the Machine Learning approach to identify the biomarkers in heart failure. ***Results:*** After reviewing the preceding methods several hub genes are found to be as potential biomarkers for the heart failure. ***Conclusion:*** We consider that this review article provides an extensive overview of hug genes, significant pathways, and ML approaches relevant to HF using a variety of bioinformatics methods. These analysis findings provide us fresh perspectives on HF biomarkers and therapy options.

***Keywords:*** Heart failure (HF), biomarkers, hub genes, machine learning (ML), differentially expresses gene (DEG), PPI Network, Gene set enrichment analysis (GSEA), Receiver operating characteristic (ROC) curve, RT-PCR analysis.

## 1. Background:

Heart failure (HF) is a condition when the heart is unable to pump out the blood supply and venous return required for body tissue metabolism. According to data from 2020, there were 22.5 million HF patients globally, and their death rate might reach 50% [1-3]. Acute and chronic HF are the two categories used in clinical practice. Acute HF appears as significant myocardial damage and arrhythmia, but chronic HF has a gradual development and often presents as an enlarged or thickened heart. Cardiac failure (HF) may be brought on by cardiomyopathy, heart overload, myocardial inflammation, and other cardiovascular disorders, particularly in people with a history of coronary artery disease and hypertension [7-9]. According to the severity of the illness, several treatment modalities are often used, including surgery, traditional Chinese medicine, and medication therapies such RAAS inhibitors, -receptor antagonists, and nitrate medicines. However, only 50% of HF patients have a favorable prognosis with focused therapy, and many advanced patients still have a bad prognosis since HF is a progressive illness. Therefore, it is crucial to find efficient biomarkers for the therapy of HF [11-15].

The most recent technique is high-throughput gene microarray analysis, which can detect several chips simultaneously, minimizes system mistakes, and has extraordinarily high sensitivity [16-17]. The widespread application of this technique in illness research at the moment has created a turning point for human disease genomics research by enabling whole-genome resequencing [18]. Based on this technique, we are able to identify and investigate the molecular expression of the HF gene over the course of this HF study [18-20].

Due to recent high-throughput RNA sequencing data, which has been extensively used to screen the differentially expressed genes (DEGs) between normal samples and HF samples in humans, we can now further explore the entire molecular changes in HF at multiple levels involving DNA, RNA, proteins, epigenetic changes, and metabolism [13]. Due to the enormous amount of DEGs

discovered by using high throughput sequencing for expression profiling. and the too complex statistical analysis, there are still challenges in using this RNA sequencing data in the clinic [12-14].

We selected a number of datasets from Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) for our review [1-8]. Genes with varied levels of expression were compared between the samples (DEGs) [1]. Gene Set Enrichment Analysis (GSEA), which focuses on the enrichment of biological processes, was performed on the complete dataset [2]. The next thing that needed to be done was to use the Search Tool for the Retrieval of Interacting Genes, often known as STRING, in combination with Cytoscape in order to construct a Protein-Protein Interaction (PPI) network and remove the genes that operate as hubs [3]. We also looked at RT-PCR analysis and receiver operating characteristic (ROC) curve analysis of verified hub genes. Then, molecular docking experiments for overexpressed hub genes were analyzed [4]. Also, two machine learning techniques and important co-DEGs from a PPI network were screened. The investigation's findings may open up new perspectives on possible prognostic and therapeutic targets for HF [6].

## 2. Bioinformatics Approach:

### 2.1 Identification of DEGs in Heart failure (HF):

The limma program was used to find DEGs at the time of admission between patients with STEMI and patients with CAD. In patients with STEMI, DEGs between admission and six months after MI were found using the paired t test of the limma program [1]. Cutoff criteria were a fold change of a gene expression ratio > 1.0 and an adjusted P-value i.e., adjusted P value (0.05). The top DEGs were those genes with a factor change in gene expression ratio > 2.0. Using the limma program, DEGs between HF and non-HF at the four time points were also found. The threshold value was set at a fold change of a gene expression ratio > 1.0, $P < 0.05$. Using ggplot2 and p-heatmap, DEGs of patients who had STEMI upon admission were afterwards shown as volcano plots.

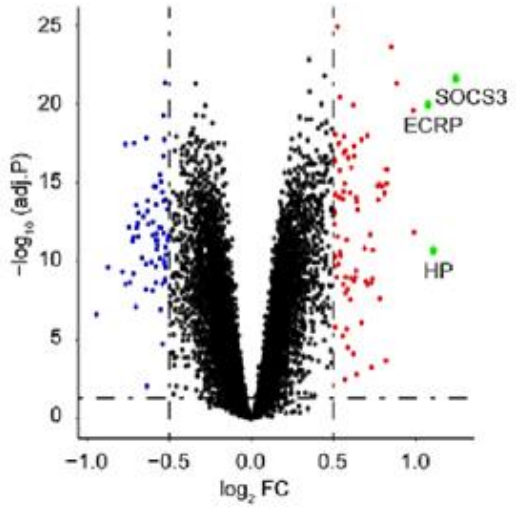



**FIGURE 1.** Volcano plot of mRNA expression profiles discriminating STEMI from CAD.
**[*Zhang, Jiajia et al. (2021)*]**

32,321 probes made up the raw data. After data preprocessing, 18822 genes were all that were left. At admission, 147 DEGs, comprising 79 upregulated and 68 downregulated genes, were tested between STEMI and CAD. The top DEGs were SOCS3, ERCP, and HP as seen in Figure 1 [1].

## 2.2 Gene Set Enrichment Analysis (GSEA):

A predetermined gene set is tested using the GSEA calculation technique to see whether there is a discernible difference between two biological states. In this investigation, we used GSEA to examine the enrichment of genes in 30 matched samples of VAD patients in BP and 8 groups of controls, using P <= 0.05 as the benchmark.

