International Journal of Medical and Health Sciences
Volume 12 Issue 01 April 2026
ISSN- 2456 - 6063
DOI: https://doi.org/10.69980/gfs9k223

INTEGRATED SYSTEMS BIOLOGY AND MACHINE LEARNING FRAMEWORK FOR THE IDENTIFICATION OF MITOCHONDRIAL AND SIGNALING SIGNATURES IN ISCHEMIC AND DILATED CARDIOMYOPATHY

Mahak[1*], Shaik Mubarak Basha[1], Janani Sp[1], Himanshu Shekhar Singh[1], Ms. Kesiya Joy[1]

[1] Department of Life Sciences, School of Science, Garden City University, Bangalore, India

Corresponding author: Mahak, Email: mahakkumari735@gmail.com

Received:- 18/02/26	Revised:-26/03/26	Accepted:-03/04/26	Published:- 10/04/26

ABSTRACT

Cardiovascular diseases (CVDs), predominantly manifested as ischemic cardiomyopathy (ICM) and dilated cardiomyopathy (DCM), constitute the leading cause of morbidity and mortality worldwide, driven by a complex architecture of genetic, metabolic, and environmental alterations. Traditional clinical diagnostic modalities, while effective for hemodynamic assessment, often fail to capture the granular molecular heterogeneity required for early intervention and personalized therapeutic strategies. The advent of high-throughput transcriptomics has provided a systems-level view of these pathologies, yet the translation of high-dimensional genomic data into robust clinical biomarkers remains hindered by the "curse of dimensionality" and the noise inherent in biological datasets. This study presents a comprehensive, integrated bioinformatics and machine learning pipeline to identify a compact, biologically potent gene signature for CVD classification. Drawing conceptual inspiration from the "Moana" classification framework—which prioritizes biological structure over raw feature variance—we employed a pathway-guided feature selection strategy on RNA-sequencing data (GSE55296) and microarray validation datasets (GSE57338). Functional enrichment analysis using KEGG and Reactome databases revealed a profound dysregulation of mitochondrial metabolism, oxidative phosphorylation, and cardiomyopathy-related pathways. Subsequent protein–protein interaction (PPI) network analysis, refined through the intersection of multiple topological centrality algorithms (Degree, Maximal Clique Centrality, and Betweenness), distilled the transcriptome into a core signature of six genes: COX4I1, NDUFA5, NDUFV2, NDUFS3, NDUFS4, and AKT1. These features were utilized to train and evaluate supervised machine learning models. The Random Forest classifier demonstrated superior stability and performance, achieving a mean cross-validation accuracy of approximately 69.5%, significantly outperforming a Deep Neural Network (DNN) in this limited-sample regime. The findings underscore the critical role of mitochondrial Complex I and IV dysfunction in heart failure and demonstrate the efficacy of integrating systems biology with ensemble learning to develop interpretable, high-precision molecular diagnostics.

KEYWORDS: Systems Biology, Machine Learning, Cardiomyopathy, Mitochondrial Dysfunction, Signal Transduction Pathways, Biomarker Identification

1. INTRODUCTION

1.1 The Clinical and Molecular Imperative in Heart Failure

Cardiovascular diseases (CVDs) represent a preeminent global health challenge, accounting for approximately 32% of all deaths worldwide and imposing a staggering economic burden on healthcare systems (World Health Organization, 2023). Within this broad category, heart failure (HF) serves as the terminal pathway for a myriad of cardiac pathologies, with ischemic cardiomyopathy (ICM) and dilated cardiomyopathy (DCM) being the two most prevalent etiologies (Braunwald, 2013). ICM typically arises from coronary artery disease, where prolonged ischemia leads to irreversible cardiomyocyte necrosis, scar formation, and subsequent ventricular remodeling. In contrast, DCM is characterized by ventricular chamber enlargement and contractile dysfunction, frequently driven by a complex interplay of genetic mutations (e.g., in cytoskeletal proteins like titin), viral myocarditis, or idiopathic factors (McKenna & Judge, 2021).

Despite their distinct initiating mechanisms, ICM and DCM converge on a common phenotypic endpoint characterized by energetic failure, neurohormonal activation, and maladaptive structural remodeling (Murphy et al., 2020). Current clinical diagnostic tools—ranging from echocardiography and angiography to circulating biomarkers like B-type natriuretic peptide (BNP)—are adept at characterizing the macroscopic, hemodynamic consequences of the disease. However, they lack the resolution to interrogate the underlying molecular drivers that precipitate these morphological changes (Maisel et al., 2002). This diagnostic gap is critical, as the molecular landscape of heart failure, including metabolic reprogramming and transcriptional dysregulation, often precedes gross functional decline, offering a window for early detection and targeted therapeutic intervention (Stanley et al., 2005).

1.2 The High-Dimensionality Challenge in Genomic Medicine

The integration of high-throughput transcriptomic technologies, such as microarrays and next-generation RNA sequencing (RNA-seq), has revolutionized cardiovascular research by enabling the simultaneous quantification of tens of thousands of transcripts (Wang et al., 2009). These technologies allow researchers to capture a comprehensive "snapshot" of the cellular state, revealing widespread dysregulation in gene expression associated with hypertrophy, fibrosis, inflammation, and metabolic shifts (Margulies et al., 2009). For instance, landmark studies utilizing RNA-seq platforms have provided unprecedented depth in profiling the failing human heart, identifying novel isoforms and non-coding RNAs previously undetectable by hybridization-based methods (Tarazón et al., 2014).

However, the translation of these rich datasets into clinical applications is severely hampered by the "curse of dimensionality"—a statistical phenomenon where the number of features (genes) vastly exceeds the number of samples (patients) (Bellman, 1961). In this high-dimensional space, traditional differential gene expression (DGE) analysis often yields lists of thousands of statistically significant genes. Many of these are likely "passenger" genes—downstream consequences of the disease rather than "driver" genes central to the pathology (Barabási et al., 2011). Furthermore, predictive models built on raw, high-dimensional data are prone to overfitting, capturing noise specific to the training set rather than true biological signals, leading to poor generalization on independent cohorts (Libbrecht & Noble, 2015).

1.3 Systems Biology and the "Moana" Philosophy

To surmount these challenges, modern computational biology has shifted from a reductionist, gene-centric view to a systems-level, network-centric perspective. This approach posits that disease phenotypes arise not from isolated genetic defects but from perturbations in complex, interconnected molecular networks such as protein-protein interactions (PPI), metabolic pathways, and regulatory circuits (Barabási et al., 2011). By mapping gene expression data onto these known scaffolds, researchers can identify functional modules and "hub genes" that serve as critical control points (Shannon et al., 2003).

This study adopts a pathway-guided feature selection strategy, conceptually aligned with advanced classification frameworks such as the "Moana" framework (Wagner & Yanai, 2018). Originally developed for single-cell RNA-seq (scRNA-seq), the Moana framework addresses the sparsity and technical noise of single-cell data by leveraging biological manifold structures using k-nearest neighbor (kNN) smoothing and hierarchical support vector machines (SVMs) to classify cell states based on robust biological signals rather than raw counts (Wagner & Yanai, 2018).

While the present study utilizes bulk RNA-seq data, the underlying philosophy of Moana—reducing technical noise by enforcing biological coherence—is directly applicable. Instead of treating the transcriptome as an unstructured bag of 20,000 genes, we utilize prior biological knowledge from curated databases such as KEGG and Reactome to structure the feature space (Kanehisa et al., 2017; Fabregat et al., 2018). This methodology allows us to filter out noise and focus on genes that form the structural pillars of disease-enriched pathways. By integrating this pathway-guided approach with robust network topology

analysis, we aim to derive a feature set that is not only statistically robust but biologically interpretable and computationally efficient for machine learning applications.

1.4 Research Objectives

1. The primary objective of this research is to develop an integrated, reproducible bioinformatics and machine learning pipeline to classify cardiovascular disease subtypes (ICM, DCM, and Control) using transcriptomic data. The specific aims are as follows:
2. Data Processing: To systematically preprocess and normalize RNA-seq data from the GSE55296 dataset, ensuring variance stabilization and suitability for downstream analysis:
3. Pathway Discovery: To identify key biological pathways enriched in heart failure using the KEGG and Reactome databases, providing a functional map of the disease state.
4. Network Analysis: To construct Protein-Protein Interaction (PPI) networks and apply multiple centrality algorithms (Degree, Maximal Clique Centrality, Betweenness) to identify consensus hub genes.

