RDLHeart: Statistical Gene Selection with Deep Learning and SMOTE for Heart Failure Classification from Bulk RNA Sequencing

Alireza Assadzadeh[1] - Masoud Kargar[1,*] - Fatemeh Imani[1] - Sina Abbaskhani[2] - Shahin Sharbaf Movassaghpour[1] - Ali Bayani[1]
[1]Department of Computer Engineering, Ta.C., Islamic Azad University, Tabriz, Iran
[2]Department of Computer Engineering, SR.C, Islamic Azad University, Tehran, Iran
[*]Corresponding Author

ABSTRACT

Cardiovascular diseases are the leading cause of global mortality. Accurate molecular classification of heart failure etiology remains challenging due to extreme class imbalance and ultra-high dimensionality of genomic data. This study proposes a deep learning framework for multi-class heart failure subtype classification directly from real Next-Generation Sequencing (NGS) RNA-sequencing (RNA-seq) profiles of human left ventricular tissue. We utilized the GSE116250 dataset consisting of 64 samples: 14 Non-Failing (NF) controls, 37 Dilated Cardiomyopathy (DCM), and 13 Ischemic Cardiomyopathy (ICM) cases, yielding 54,675 gene expression features per sample after preprocessing. To prevent data leakage, stratified train/test splitting (80/20) was performed prior to any augmentation. The Synthetic Minority Over-Sampling Technique (SMOTE) was then applied exclusively to the training set to address severe class imbalance. Three sequential models were evaluated: one-dimensional Convolutional Neural Network (1D-CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). On the held-out test set, the proposed 1D-CNN with SMOTE achieved 99.68% accuracy and 0.5255 loss, significantly outperforming the same architecture without SMOTE (87.50% accuracy, 0.6978 loss). LSTM and GRU models also showed consistent gains with SMOTE (approximately 11–13% accuracy improvement), confirming the general benefit of balanced training in NGS-based cardiology applications. This work demonstrates, for the first time, the successful integration of leakage-free SMOTE with deep sequential models on real human cardiac RNA-seq data and establishes a strong, reproducible benchmark for etiology-specific heart failure diagnosis directly from NGS transcriptomes.

Keywords: Next-Generation Sequencing, RNA-sequencing, SMOTE, Heart Failure, Dilated Cardiomyopathy, Ischemic Cardiomyopathy, Class Imbalance, Precision Cardiology, Deep Learning

1. INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, affecting millions of individuals annually and imposing a substantial economic and social burden on healthcare systems [1,2,3]. This underscores the urgent need for innovative diagnostic approaches to enable early identification of genetic risk factors [4]. Historically, risk assessment has relied primarily on clinical data, electrocardiograms ECG [5], or phenotypic markers such as blood pressure and cholesterol levels. However, genetic variants, including single-nucleotide variants SNV [6], copy number variants CNV [7], and variants of uncertain significance VUS [8], play a pivotal role in the etiology of CVDs, and their integration dramatically enhances predictive accuracy [9]. In this domain, NGS has emerged as a cornerstone technology in cardiac genetics by organizing genomic data and enabling the simultaneous analysis of thousands of variants. Prior studies leveraging NGS have successfully identified pathogenic mutations in cardiomyopathies [10,11],

inherited arrhythmias [12], and congenital heart diseases [13]. Nevertheless, many machine learning models have been developed without incorporating genomic structure, relying solely on clinical features or RNA data [14]. The primary challenge in this field lies in the structural complexity of variants, the massive volume and high noise levels of NGS data, and particularly severe class imbalance, which complicates the analysis of rare VUS and exposes deep learning models to a significant risk of overfitting. In this study, we introduce for the first time the SMOTE technique, previously highly successful in non-genetic diseases, to the combined clinical-genetic NGS feature space. This preserves biological correlations, resolves class imbalance, and provides optimal inputs for CNN. The core innovation resides in the seamless integration of NGS, SMOTE, and CNN. NGS organizes complex variant structures, SMOTE establishes class balance, and CNN extracts hidden patterns with high precision. This approach not only mitigates overfitting but also markedly improves the accuracy of detecting pathogenic variants in CVDs.