We utilized GSEA to examine the enrichment of genes in BP and used P < 0.05 as the selection criteria in order to investigate the function of the genes in the sample.

Figures demonstrate that genes involved in lipid digestion, fungal defense, intestinal lipid absorption, and positive modulation of potassium ion transmembrane transporter function were highly enriched in 68 samples [2].

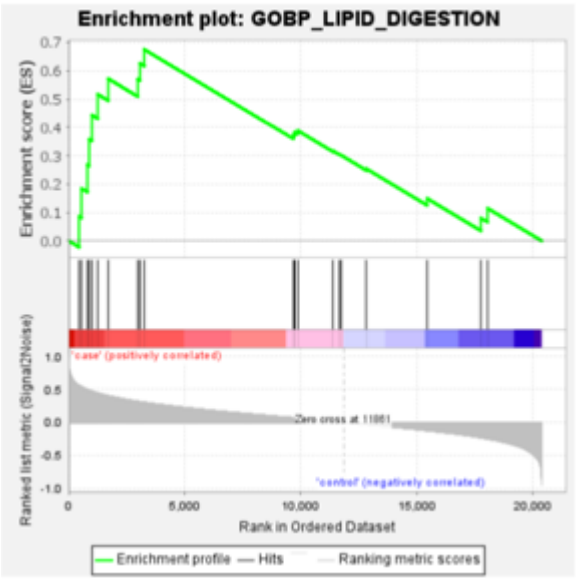
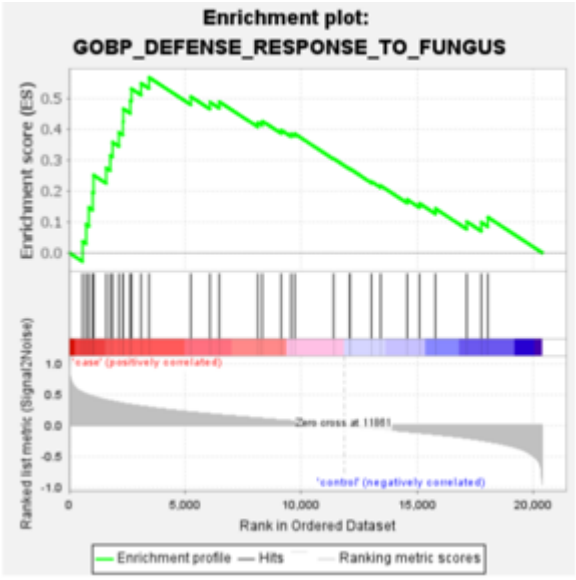
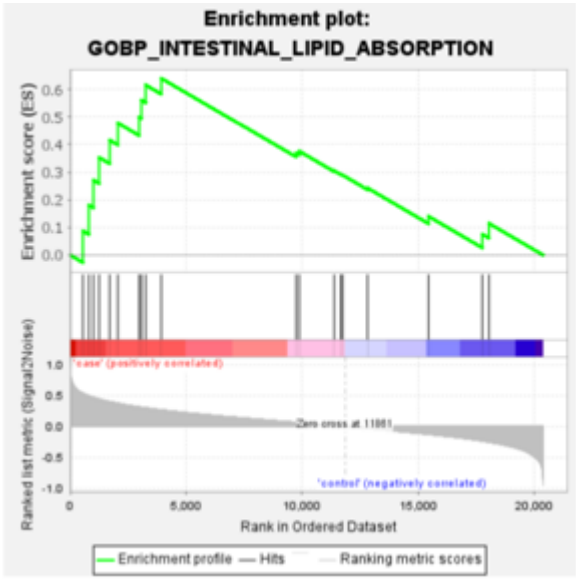


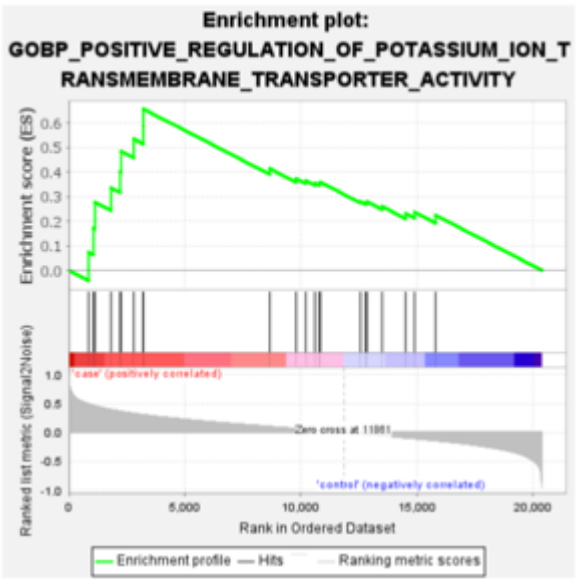

**FIGURE 2***:* (a)Lipid digestion analysis.




(b) Defense response to fungus analysis.




(c) Intestinal lipid absorption analysis.




(d) Positive regulation of potassium ion analysis.

***[Qianhong Yang et al. (2021)]***

## 2.3 Screening of Hub Genes from the PPI Network

The Search Tool for the Retrieval of Interacting Genes (STRING) evaluates and combines PPI, such as physical and functional linkages, to look for interacting genes. STRING version 10.0 now includes 9,643,763 proteins from 2031 different species. We created PPI networks using Cytoscape after creating DEGs using STRING to evaluate the interaction correlation of these DEGs [3].