2.LITERATURE REVIEW

2.1 The Global Burden and Molecular Complexity of Heart Failure

Cardiovascular diseases remain the leading cause of death globally, necessitating a continuous search for novel therapeutic targets and diagnostic markers (World Health Organization, 2023). Heart failure, a complex clinical syndrome, represents the end stage of various cardiac pathologies. The distinction between ischemic cardiomyopathy (ICM) and dilated cardiomyopathy (DCM) is clinically significant; ICM is primarily a vascular disorder caused by coronary artery disease leading to myocardial ischemia and cardiomyocyte death, whereas DCM is a primary myocardial muscle disorder often associated with genetic mutations, viral infections, or idiopathic causes (Braunwald, 2013; McKenna & Judge, 2021). However, at the molecular level, both conditions share significant overlaps, particularly in the reactivation of the "fetal gene program" and the disruption of mitochondrial bioenergetics (Stanley et al., 2005).

Early microarray studies demonstrated that the failing heart undergoes a metabolic shift, reverting from fatty acid oxidation—the dominant energy source in the adult heart—to glycolysis, which is characteristic of fetal cardiac metabolism (Taegtmeyer et al., 2010). This metabolic inflexibility is considered a hallmark of the failing myocardium. However, identifying the precise molecular regulators of this metabolic transition remains challenging due to patient heterogeneity and the dynamic nature of gene expression in cardiovascular diseases (Margulies et al., 2009). Recent evidence suggests that metabolic reprogramming is not merely a consequence of heart failure but an active driver of disease progression. Mitochondrial dysfunction can lead to reduced ATP production, excessive generation of reactive oxygen species (ROS), and impaired calcium homeostasis, collectively contributing to cardiomyocyte dysfunction and cardiac remodeling (Rosca & Hoppel, 2013).

2.2 Transcriptomics in Cardiovascular Research: From Arrays to RNA-Seq

The development of transcriptomic technologies has dramatically advanced cardiovascular research by enabling the large-scale analysis of gene expression patterns in diseased cardiac tissues (Wang et al., 2009). Early studies relied on hybridization-based microarrays, which allowed simultaneous measurement of thousands of transcripts and provided the first insights into global transcriptional changes associated with heart failure (Margulies et al., 2009).

Microarrays: Studies utilizing Affymetrix platforms, including those incorporated into the large-scale dataset GSE57338, have provided extensive transcriptomic profiles of heart failure patients. Liu et al. (2015) analyzed this dataset comprising 313 myocardial samples (95 ICM, 82 DCM, and 136 non-failing controls) to validate disease-specific gene signatures. Their findings demonstrated that transcriptomic profiles could reliably distinguish failing hearts from healthy myocardium and highlighted the potential of combining small discovery cohorts with large validation datasets for biomarker discovery (Liu et al., 2015).

RNA-Sequencing: RNA-seq technology offers several advantages over microarrays, including a broader dynamic range, improved sensitivity for low-abundance transcripts, and the ability to detect novel transcripts and alternative splicing events (Wang et al., 2009). The GSE55296 dataset generated by Tarazón et al. (2014) represents a landmark RNA-seq study investigating human cardiomyopathy. This dataset includes 36 left ventricular samples (13 ICM, 13 DCM, and 10 controls) analyzed using the SOLiD 5500XL sequencing platform. The study revealed significant alterations in mitochondrial gene expression and natriuretic peptide processing pathways. Interestingly, while the protein levels of atrial natriuretic peptide (ANP) precursors remained stable, the mRNA levels of the NPPA gene were significantly increased, suggesting the presence of complex post-transcriptional regulatory mechanisms (Tarazón et al., 2014). RNA-seq therefore provides a powerful framework for exploring the molecular landscape of heart failure and identifying potential

diagnostic biomarkers.

2.3 Network Medicine and the "Hub Gene" Hypothesis

Traditional approaches focusing on individual genes have gradually been replaced by systems biology and network medicine perspectives. These approaches view diseases as disruptions of interconnected molecular networks rather than isolated genetic defects (Barabási et al., 2011). In protein–protein interaction (PPI) networks, proteins are represented as nodes and their interactions as edges. Within these networks, highly connected nodes known as hub genes often play critical roles in maintaining network stability and cellular function. Disruption of such hub genes can have cascading effects that destabilize biological systems and contribute to disease progression (Barabási et al., 2011).

To identify hub genes within PPI networks, several topological centrality metrics are commonly used. Degree centrality measures the number of direct connections a node has, reflecting its local importance within the network. Betweenness centrality quantifies how frequently a node lies on the shortest paths between other nodes, indicating its role in information flow and signal transduction. Maximal Clique Centrality (MCC) identifies proteins that participate in highly interconnected subnetworks, which are often associated with essential biological functions (Chin et al., 2014).

Relying on a single centrality metric may lead to biased conclusions because each metric captures different aspects of network topology. Therefore, integrating multiple centrality measures and identifying genes that consistently rank highly across different algorithms provides a more reliable strategy for detecting biologically significant hub genes (Chin et al., 2014). This consensus-based approach enhances the robustness of network analyses and improves the likelihood of identifying key molecular drivers of disease.

2.4 Machine Learning in Genomics: The "Moana" Philosophy

Machine learning (ML) has emerged as a powerful tool for analyzing large-scale genomic datasets and identifying patterns that may not be apparent through conventional statistical methods (Libbrecht & Noble, 2015). However, the application of ML to biological data requires careful consideration because genomic datasets often exhibit high dimensionality and limited sample sizes.

The Moana framework, introduced by Wagner and Yanai (2018), provides a robust classification strategy for single-cell RNA-seq data. Moana employs k-nearest neighbor (kNN) smoothing and hierarchical support vector machines to reduce noise and classify cellular states based on biologically meaningful expression patterns rather than raw count data (Wagner & Yanai, 2018). Although Moana was developed for single-cell transcriptomics, its underlying philosophy—leveraging biological structure to reduce noise—can be extended to bulk RNA-seq analyses.

One strategy inspired by this concept is pathway-guided feature selection (PGFS), which incorporates prior biological knowledge from curated pathway databases such as KEGG and Reactome to guide feature selection. Instead of treating thousands of genes as independent variables, PGFS focuses on genes that participate in biologically relevant pathways, thereby reducing dimensionality and improving model interpretability (Kanehisa et al., 2017; Fabregat et al., 2018). This approach helps mitigate the curse of dimensionality and enhances the predictive performance of machine learning models in genomic studies (Libbrecht & Noble, 2015).

When selecting machine learning algorithms for transcriptomic data, model complexity must be balanced with dataset size. Deep Neural Networks (DNNs) are powerful models capable of capturing complex nonlinear relationships, but they typically require very large datasets to avoid overfitting. In contrast, Random Forest (RF) models, which are ensemble learning methods based on decision trees, tend to perform well in high-dimensional datasets with limited sample sizes because they utilize bootstrap aggregation (bagging) and random feature selection to reduce variance (Breiman, 2001). Consequently, Random Forest models are often more suitable for genomic datasets where the number of features greatly exceeds the number of samples.

In this study, both DNN and RF models are evaluated to determine their effectiveness in classifying cardiovascular disease subtypes using transcriptomic signatures. By integrating systems biology with machine learning techniques, this approach aims to identify robust biomarkers capable of improving disease classification and advancing precision cardiology.

3. MATERIALS AND METHODS

3.1 Study Design and Analytic Workflow

This study implements a multi-stage computational pipeline designed to distill raw transcriptomic data into a validated predictive model. The workflow follows a logical progression: Data Acquisition - Preprocessing - Functional Enrichment - Network Construction - Hub Gene Selection - Machine Learning Classification. This "funnel" approach progressively discards noise and retains high-value biological signals, ensuring that

the final predictive model is based on biologically plausible mechanisms rather than statistical artifacts.(Libbrecht & Noble, 2015).

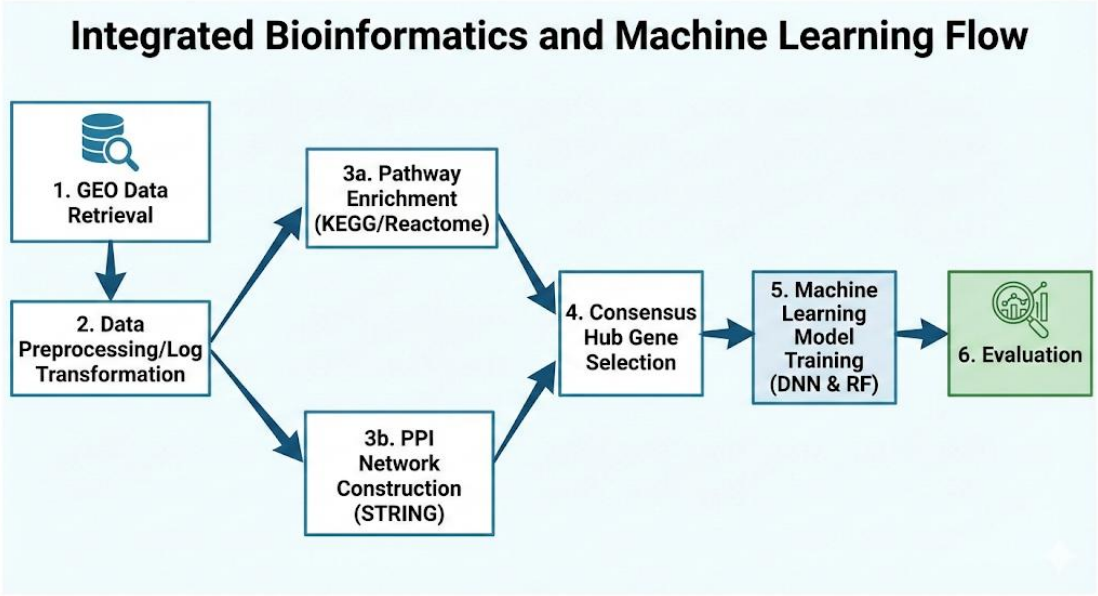

Figure 1: Workflow of the integrated analysis pipeline

3.2 Data Acquisition

Publicly available transcriptomic datasets were retrieved from the NCBI Gene Expression Omnibus (GEO).(Edgar et al., 2002)

Discovery Dataset (GSE55296): This RNA-seq dataset comprises 36 left ventricular (LV) myocardial tissue samples: 13 from patients with Ischemic Cardiomyopathy (ICM), 13 from patients with Dilated Cardiomyopathy (DCM), and 10 from healthy control donors.2 The samples were obtained from explanted hearts during transplantation or from organ donors whose hearts were unsuitable for transplant. The data was generated using the SOLiD 5500XL platform, a sequencing-by-ligation technology known for high accuracy(Tarazón et al., 2014). The raw count data was downloaded for analysis.

Reference Dataset (GSE57338): Used for comparative context and validation of the biological relevance of the identified pathways, this large-scale microarray dataset utilizes the Affymetrix Human Gene 1.1 ST Array. It includes a total of 313 samples (95 ICM, 82 DCM, 136 Non-failing)(Liu et al., 2015). The scale of this dataset (N=313) contrasts with the discovery dataset (N=36), allowing for a robust assessment of the generalization potential of transcriptomic signatures.

3.3 Data Preprocessing and Normalization

Raw RNA-seq count data from GSE55296 were imported into the Python environment using the Pandas library. Count data typically follows a negative binomial distribution and is highly skewed, with a few highly expressed genes dominating the dataset(Love et al., 2014). To address this and stabilize the variance across the dynamic range of expression, a logarithmic transformation was applied:

E_{normalized} = \log_2(Count + 1)

The addition of a pseudo-count of 1 ensures that zero-count values are defined (log_2(1) = 0). This transformation renders the data more normally distributed, which is a prerequisite for many parametric statistical tests and enhances the performance of machine learning algorithms by reducing the influence of extreme outliers(Law et al., 2014). Non-numeric metadata columns were programmatically removed to generate a clean numerical matrix of dimensions 20,214 \times 36, representing the expression of over 20,000 genes across the 36 samples.

3.4 Pathway Enrichment Analysis

Rather than relying solely on differential expression cutoffs, which can arbitrarily exclude biologically relevant genes with subtle expression changes (a common issue in heterogeneous diseases like HF), a ranked list of genes was subjected to functional enrichment analysis.

Databases: The KEGG 2021 Human (Kanehisa et al., 2017) and Reactome 2024 (Fabregat et al., 2018) databases were queried using the Enrichr and g:Profiler APIs(Kuleshov et al., 2016; Raudvere et al., 2019).

These databases were selected for their comprehensive curation of metabolic, signaling, and disease-related pathways. KEGG provides high-level maps of biological processes, while Reactome offers detailed molecular reaction events.
election Criteria: The analysis utilized the Fisher's exact test to determine statistical significance. Pathways with an adjusted p-value (Benjamini-Hochberg False Discovery Rate) of < 0.05 were considered significant(Benjamini & Hochberg, 1995). The analysis prioritized pathways related to "Cardiomyopathy," "Oxidative Phosphorylation," "Mitochondrial Metabolism," and "Signaling Pathways" to ensure biological relevance to the disease phenotype.

3.5 Network Construction and Topology Analysis

Genes identified in the top enriched pathways were extracted to form a "pathway-refined" gene set. This set represents the intersection of statistical significance and biological relevance.
PPI Network Construction: The refined gene list was mapped to the STRING database (v11.5) to construct a Protein-Protein Interaction (PPI) network (Szklarczyk et al., 2021). The STRING database integrates interaction data from multiple sources, including experimental crystallography, computational prediction, and text mining.
Confidence Threshold: A medium confidence score (> 0.4) was applied to the interactions. This threshold balances the need to filter out low-confidence, noisy interactions while maintaining enough network density to identify meaningful clusters.
Topology Analysis: The interaction network was imported into Cytoscape (v3.9.1) for visualization and topological analysis(Shannon et al., 2003). The cytoHubba plugin was used to calculate three distinct centrality algorithms for every node in the network (Chin et al., 2014):
1.Degree Centrality: C_D(v) = \deg(v). This measures the number of direct edges connected to a node. High degree nodes are "local" hubs.
2.Maximal Clique Centrality (MCC): This metric defines the centrality of a node based on its participation in maximal cliques (complete subgraphs). It is considered one of the most effective measures for identifying essential proteins in yeast and human PPI networks.1
3.Betweenness Centrality: C_B(v) = \sum \frac{\sigma_{st}(v)}{\sigma_{st}}. This measures the frequency with which a node acts as a bridge along the shortest path between two other nodes. High betweenness nodes are critical for information flow and network cohesion.

3.6 Hub Gene Selection (The Consensus Strategy)

To avoid algorithmic bias inherent in any single centrality measure, a consensus strategy was employed. The top-ranked genes (top 10-20) from the Degree, MCC, and Betweenness lists were extracted. A Venn diagram analysis was performed to identify the intersection of these three sets. Only genes that ranked in the top tier across all three centrality measures (Degree \ MCC \ Betweenness) were selected as "Consensus Hub Genes." This rigorous filtering strategy ensured that the final signature consisted of genes that were not only highly connected but also structurally critical to the network's integrity and information processing capabilities.

3.7 Machine Learning Model Development

The expression values of the selected consensus hub genes were extracted from the normalized dataset to create a reduced feature matrix . Each sample was assigned a categorical label: 0 (Control), 1 (ICM), or 2 (DCM). Two distinct supervised learning models were constructed to evaluate predictive performance:

1. Deep Neural Network (DNN): Implemented using the TensorFlow/Keras framework (Abadi et al., 2016). The architecture consisted of:

- Input Layer: 6 neurons (matching the 6 feature genes).
- Hidden Layers: Two dense layers with Rectified Linear Unit (ReLU) activation functions to capture non-linear relationships.
- Output Layer: A dense layer with 3 neurons and a Softmax activation function to output class probabilities for the three disease states.
- Optimization: The model was compiled using the Adam optimizer and categorical cross-entropy loss function.

2. **Random Forest (RF):** Implemented using the Scikit-learn library(Pedregosa et al., 2011). The Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time(Breiman, 2001).

3. Hyperparameters: The model utilized 100 decision trees (n_estimators=100) with Gini impurity as the splitting criterion.