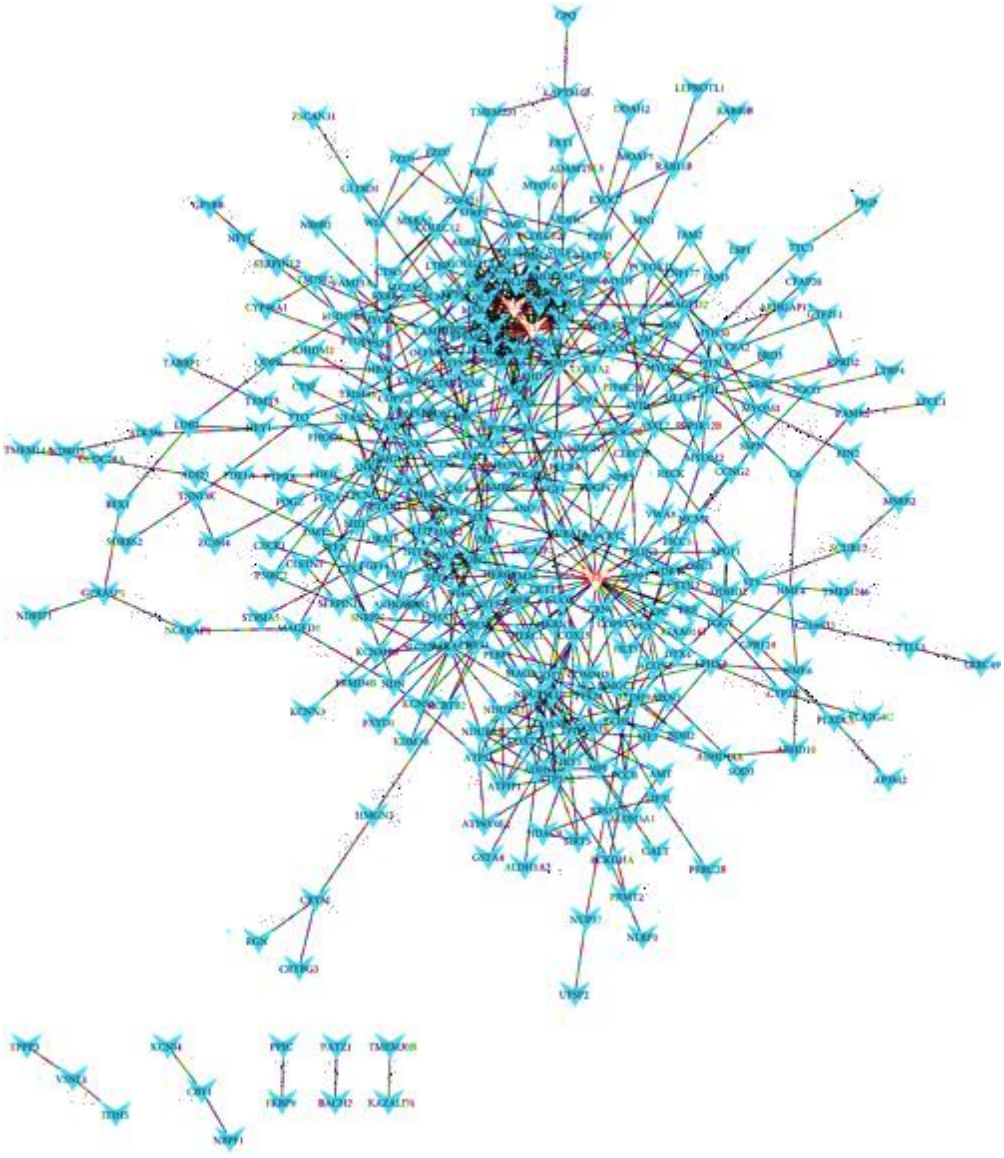


**FIGURE 3*:*** The PPI network of DEGs ***[Xiaodong Sheng et al. (2022)]***

The order of degree, from greatest to least significant, served as one of the primary considerations in the process of identifying hub genes.

Based on the connection, we built a PPI network of genes. We excluded the 3 hub genes COL1A1 (degree = 33), UBB (degree = 32), and COL3A1 (degree = 31) that were generated by the elevated DEGs.