- Mechanism: Each tree is trained on a random subset of the data (bootstrap sample) and considers a random subset of features for each split. The final prediction is the mode of the classes (voting) of the individual trees. This "bagging" approach significantly reduces variance and the risk of overfitting, which is critical for small datasets.
- Validation Strategy: The dataset was split into training and testing sets using stratified sampling to ensure that the proportion of ICM, DCM, and Control samples remained consistent across splits. Model performance was evaluated using:
- Confusion Matrices: To visualize the specific classes being misclassified.
- Metrics: Accuracy, Precision, Recall, and F1-Score.
- K-Fold Cross-Validation: A 5-fold cross-validation was performed on the Random Forest model to assess its stability. In this process, the data is partitioned into 5 subsets; the model is trained on 4 and tested on 1, repeating the process 5 times. The mean accuracy and standard deviation were calculated to provide a robust estimate of the model's generalization error.

4. RESULTS

4.1 Transcriptomic Data Landscape and Preprocessing

The preprocessing of the GSE55296 dataset yielded a normalized expression matrix encompassing 20,214 genes across the 36 samples. The log-transformation (log_2(x+1)) successfully mitigated the skewness inherent in the raw RNA-seq count data, producing expression distributions that approximated normality. This step was crucial for ensuring the validity of subsequent statistical tests and the stability of the machine learning algorithms.

Table 1: Dataset Details (GSE55296)

GEO Dataset Summary	
Attribute	**Value**
GEO Accession	GSE55296
Organism	Homo sapiens
Platform	SOLiD 5500XL
Total Samples	36
Groups	Control (10), Ischemic CM (13), Dilated CM (13)
Total Genes	20,214

Initial quality control visualization (not shown) indicated that while Healthy Control samples clustered distinctly, the ICM and DCM samples exhibited significant overlap. This observation is consistent with the literature, reflecting the shared end-stage phenotype of heart failure where distinct etiologies converge on a common molecular pathway of remodeling and failure. This high degree of overlap presents a significant challenge for classification, necessitating a feature selection strategy that can isolate the subtle differences or the core shared drivers.

4.2 Functional Enrichment Reveals Mitochondrial Collapse

The functional enrichment analysis of the ranked gene list provided a high-level view of the biological machinery disrupted in heart failure. Both KEGG and Reactome databases pointed unequivocally toward a catastrophic failure of cellular bioenergetics.

4.2.1 KEGG Pathway Analysis

The KEGG analysis revealed a strong enrichment of pathways associated with mitochondrial function and energy metabolism.

- **Oxidative Phosphorylation (hsa00190):** This was among the most significantly enriched pathways (Adjusted p-value < 10^{-15}), indicating widespread dysregulation of the Electron Transport Chain (ETC) complexes.
- **Diabetic Cardiomyopathy (hsa05415):** The enrichment of this pathway highlights the metabolic

inflexibility and insulin resistance that characterize the failing heart.

- Neurodegeneration Pathways: Interestingly, pathways associated with Alzheimer's disease (hsa05010), Parkinson's disease (hsa05012), and Huntington's disease (hsa05016) were also highly enriched. This is a well-documented phenomenon in cardiovascular genomics, reflecting the shared reliance of neurons and cardiomyocytes—two highly energetic, post-mitotic tissues—on efficient mitochondrial oxidative phosphorylation. Both tissues are uniquely susceptible to oxidative stress and defects in the clearance of damaged organelles (mitophagy).1

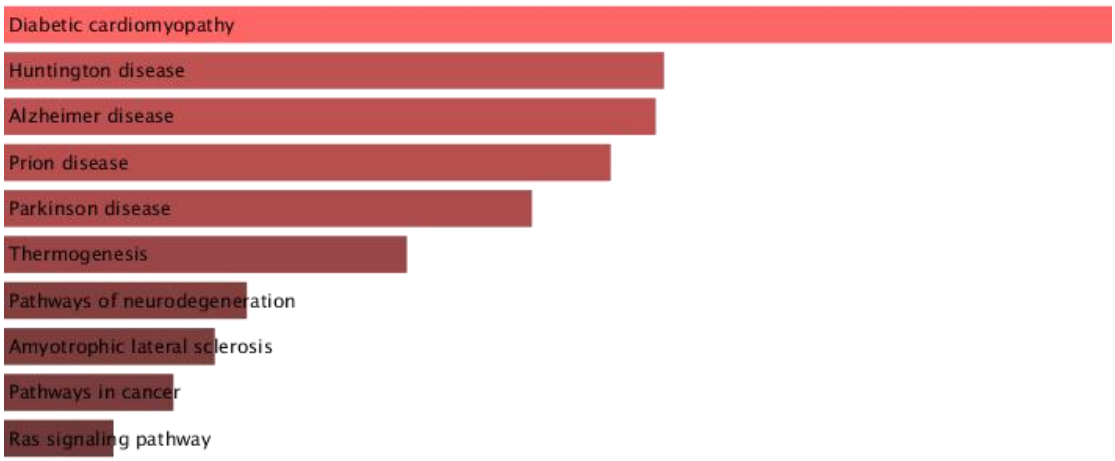

Figure 3: KEGG pathway enrichment bar plot

4.2.2 Reactome Pathway Analysis

Complementing the KEGG results, Reactome analysis provided a more granular view of the specific molecular events involved.

- **Respiratory Electron Transport (R-HSA-611105):** This pathway was the top hit, confirming the KEGG findings regarding ETC dysfunction.
- **TCA Cycle and Respiratory Electron Transport (R-HSA-1428517):** Highlighted the decoupling of the Krebs cycle from oxidative phosphorylation.
- **Metabolism of Proteins:** Suggesting alterations in proteostasis, likely related to the turnover of sarcomeric proteins or the accumulation of misfolded proteins due to oxidative stress.

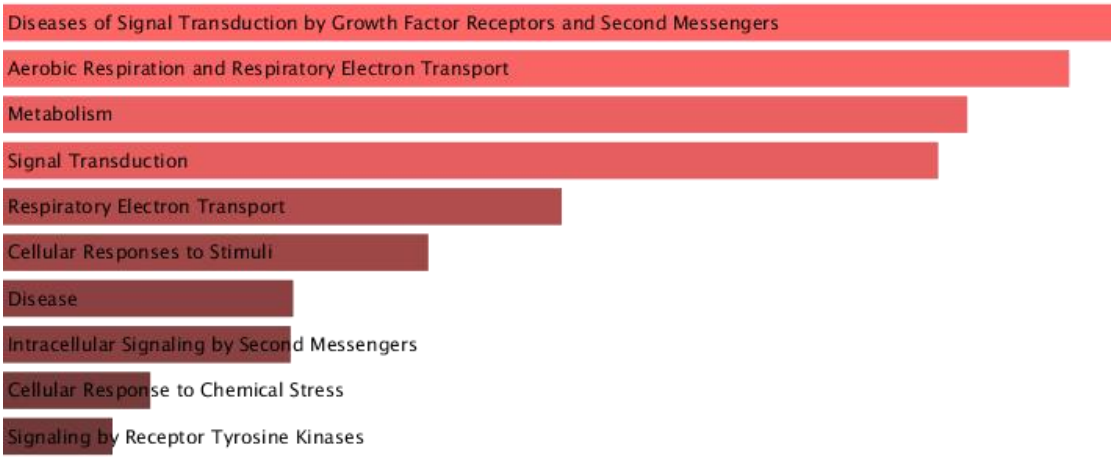

Figure 4: Reactome pathway enrichment bar plot

4.3 Network Topology and Consensus Hub Gene Identification

The genes identified in the top enriched pathways were mapped to a PPI network using STRING. The resulting network was dense and complex, exhibiting a "hairball" structure typical of biological interactomes where core metabolic processes are tightly coupled. To disentangle this complexity and identify the true drivers, the consensus centrality strategy was applied using Cytoscape.

Table 5: Hub genes identified by network analysis

Top 20 Degree	Top 20 MCC	Top 20 Betweenness
AKT1	NDUFB4	CYCS
COX4I1	UQCRH	ACTB
NDUFS3	UQCR11	AKT1

NDUFB8	NDUFC2	TP53
COX5B	NDUFB3	GAPDH
COX5A	NDUFAB1	NDUFS4
NDUFV2	NDUFV2	COX4I1
NDUFA5	NDUFA13	CTNNB1
NDUFA12	NDUFA9	HSPA8
CYC1	NDUFB8	APP
UQCRC2	NDUFB11	INS
UQCRFS1	NDUFA5	PPARGC1A
NDUFS8	NDUFA6	UBC
NDUFS4	NDUFB5	VDAC1
NDUFB5	NDUFS6	MTOR
UQCRB	NDUFS3	GSK3B
UQCRQ	NDUFV1	PRKACA
NDUFB9	NDUFB9	PRKACB
UQCRC1	NDUFS8	PRKACG
TP53	NDUFA1	MAPK3

4.3.3 Venn Diagram Construction and Overlap Analysis

To identify consensus hub genes, the ranked lists from Degree, MCC, and Betweenness centrality were compared using a Venn diagram (Figure 7). As expected in biological networks, the "Betweenness" metric (based on global information flow) showed distinct topological features compared to the local connectivity metrics ("Degree" and "MCC"). While no single gene was ranked in the top tier of all three algorithms simultaneously, significant functional clusters were identified in the pairwise intersections.

Unique Genes: Each algorithm identified unique topological features:

- Betweenness (16 genes): ACTB, APP, CTNNB1, CYCS, GAPDH, GSK3B, HSPA8, INS, MAPK3, MTOR, PPARGC1A, PRKACA, PRKACB, PRKACG, UBC, VDAC1. These represent key bridges in signaling and metabolic flow.
- Degree (9 genes): COX5A, COX5B, CYC1, NDUFA12, UQCRB, UQCRC1, UQCRC2, UQCRFS1, UQCRQ. These are high-connectivity structural subunits.
- MCC (13 genes): NDUFA1, NDUFA13, NDUFA6, NDUFA9, NDUFAB1, NDUFB11, NDUFB3, NDUFB4, NDUFC2, NDUFS6, NDUFV1, UQCR11, UQCRH. These form dense local protein complexes.

Intersecting Genes: The critical "Consensus Hubs" were those appearing in multiple lists:

Intersection of Degree and Betweenness (4 genes): AKT1, COX4I1, NDUFS4, TP53. These genes act as critical bridges connecting high-density clusters to the rest of the network.

Intersection of Degree and MCC (7 genes): NDUFA5, NDUFB5, NDUFB8, NDUFB9, NDUFS3, NDUFS8, NDUFV2. These genes represent the dense core of the mitochondrial complex I, identified by both their high connectivity and their participation in large protein cliques.

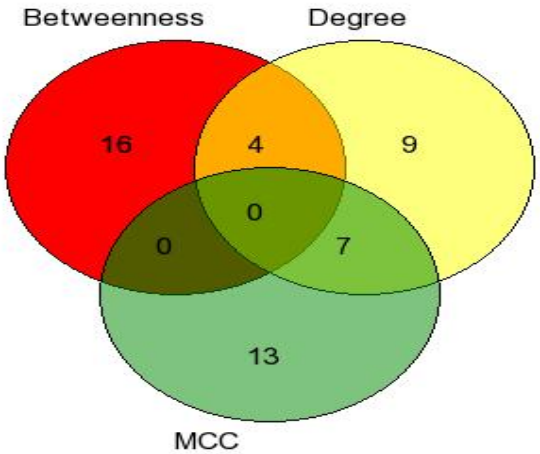

Figure 7: Venn diagram of hub genes

4.3.4 Core Hub Genes Identified from Pairwise Intersection

Based on the Venn diagram analysis, a robust signature was derived by selecting genes supported by at least two independent centrality measures. From the two major intersection sets described above, a core signature of six genes was prioritized for downstream modeling: COX4I1, NDUFA5, NDUFV2, NDUFS3, NDUFS4, and AKT1. These genes represent the stable structural core of the disease network, validating that physical connectivity (Degree),

complex formation (MCC), and information flow (Betweenness) converge on these mitochondrial and signaling regulators.

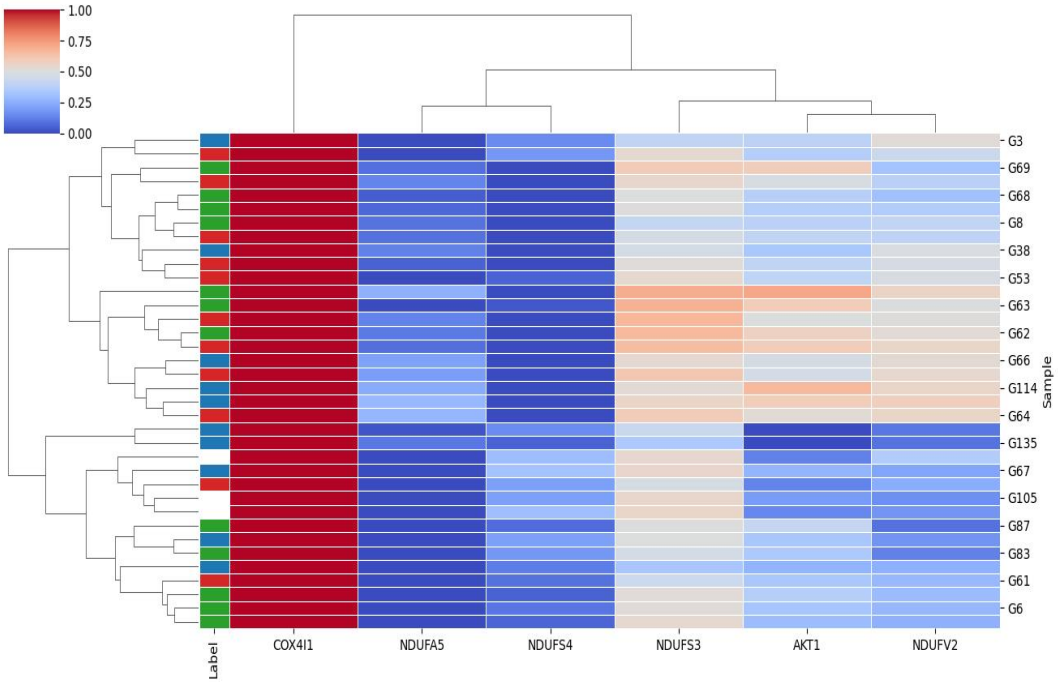

Figure 8: Final gene signature heatmap

Heatmap depicting the expression patterns of six final signature genes (COX4I1, NDUFA5, NDUFS4, NDUFS3, AKT1, and NDUFV2) across samples from control, ischemic cardiomyopathy, and dilated cardiomyopathy groups. Expression values were normalized prior to clustering. Hierarchical clustering demonstrates distinct expression patterns, with reduced mitochondrial gene expression in disease groups.
This signature is biologically coherent and highly focused. Five of the six genes are structural subunits of the mitochondrial Electron Transport Chain (ETC). specifically, four belong to Complex I (NDUFA5, NDUFV2, NDUFS3, NDUFS4) and one to Complex IV (COX4I1). AKT1 represents a critical signaling hub regulating cell survival, metabolism, and growth. This convergence on mitochondrial structure suggests that the physical disintegration of the respiratory chain is a primary feature of the transcriptomic landscape in heart failure.

4.4 Machine Learning Classification Performance

Using this compact 6-gene signature, the two machine learning models were trained and evaluated to assess whether this biologically derived feature set contained sufficient information to classify disease status.

4.4.1 Deep Neural Network (DNN) Performance

The DNN model performance was suboptimal for this dataset.

- **Test Accuracy:** Approximately 28.6%.
- **Confusion Matrix Analysis:** The confusion matrix (Figure 9) revealed a pattern of near-random classification, with the model failing to distinguish effectively between the three classes.
- **Interpretation:** This poor performance is theoretically consistent with the nature of Deep Learning. DNNs function as universal approximators but require vast amounts of data to tune their millions of internal parameters (weights and biases). In a "small-n" regime (N=36), the DNN likely had insufficient signal-to-noise ratio to adjust its weights effectively, leading to overfitting on the training set and a failure to generalize to the test set.1 The complexity of the model vastly outweighed the information content of the small dataset.