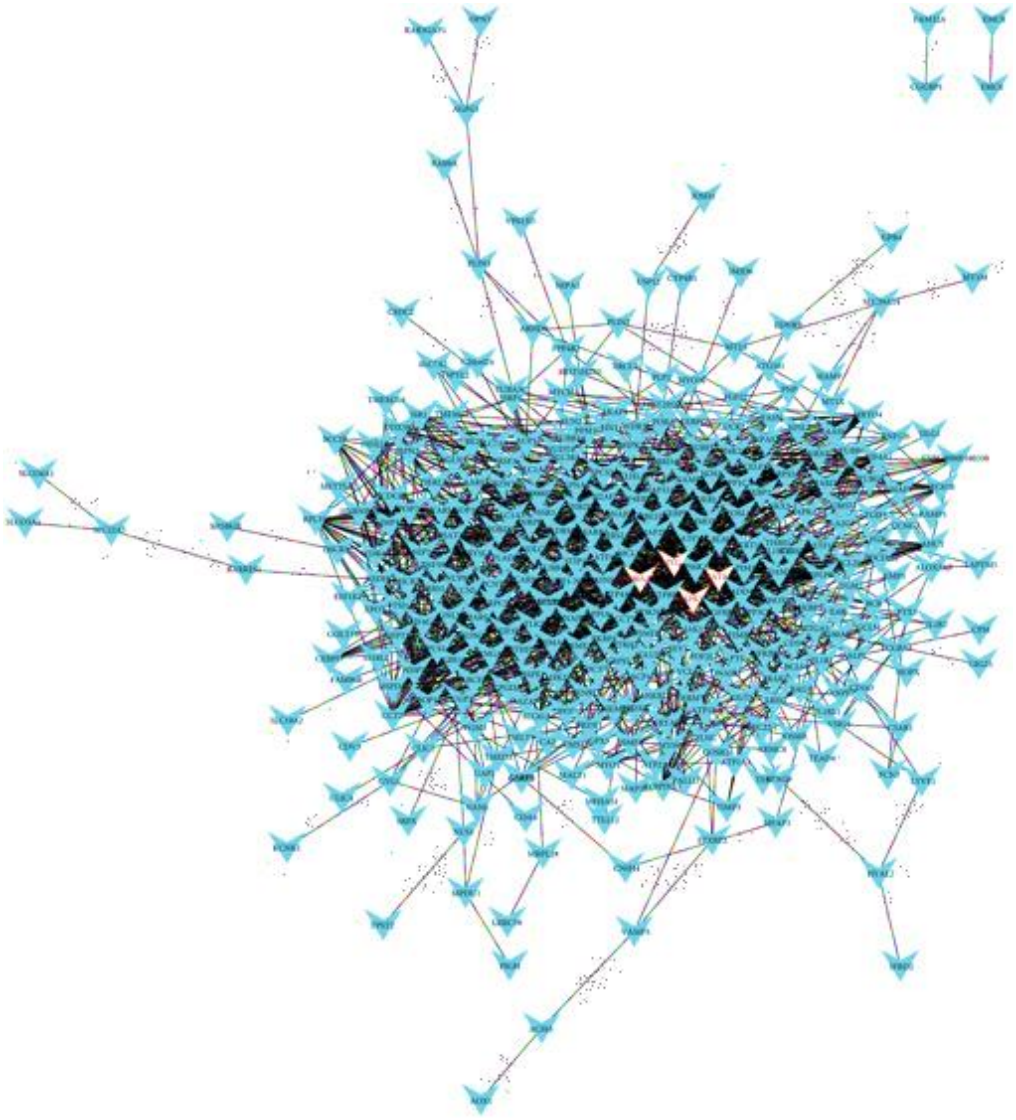


**FIGURE 4*:*** PPI network of downregulated DEGs ***[Xiaodong Sheng et al. (2022)]***

Four hub genes, including HSP90AA1 (degree = 73), MYC (degree = 59), STAT3 (degree = 49), and MAPK1 (degree = 49), were chosen from the PPI network

map created by the downregulated DEGs [3].

## 2.4: ROC curve analysis

ROC curve analysis was used to identify the sensitivity and specificity of the hub genes for HF diagnosis, and we assessed how big the area under the curve (AUC) was by making use of the statistical tool pROC that is included in the R software [4].

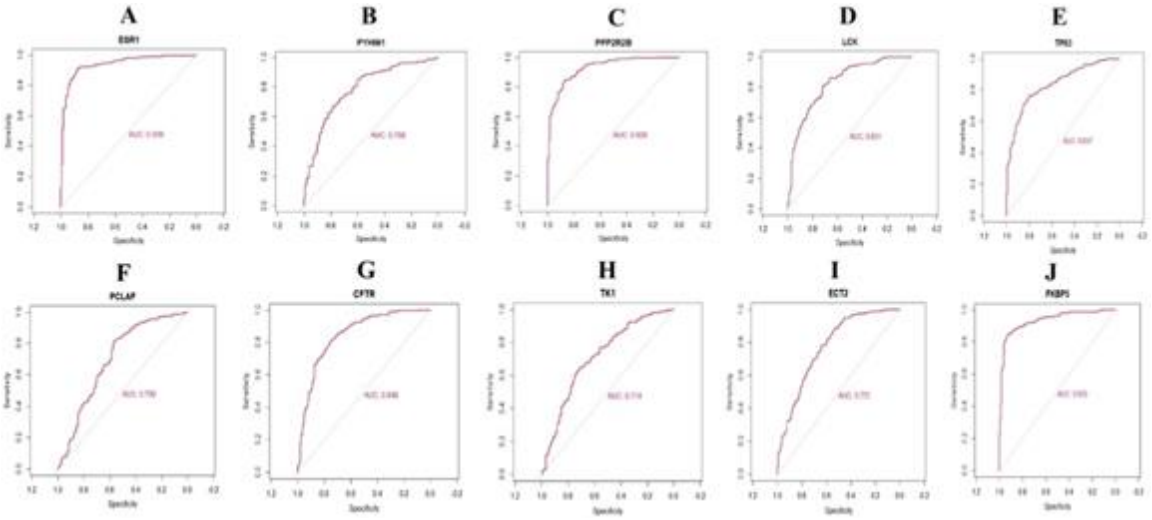



**FIGURE 5:** ROC curve analyses of hub genes. ***[Kolur, V., Vastrad, B., Vastrad, C. et al. (2021)]***
**A.** ESR1, **B.** PYHIN1, **C.** PPP2R2B, **D.** LCK, **E.** TP63, **F.** PCLAF, **G.** CFTR, **H.** TK, **I.** ECT2, **J.** FKBP5

In the first place, we carried out a study of the ROC curve using the GSE141910 to compare the 10 hub genes. ESR1, PYHIN1, PPP2R2B, LCK, TP63, PCLAF, CFTR, TK1, ECT2 and FKBP5 achieved an AUC value of > 0.7, proving that these 10 genes may be used as biomarkers for HF diagnosis because of their high sensitivity and specificity for the disease. The results showed that an AUC value of > 0.7 was achieved by all of these genes [4].