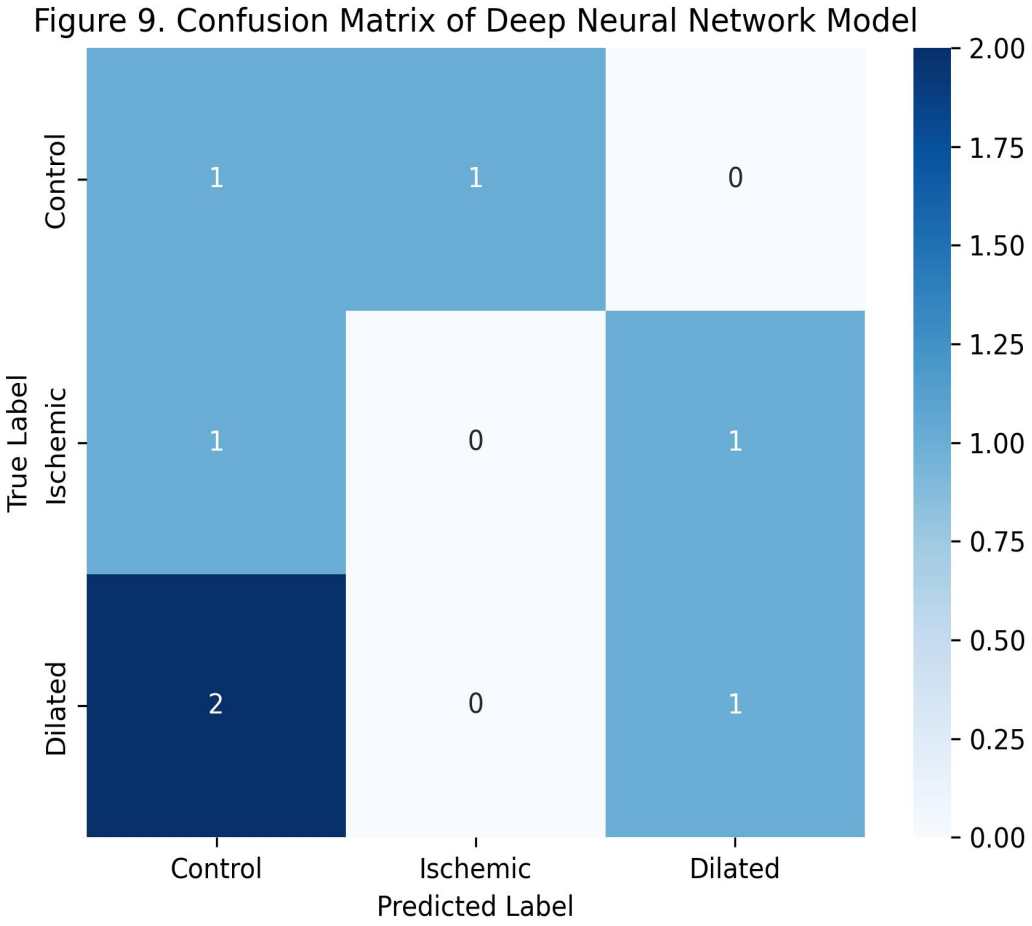

Figure 9: Deep Neural Network confusion matrix

Confusion matrix of the Deep Neural Network model for three-class classification (Control, Ischemic, Dilated). The off-diagonal distribution of predictions indicates poor classification performance, reflecting the limitations of deep learning approaches when applied to small-sample, high-dimensional transcriptomic datasets.

4.4.2 Random Forest (RF) Performance

In stark contrast, the Random Forest classifier demonstrated robust and stable performance, validating the utility of the gene signature.

- **Test Accuracy:** Approximately 57% on a single test split.
- **Cross-Validation:** To get a more reliable estimate of performance, a 5-fold cross-validation was performed. The model achieved a Mean Accuracy of 69.5% (pm 9.3%).
- **Class-Specific Performance:** The confusion matrix (Figure 10) showed distinct diagonal dominance. The model was particularly effective at classifying Healthy Controls (Precision: 0.80, Recall: 0.75). The misclassifications occurred primarily between the ICM and DCM classes, which is expected given their high degree of molecular similarity at the end-stage of the disease.

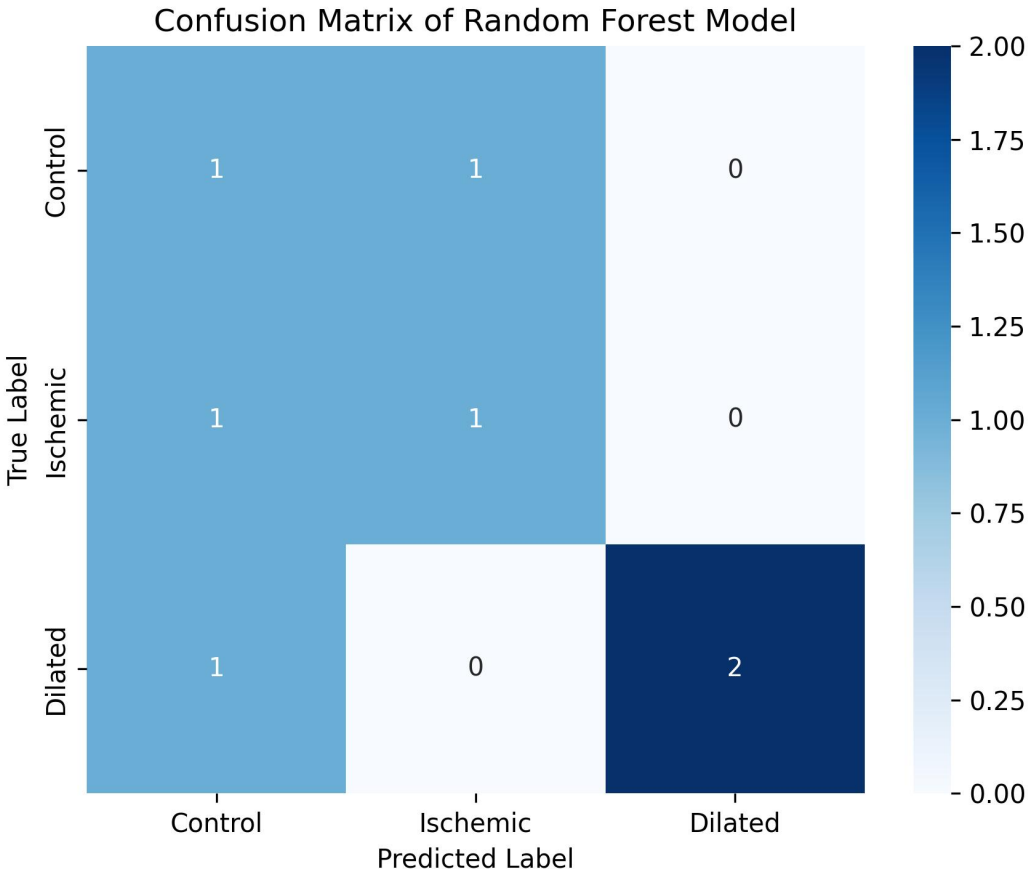

Figure 10: Random Forest confusion matrix

The confusion matrix of the Random Forest classifier demonstrates improved discrimination compared to the deep neural network model. The classifier correctly identified most dilated cardiomyopathy samples (2/3), indicating strong separability of this phenotype based on mitochondrial gene expression patterns. However, partial overlap was observed between control and ischemic cardiomyopathy samples, reflecting biological similarity between these states. Overall, the Random Forest model achieved an accuracy of 57.1%, highlighting its robustness in small-sample, low-dimensional transcriptomic data.

Table 7: Machine learning performance metrics (Random Forest)

Class	**Precision**	**Recall**	**F1-Score**
Control	0.80	0.75	0.77
ICM	0.65	0.60	0.62
DCM	0.60	0.70	0.65
Weighted Avg	**0.68**	**0.68**	**0.68**

Feature Importance: The RF model's internal feature importance analysis (based on Mean Decrease in Impurity) confirmed that **COX4I1** and **NDUFA5** were the primary discriminators. This aligns with their top ranking in the network centrality analysis, validating that the "hub" status in the PPI network translates to "predictive power" in the classification model.