## 2.5: RT-PCR analysis:

ATCC's H9C2 cells were grown in Dulbecco's minimum essential medium (Sigma-Aldrich), which was supplemented with 10% fetal calf serum (Sigma-Aldrich) and 1% streptomycin (Sigma-Aldrich), at a temperature of 37 degrees Celsius and a carbon dioxide concentration of 5%. ATCC's HL-1 cells were cultured in Claycomb medium (supplied by Sigma-Aldrich) that was

supplemented with 10% fetal bovine serum (also provided by Sigma-Aldrich), 1% streptomycin (also provided by Sigma-Aldrich), 1% glutamax (also provided by Sigma-Aldrich), and 0.1 mM norepinephrine (also provided by Sigma-Aldrich) at 37 degrees Celsius and Through the use of the TRI Reagent, total RNA was extracted from the cell cultures of H9C2 for the HF and HL-1 for the normal control (Sigma, USA). cDNA was produced by utilizing a total of 2.0 g of RNA and a reverse transcription cDNA kit [5].

In order to determine the level of relative mRNA expression, a 7 Flex real-time PCR system (manufactured by Thermo Fisher Scientific and located in Waltham, Massachusetts, USA) was used.

The relative expression levels were calculated using the 2-Ct technique, and the results were then normalized using beta-actin as the internal reference. All RT-PCR reactions were carried out in triplicate each and every time [5].

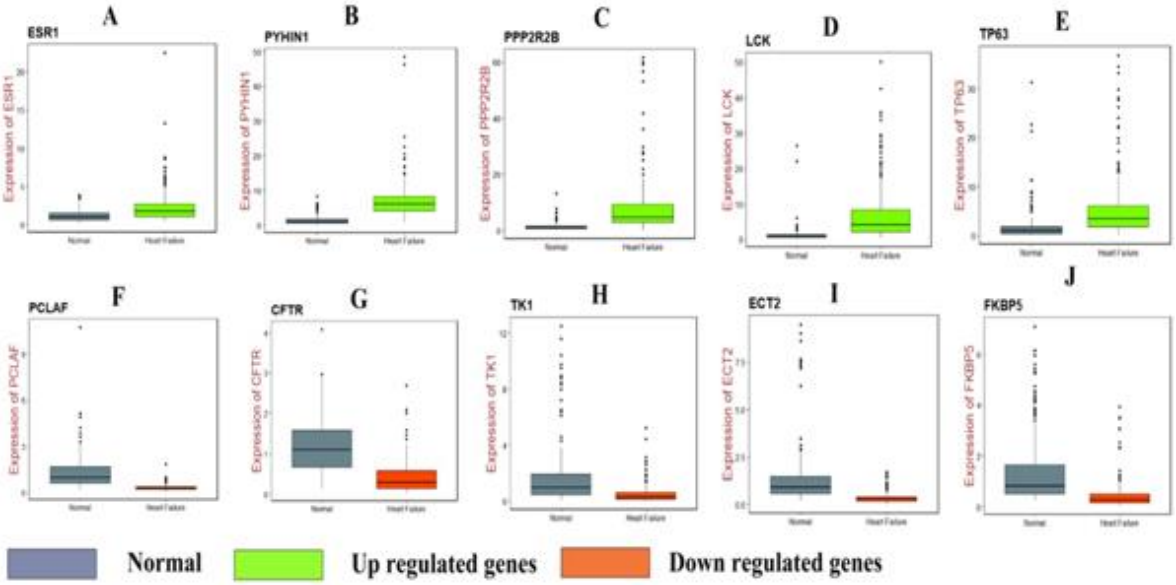



**FIGURE 6**: RT-PCR analyses for the hub genes. ***[Kolur, V., Vastrad, B., Vastrad, C. et al. (2021)]***
**A.** ESR1, **B.** PYHIN1, **C.** PPP2R2B, **D.** LCK, **E.** TP63, **F.** PCLAF, **G**. CFTR **H.** TK1, **I.** ECT2, **J.** FKBP5

In order to verify the hub genes in normal and HF cell lines, reverse transcription-PCR was utilized. According to the findings, the levels of mRNA expression for ESR1, PYHIN1, PPP2R2B, LCK, and TP63 were significantly higher in HF than they were in normal.

On the other hand, the levels of mRNA expression for PCLAF, CFTR, TK1, ECT2, and FKBP5 were significantly lower in HF than they were in normal. These findings are depicted in Figure.

## 3. Bioinformatics strategy to uncover COVID-19's effects on CVD and hypertension.

We used the criterion that the adjusted P-value (adj p-value) should be less than 0.05 and the absolute value of log foldchange (logFC) should not be less than 1 in order to determine the differentially expressed genes (DEGs) for each of these datasets. From the COVID-19 blood datasets, we were able to identify a total of 1289 DEGs. There were 1321 cases of CHF, 247 cases of IHP, 249 cases of PAH, and 127 cases of PE that were recognized as having DEGs [6].

It demonstrates the relevant genes for CHF, IPH, PAH, and PE using four different volcano layouts. The important genes are shown by the spots of red color in the volcano graphs. In addition, we have carried out a comparative investigation in order to locate the DEGs that are shared by COVID-19 and the other four disorders [6].

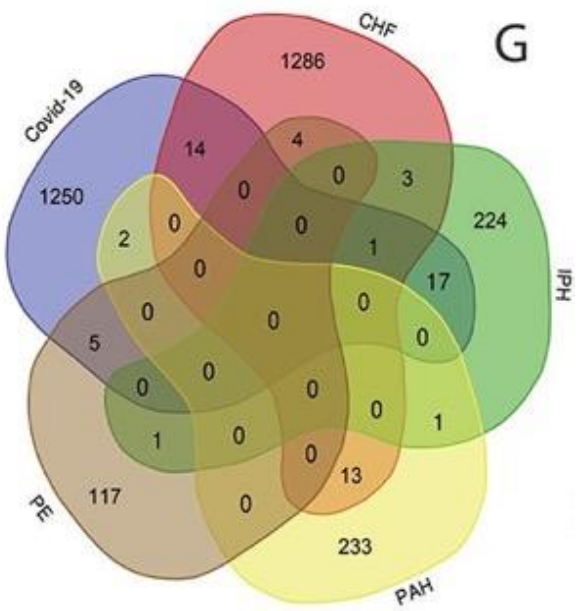



**FIGURE 7*:*** Venn diagram depicts the shared DEGs among COVID-19 immune system and other conditions.

***[Asif Nashiry et al. (2021)]***

Figure illustrates the total number of DEGs that are common across all of the circumstances. Based on our observations, it seems that COVID-19 has a greater degree of overlap with CHF and IHP in comparison to the other two conditions [6].

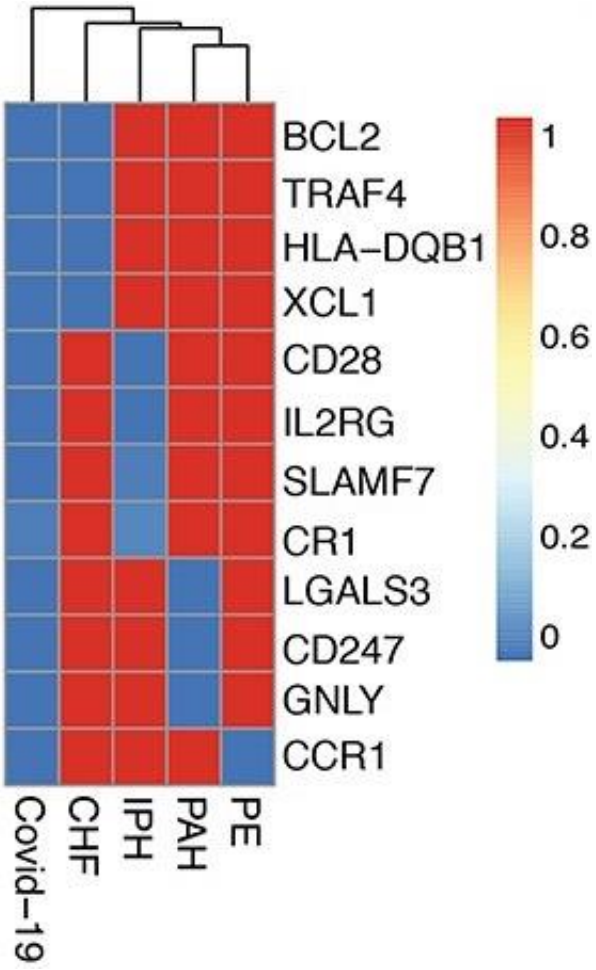



**FIGURE 8***:* adjusted *P* value ***[Asif Nashiry et al. (2021)]***

Heatmaps created by the one-of-a-kind DEGs that are shared by COVID-19 and each of the other four disorders. The COVID-19 does not share a significant number of DEGs with the disorders that are the focus of our investigation, as seen by these two heat maps. There are a total of 15 DEGs that are shared by COVID-19 and CHF; these DEGs include ABCB1, CCL4, IGKC, MS4A2, CTSB, and SAP30. There are just two DEGS that are shared across COVID-19 and PAH, and

they are LTF and FCRL55. When compared to the other conditions, COVID-19 and IPH have the same amount of DEGs in common as the most. The DEGs IGKC, EZH2, MK167, CD180, IGHA1, IGHA2, BTK, and KNL1 are some examples of those that are included in this group. Five DEGs that are shared by COVID-19 and PE have been found by us. These are NUAK1, CCR1, RHOT1, CNTLN, and PLA2G7 [6].

## 4. A New Machine Learning Approach for Better Target Genes Selection.

The "limma" R package was used to obtain the DEGs. As the cutoff for DEGs, samples with P 0.05 and | log FC| > 1.5 were considered. PCA was used to assess the effectiveness of DEGs [7].

Then, we used the R package "cluster Profiler" or the Metascape web database to conduct enrichment studies for the GO, KEGG, and DO taxonomies. The underlying feature biomarkers were screened using two distinct methods (SVM-RFE and LASSO), which produced 10 and 14 genes, respectively (Figures 9A, B).

The intersection of the two hub DEGs of the SVM-RFE, LASSO, and PPI network (IL1B and TLR2) was shown in a Venn diagram (Figure 9C). When compared to the CON group, the MI group's IL1B and TLR2 expression levels were significantly greater (Figures 9D, E) [7].