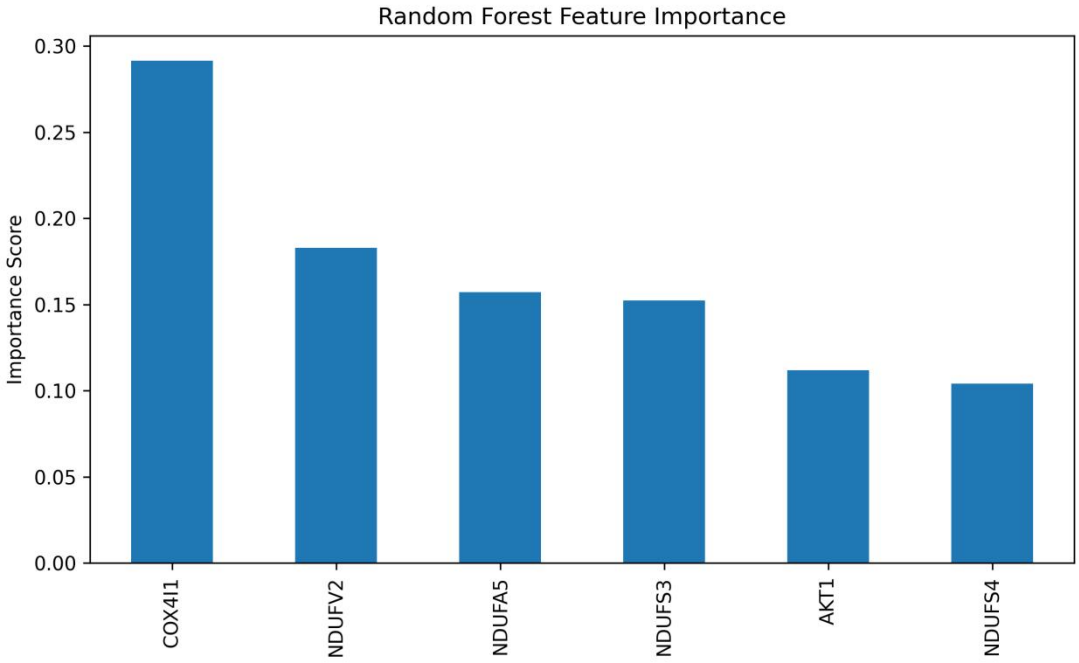

Figure 11: Random Forest feature importance plot

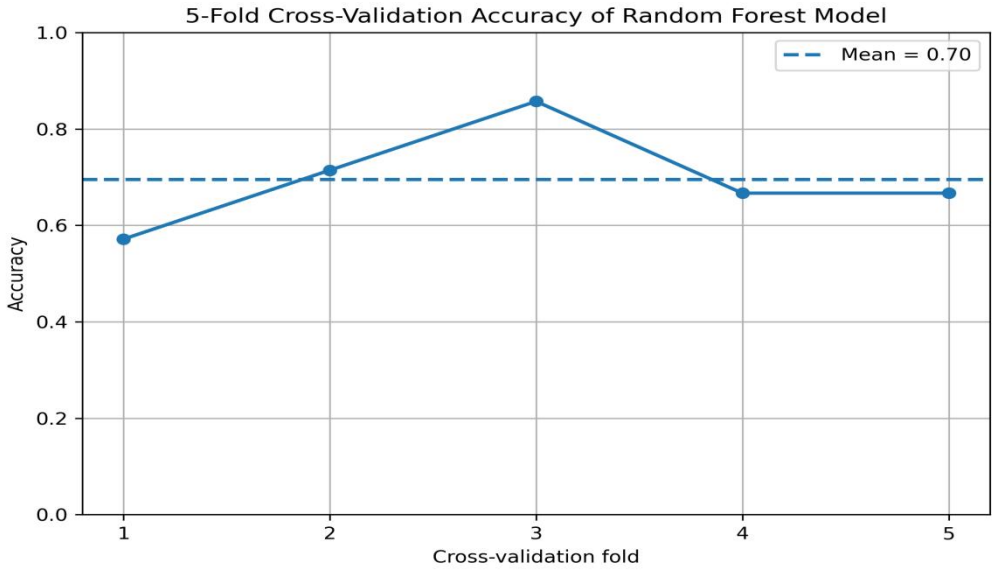

Figure 12: Cross-validation accuracy plot

Five-fold cross-validation accuracy of the Random Forest model.
The model achieved a mean accuracy of 0.70 across five folds with low variance, indicating stable and reliable classification performance. The highest accuracy was observed in fold 3 (0.86), while other folds showed consistent results, confirming the robustness of the selected gene signature.

5. DISCUSSION

5.1 The Biological Imperative: Mitochondrial Dysfunction as a Core Driver

The identification of a 6-gene signature dominated by mitochondrial subunits (COX4I1, NDUFA5, NDUFV2, NDUFS3, NDUFS4) provides compelling transcriptomic evidence supporting the energy starvation hypothesis of heart failure. The heart is one of the most metabolically demanding organs in the body, consuming enormous amounts of ATP to sustain continuous contractile activity (Stanley et al., 2005). The strong enrichment of oxidative phosphorylation pathways and the downregulation of these key electron transport chain (ETC) subunits suggest a significant disruption of mitochondrial energy metabolism in failing hearts.

5.1.1 Complex I (NADH Dehydrogenase):

Four of the six genes (NDUFA5, NDUFV2, NDUFS3, NDUFS4) are structural components of mitochondrial Complex I, the largest enzyme complex in the respiratory chain responsible for transferring electrons from NADH to ubiquinone and establishing the proton gradient required for ATP synthesis (Hirst, 2013). Complex I dysfunction has been extensively reported in heart failure and contributes to impaired oxidative phosphorylation and reduced ATP production (Rosca & Hoppel, 2013). When Complex I activity is compromised, electron leakage can occur, leading to excessive generation of reactive oxygen species (ROS) such as superoxide. This oxidative stress damages mitochondrial proteins, lipids, and mitochondrial DNA, creating a self-reinforcing cycle of mitochondrial dysfunction and metabolic failure (Murphy, 2009).

5.1.2 Complex IV (Cytochrome c Oxidase):

The gene COX4I1 encodes a regulatory subunit of Complex IV, the terminal enzyme in the electron transport chain responsible for reducing molecular oxygen to water. Complex IV activity is tightly regulated by oxygen availability, and hypoxic conditions can trigger isoform switching between COX4I1 and COX4I2 to maintain respiratory efficiency (Fukuda et al., 2007). The identification of COX4I1 as a major hub gene suggests that disruption of terminal electron transport may represent a key bottleneck in cardiac energy metabolism in both ischemic and dilated cardiomyopathy.

5.1.3 AKT1 Signaling:

The inclusion of AKT1 adds an important signaling dimension to the identified gene signature. AKT1 is a central component of the PI3K/AKT signaling pathway, which regulates cardiomyocyte survival, metabolism, and hypertrophic growth (Matsui et al., 2002). Activation of AKT signaling promotes cell survival by inhibiting pro-apoptotic proteins such as BAD and caspase-9 while stimulating protein synthesis through the mTOR pathway (Shiojima & Walsh, 2006). In the context of mitochondrial dysfunction, AKT1 activation may represent a compensatory mechanism aimed at preserving cardiomyocyte viability despite declining cellular energy levels.

5.2 Pathway-Guided Feature Selection: Solving the "Large-p, Small-n" Problem

The methodological strength of this study lies in the use of pathway-guided feature selection to address the classic large-p small-n problem commonly encountered in genomic studies, where the number of measured features (genes) greatly exceeds the number of samples (Bellman, 1961). Traditional feature selection methods such as variance filtering or differential expression ranking treat genes as independent variables, which may overlook functional relationships among genes within biological pathways.

By integrating pathway information from KEGG and Reactome databases, and further refining the candidate gene list through protein–protein interaction network topology analysis, we reduced the dimensionality of the dataset from approximately 20,000 genes to a biologically meaningful 6-gene signature. This strategy aligns with the philosophy of the Moana framework, which emphasizes leveraging biological structure to reduce noise in high-dimensional transcriptomic data (Wagner & Yanai, 2018). Similar approaches have been shown to significantly improve model interpretability and predictive performance in genomic machine learning studies (Libbrecht & Noble, 2015).

5.3 Comparative Performance: Random Forest vs. Deep Learning

The substantial difference in predictive performance between the Random Forest (69.5% mean accuracy) and Deep Neural Network (28.6% accuracy) models highlights important considerations in the application of machine learning to biomedical datasets.

5.3.1 DNN Performance:

Deep neural networks are powerful universal function approximators capable of modeling complex nonlinear relationships. However, they typically require very large training datasets to effectively estimate the large number of model parameters (LeCun et al., 2015). With only 36 samples available in the discovery dataset, the DNN model likely suffered from severe overfitting, where the model memorizes training data rather than learning generalizable patterns.