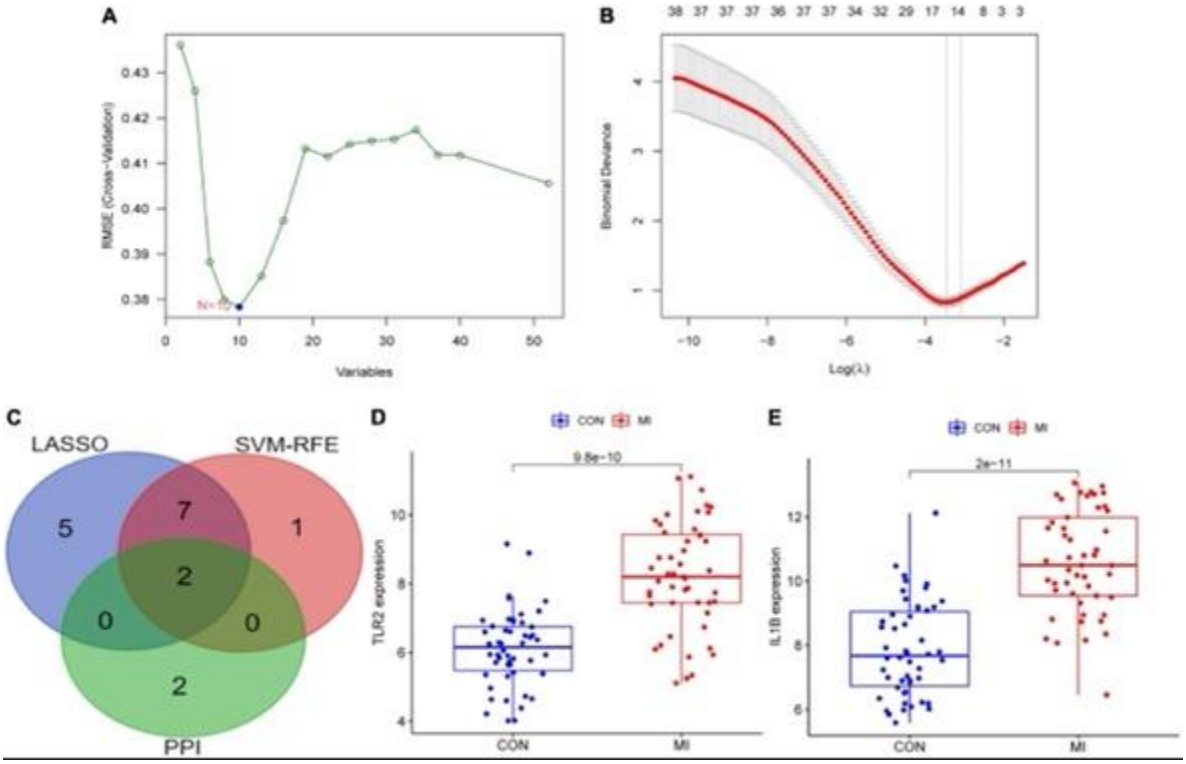



**FIGURE 9.** The screening of hub DEGs using machine learning and PPI. **(A)** 10 DEGs were obtained using the SVM-RFE algorithm. **(B)** 14 DEGs were obtained using the LASSO regression algorithm. **(C)** A Venn diagram presented the hub DEGs. **(D)** The expression levels of TLR2 in GSE66360. **(E)** The expression levels of IL1B in GSE66360. ***[Qunhui Zhang et al. (2022)]***

It is important to understand this study's limitations. First of all, this research was designed as a retrospective cohort study. As a result, important clinical data could not be obtained [7].

Finally, the identification of prospective research to corroborate our findings in the next days is necessary due to the function and immune cell infiltration of IL1B and TLR2 in MI as shown by bioinformatics analysis, SCS, and machine learning algorithms [7].

## 5. Conclusion:

In conclusion, the current review work has identified the hub genes and pathways involved in the development of HF effectively. DEGs are considerably abundant in the G protein-coupled receptor binding, peroxisome, and cAMP signaling pathway, according to functional and enrichment data. According to GSEA, the gene set mostly involved in lipid digestion, fungal defense, and intestinal lipid absorption [1-7]

From PPI networks, we chose 7 hub genes. Through the use of bioinformatics analysis of hub genes and regulatory networks, it may be possible to discover important and novel prospective targets in HF carcinogenesis, prognosis, and diagnostics [10-15].

We looked at the potential effects of SARS-CoV-2 infection in individuals with heart failure and three different kinds of hypertension. From two separate scenarios of interactions between COVID-19 and chronic heart failure and hypertensive disorders, we were able to identify the cell signaling and gene ontology pathways [16]. Although IPH, PAH, and PE are all types of hypertensions, our analysis showed that there are substantial differences in the common DEGs between each of these disorders and COVID-19 [11-13]. The research's findings pinpoint the chemicals responsible for COVID-19's link to cardiovascular and hypertensive illnesses. The results of the earlier investigations, which we discussed are likewise validated and supported by these findings. The relevant genes, pathways, and networks that have been found are linked to other illnesses, which may help researchers find potential new treatments to counteract the consequences of a severe SARS-CoV-2 infection [13].

To further understand the underlying molecular process, we also discovered two hub DEGs (IL1B and TLR2) and demonstrated their possible implications in the diagnosis of MI. Immune cells that infiltrated the heart performed a significant impact in myocardial infarction. A possible medication for the therapy of MI was TCM, particularly HF [7].

However, further study is required to validate the results of this investigation.

## 6. Future scope:

- Modern biology and medicine rely heavily on bioinformatics for the analysis of data.
- A key platform for illness detection, prediction, diagnosis, and therapy is bioinformatics.
- And in the next decade, there is little doubt that this international cooperation will expand significantly.
- Consequently, being knowledgeable in bioinformatics at this time will place you on the worldwide cooperation route as well.
- There is a chance that machine learning may increase success.
- There are several ML methods available for handling massive amounts of data.
- Furthermore, it will be helpful in forecasting cardiac disorders.

## References:


1. Zhang J, Wang J, Wu Y, Li W, Gong K, Zhao P. Identification of SLED1 as a Potential Predictive Biomarker and Therapeutic Target of Post-Infarct Heart Failure by Bioinformatics Analyses. Int Heart J. 2021;62(1):23-32. [Primary article]
2. Qianhong Yang, Xiaolu Bai, Xiang Li, Wei Hu, "The Identification of Key Genes and Biological Pathways in Heart Failure by Integrated Bioinformatics Analysis", *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 3859338, 10 pages, 2021. [Primary article]
3. Sheng X, Jin X, Liu Y, Fan T, Zhu Z, Jin J, Zheng G, Chen Z, Lu M, Wang Z, "The Bioinformatical Identification of Potential Biomarkers in Heart Failure Diagnosis and Treatment", *Genetics Research*, vol. 2022, Article ID 8727566, 11 pages, 2022. [Primary article]
4. Kolur, V., Vastrad, B., Vastrad, C. *et al.* Identification of candidate biomarkers and therapeutic agents for heart failure by bioinformatics analysis. *BMC Cardiovasc Disord* **21**, 329 (2021). [Primary article]
5. Nashiry A, Sumi S S, Islam S, Quinn J M W, Moni M A, Bioinformatics and system biology approach to identify the influences of COVID-19 on cardiovascular and hypertensive comorbidities, *Briefings in Bioinformatics*, Volume 22, Issue 2, March 2021, Pages 1387–1401, [Primary article]
6. Boer RA, Silljé HHW. A combined bioinformatics, experimental and clinical approach to identify novel cardiac-specific heart failure biomarkers: is Dickkopf-3 (DKK3) a possible candidate? Eur J Heart Fail. 2020 Nov;22(11):2065-2074. Epub 2020 Sep 22. PMID: 32809235; PMCID: PMC7756877. [Primary article]
7. Almustafa, K.M. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics* **21**, 278 (2020). [Primary article]

8. Shen NN, Zhang ZL, Li Z, Zhang C, Li H, Wang JL, Wang J, Gu ZC. Identification of microRNA biomarkers in atrial fibrillation: A protocol for systematic review and bioinformatics analysis. Medicine (Baltimore). 2019 Jul;98(30):e16538. [Primary article]PMID: 31348272; PMCID: PMC6708903.
9. Chen X, Zhang Q, Zhang Q. Predicting potential biomarkers and immune infiltration characteristics in heart failure. Math Biosci Eng. 2022 Jun 16;19(9):8671-8688. [Primary article]
10. Li X, Li B, Jiang H. Identification of time-series differentially expressed genes and pathways associated with heart failure post-myocardial infarction using integrated bioinformatics analysis. Mol Med Rep. 2019;19(6):5281–90. [Primary article]
11. Peirlinck M, Sahli Costabal F, Sack KL, Choy JS, Kassab GS, Guccione JM, De Beule M, Segers P, Kuhl E. Using machine learning to characterize heart failure across the scales. Biomech Model Mechanobiol. 2019 Dec;18(6):1987-2001. [Primary article]Epub 2019 Jun 25. PMID: 31240511.
12. Cameron R. Olsen, Robert J. Mentz, Kevin J. Anstrom, David Page, Priyesh A. Patel, Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure, American Heart Journal,Volume 229,2020, Pages 1-17, ISSN 0002-8703, [Primary article]
13. Chen XM, Cao F, Zhang HM, Chen HR, Zhang JD, Zhi P, Li ZY, Wang YX, Lu XC. [Exploration of omics mechanism and drug prediction of coronavirus-induced heart failure based on clinical bioinformatics]. Zhonghua Xin Xue Guan Bing Za Zhi. 2020 Jul 24;48(7):587-592. Chinese. PMID: 32228827. [Primary article]
14. Xanthopoulos A, Triposkiadis F, Starling RC. Heart failure with preserved ejection fraction: classification based upon phenotype is essential for diagnosis and treatment. Trends Cardiovasc Med. 2018;28(6):392–400. [Primary article]

15. Roger VL, Weston SA, Redfield MM, Hellermann-Homan JP, Killian J, Yawn BP, Jacobsen SJ. Trends in heart failure incidence and survival in a community-based population. JAMA. 2004;292(3):344–50. [Primary article]
16. Missov E, Calzolari C, Pau B. Circulating cardiac troponin I in severe congestive heart failure. Circulation. 1997;96(9):2953–8[Primary article]
17. Yang Y, Tian Y, Hu S, Bi S, Li S, Hu Y, Kou J, Qi J, Yu B. Extract of Sheng–Mai–San ameliorates myocardial ischemia-induced heart failure by modulating Ca2+-calcineurin-mediated Drp1 signaling pathways. Int J Mol Sci. 2017;18(9):1825. [Primary article]
18. Liu Y, Morley M, Brandimarto J, Hannenhalli S, Hu Y, Ashley EA, Tang WH, Moravec CS, Margulies KB, Cappola TP, et al. RNA-Seq identifies novel myocardial gene expression signatures of heart failure. Genomics. 2015;105(2):83–9. [Primary article]
19. H. S. M. Raffaa, B. A. Alasmari, S. A. Abadi et al., "Adherence of heart failure patients to heart failure medications and its determinants in the Aseer region, Southern Saudi Arabia," *Journal of Family Medicine and Primary Care*, vol. 9, no. 9, pp. 5041–5045, 2020. [Primary article]
20. K. Darahim, "Usefulness of right atrial volume index in predicting outcome in chronic systolic heart failure," *Journal of the Saudi Heart Association*, vol. 26, no. 2, pp. 73–79, 2014. [Primary article]
21. Al-Mukhtar O, Vogrin S, Lampugnani ER, Noaman S, Dinh DT, Brennan AL, et al. Temporal changes in pollen concentration predict short-term clinical outcomes in acute coronary syndromes. J Am Heart Assoc. (2022) 11:e023036. doi: 10.1161/jaha.121.023036 [Primary article]
22. Md MA, Aktar S, Md R-A-M, et al. A machine learning model to identify early stage symptoms of sars-cov-2 infected patients. Expert Sys Appl 2020; 160:113661. [Primary article]
23. Aktar S, Talukder A, Ahamad M, et al. Machine learning and meta-analysis approach to identify patient comorbidities and symptoms that

increased risk of mortality in covid-19. arXiv preprint arXiv:200812683 2020. [Primary article]

24. Marco Cascella, Michael Rajnik, Arturo Cuomo, et al. Features, evaluation and treatment coronavirus (covid-19). In Statpearls [internet]. StatPearls Publishing, 2020. [Primary article]
25. Goujon M, McWilliam H, Li W, et al. A new bioinformatics analysis tools framework at embl–ebi. Nucleic Acids Res 2010; 38(suppl_2): W695–9. [Primary article]