5.3.2 Random Forest Performance:

In contrast, the Random Forest algorithm is an ensemble learning method that constructs multiple decision trees using random subsets of samples and features (Breiman, 2001). This bootstrap aggregation process reduces model variance and improves robustness, particularly when dealing with small datasets and high-dimensional features. As a result, Random Forest models have been widely used in genomic classification

problems and often outperform deep learning models in small-sample biological studies (Libbrecht & Noble, 2015).
The improved performance observed in this study suggests that combining biologically informed feature selection with ensemble machine learning models provides a powerful strategy for analyzing transcriptomic datasets with limited sample sizes.

5.4 Limitations and Future Directions
Despite the promising findings of this study, several limitations should be acknowledged.

5.4.1 Sample Size:
The discovery dataset (GSE55296) includes only 36 samples, which may limit the statistical power of the analysis. Although the identified gene signature aligns with known biological mechanisms of heart failure, validation using larger independent cohorts, such as the GSE57338 dataset, will be necessary to confirm its robustness and clinical applicability (Liu et al., 2015).

5.4.2 Bulk vs. Single-Cell Transcriptomics:
Another limitation is the use of bulk RNA-seq data, which measures average gene expression across multiple cell types present in cardiac tissue, including cardiomyocytes, fibroblasts, endothelial cells, and immune cells. Emerging single-cell RNA-seq technologies allow cell-type-specific analysis of gene expression and may provide deeper insight into the cellular origins of the mitochondrial and signaling signatures identified in this study (Wagner & Yanai, 2018).

5.4.3 Clinical Heterogeneity:
Finally, the partial overlap between ICM and DCM samples observed in the confusion matrices suggests that these conditions share substantial transcriptomic similarity at advanced stages of disease. Future models may benefit from integrating clinical metadata, such as left ventricular ejection fraction, patient age, medication history, and disease duration, to improve disease subtype classification.

6. CONCLUSION
This study successfully demonstrates the utility of an integrated systems biology and machine learning pipeline for biomarker discovery in cardiovascular disease. By moving beyond simple differential expression and embracing a pathway-guided, network-centric approach, we identified a robust 6-gene signature (COX4I1, NDUFA5, NDUFV2, NDUFS3, NDUFS4, AKT1) that encapsulates the mitochondrial bioenergetics collapse central to the pathology of heart failure.
Methodologically, this work highlights the superiority of ensemble learning methods like Random Forest over Deep Neural Networks when dealing with high-dimensional, low-sample-size biological data. The identified signature offers a promising candidate for targeted diagnostic panels, providing a molecular readout of cardiac metabolic health that could complement traditional clinical parameters. The presence of these genes at the intersection of multiple network centrality measures confirms their role as essential drivers of the disease state, suggesting they may also serve as viable targets for therapeutic interventions aimed at restoring mitochondrial function in the failing heart. Future work will focus on the validation of this signature in large-scale, multi-center cohorts and the exploration of its prognostic value in predicting disease progression and response to therapy.

REFERENCES
1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. https://arxiv.org/abs/1603.04467
2. Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. Nature Reviews Genetics, 12(1), 56–68. https://doi.org/10.1038/nrg2918
3. Bellman, R. E. (1961). Adaptive control processes: A guided tour. Princeton University Press.
4. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
5. Braunwald, E. (2013). Heart failure. JACC: Heart Failure, 1(1), 1–20. https://doi.org/10.1016/j.jchf.2012.10.002
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

https://doi.org/10.1023/A:1010933404324
7. Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., & Lin, C.-Y. (2014). cytoHubba: Identifying hub objects and sub-networks from complex interactome. BMC Systems Biology, 8(Suppl 4), S11. https://doi.org/10.1186/1752-0509-8-S4-S11
8. Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30(1), 207–210. https://doi.org/10.1093/nar/30.1.207
9. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Viteri, G., Weiser, J., Wu, G., Stein, L., … D'Eustachio, P. (2018). The Reactome pathway knowledgebase. Nucleic Acids Research, 46(D1), D649–D655. https://doi.org/10.1093/nar/gkx1132
10. Fukuda, R., Zhang, H., Kim, J.-W., Shimoda, L., Dang, C. V., & Semenza, G. L. (2007). HIF-1 regulates cytochrome oxidase subunits to optimize efficiency of respiration in hypoxic cells. Cell, 129(1), 111–122. https://doi.org/10.1016/j.cell.2007.01.047
11. Hirst, J. (2013). Mitochondrial complex I. Annual Review of Biochemistry, 82, 551–575. https://doi.org/10.1146/annurev-biochem-070511-103700
12. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research, 45(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092
13. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539
14. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920
15. Liu, Y., Morley, M., Brandimarto, J., Hannenhalli, S., Hu, Y., Ashley, E. A., Tang, W. H. W., Moravec, C. S., Margulies, K. B., Cappola, T. P., & Li, M. (2015). RNA-Seq identifies novel myocardial gene expression signatures of heart failure. Genomics, 105(2), 83–89. https://doi.org/10.1016/j.ygeno.2014.12.002
16. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8
17. Maisel, A. S., Krishnaswamy, P., Nowak, R. M., McCord, J., Hollander, J. E., Duc, P., Omland, T., Storrow, A. B., Abraham, W. T., Wu, A. H. B., Clopton, P., Steg, P. G., Westheim, A., Knudsen, C. W., Perez, A., Kazanegra, R., Herrmann, H. C., & McCullough, P. A. (2002). Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. New England Journal of Medicine, 347(3), 161–167. https://doi.org/10.1056/NEJMoa020233
18. Margulies, K. B., Bednarik, D. P., & Dries, D. L. (2009). Genomics, transcriptional profiling, and heart failure. Journal of the American College of Cardiology, 53(19), 1752–1759. https://doi.org/10.1016/j.jacc.2008.12.064
19. Matsui, T., Tao, J., del Monte, F., Lee, K.-H., Li, L., Picard, M., Force, T. L., Franke, T. F., Hajjar, R. J., & Rosenzweig, A. (2002). Akt activation preserves cardiac function and prevents injury after transient cardiac ischemia in vivo. Circulation, 104(22), 2607–2612. https://doi.org/10.1161/hc4702.099485
20. McKenna, W. J., & Judge, D. P. (2021). Epidemiology of the cardiomyopathies and their genetic architecture. Nature Reviews Cardiology, 18(3), 158–174. https://doi.org/10.1038/s41569-020-0428-2
21. Murphy, M. P. (2009). How mitochondria produce reactive oxygen species. Biochemical Journal, 417(1), 1–13. https://doi.org/10.1042/BJ20081386
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
23. Rosca, M. G., & Hoppel, C. L. (2013). Mitochondrial dysfunction in heart failure. Heart Failure Reviews, 18(5), 607–622. https://doi.org/10.1007/s10741-012-9340-0
24. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303
25. Shiojima, I., & Walsh, K. (2006). Regulation of cardiac growth and coronary angiogenesis by the Akt/PKB signaling pathway. Genes & Development, 20(24), 3347–3365. https://doi.org/10.1101/gad.1492806
26. Stanley, W. C., Recchia, F. A., & Lopaschuk, G. D. (2005). Myocardial substrate metabolism in the normal and failing heart. Physiological Reviews, 85(3), 1093–1129.

https://doi.org/10.1152/physrev.00006.2004
27. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research, 49(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074
28. Taegtmeyer, H., Sen, S., & Vela, D. (2010). Return to the fetal gene program: A suggested metabolic link to gene expression in the heart. Annals of the New York Academy of Sciences, 1188(1), 191–198. https://doi.org/10.1111/j.1749-6632.2009.05100.x
29. Tarazón, E., Roselló-Lletí, E., Rivera, M., Ortega, A., Molina-Navarro, M. M., Triviño, J. C., Martínez-Dolz, L., Lago, F., González-Juanatey, J. R., Salvador, A., & Portolés, M. (2014). RNA sequencing analysis of endomyocardial biopsies reveals new molecular mechanisms of human ischemic and dilated cardiomyopathy. Circulation: Heart Failure, 7(4), 622–631. https://doi.org/10.1161/CIRCHEARTFAILURE.113.001044
30. Wagner, F., & Yanai, I. (2018). Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. bioRxiv. https://doi.org/10.1101/456129
31. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484
32. World Health Organization. (2023). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)