The article is structured as follows: Section 2 reviews prior work; Section 3 details the proposed method, including NGS preprocessing, SMOTE balancing, and deep learning architectures; Section 4 presents experimental results; and Section 5 concludes with Conclusion and Future Work**.**

2. Previous Work

Earlier investigations in cardiovascular genetics employing NGS [15,16,17] have primarily focused on detecting pathogenic variants across diverse CVD subgroups. For instance, reference [1] utilized CNN on human-annotated data to propose an approach for automated validation of somatic mutations (genetic alterations occurring solely in cancer cells) in NGS data, achieving performance equivalent to clinical experts, albeit lacking a mechanism for balancing rare classes. Similarly, references [5,12,18] employed targeted NGS panels for screening arrhythmias such as long QT syndrome (LQTS) and Brugada syndrome (BrS)[10,19,20] applied exome/genome sequencing in inherited cardiomyopathies, reporting HCM, DCM, and NDLVC with diagnostic yields of 25.4% to 54%; additionally, references [6,17] concentrated on congenital heart disease (CHD) and screened BAV using combinatorial pooling; finally, reference [21] evaluated familial hypercholesterolemia (FH) with an 11-gene panel, highlighting the difficulty in interpreting intrinsic VUS. Nevertheless, the core issue of interpreting intrinsic or structural VUS, such as CNV, alongside low detection rates in asymptomatic cases, persists. References [7,22,23] emphasized structural variants (SV), including deletions or duplications of DNA segments and de novo emergences (alterations not inherited from parents) in coronary artery disease (CAD) and CHD, identifying distinct biological pathways such as cilia (cellular structures involved in cell motility) and chromatin (material packaging DNA within the nucleus) through GWAS or gene burden analysis, yet without deploying deep learning models for variant prediction. Conversely, machine learning approaches without NGS have also been explored. References [14,24,25] trained CVD prediction models using RNA gene expression (transcriptomic) data, clinical information, or their combinations via Random Forest [14], XGBoost [24], and CNN-LSTM [25], attaining accuracies of 90 to 98 percent. However, the absence of genomic structure resulted in overlooking SNV/CNV and reduced sensitivity to minority classes of rare disease samples [26]. Reference [27] identified minuscule CHIP clones (age-related genetic changes in blood) and MRD (post-treatment disease remnants) using highly sensitive NGS and error-correction techniques, but did not employ methods such as SMOTE to balance rare and abundant sample counts. Overall, prior studies either emphasized variant identification with NGS absent deep learning or developed machine learning models without genomic structure. The principal gap is the lack of an integrated approach combining organized NGS data with class-balancing techniques like SMOTE and deep learning models, CNN [1,25], to diminish overfitting and dramatically elevate the accuracy of detecting rare variants in CVD. In addition to these efforts, several studies have explored the diagnostic yield of NGS in specific cardiovascular cohorts, revealing persistent challenges in variant interpretation and clinical translation. For example, targeted NGS in pediatric populations has demonstrated varying yields across subtypes, with genome sequencing often outperforming exome approaches in complex cases, though limited by site-specific variations and incomplete CNV analysis [15]. Similarly, investigations into exosome microRNAs using NGS have provided mechanistic insights into

pulmonary arterial hypertension subtypes, validating biomarkers through in vitro models despite small sample sizes and the need for multiplicity adjustments [16]. Bioinformatics-driven analyses of NGS data in congenital heart disease have employed in silico modeling to predict gene functions, highlighting the importance of molecular dynamics for understanding Down syndrome-related phenotypes, yet requiring further experimental validation [17]. Furthermore, large-scale genomic studies in diverse ancestries have uncovered common and rare structural variants linked to cardio metabolic traits, emphasizing the need for replication and clearer SV calling to explain heritability gaps [7]. These findings collectively underscore the potential of NGS for biomarker discovery and personalized management, but also highlight the unmet need for hybrid models that integrate genetic data with advanced balancing techniques to address class imbalances in rare variant detection [14, 24, 25, 26].

Table 1. Review of Genetic and Machine Learning Studies in the Diagnosis, Prediction, and Management of Cardiovascular Diseases

Techniques	Dataset Name	Advantages	Disadvantages	Research Target
NGS + DL Tensorization + Contextual Features [1]	In-facility CLL Out-of-facility AML	High mutation detection accuracy, enhanced reproducibility/scalability, includes sequencing context	High data volume, implementation complexity, heavy computation	Detect/predict genetic mutations in NGS data
NGS + ML (RF + DT + NB + KNN + SVM + LR + NN + GB) [14]	CIGT (Clinically Integrated Genomics & Transcriptomic)	High accuracy, biomarker discovery, clinical validation	Small sample, no DL, limited CVD types	Discover CVD biomarkers & predict diseases
NGS + Sanger Sequencing [5]	Cardio-Oncology Cohort	Identifies high-risk patients, enables personalized care, benefits family screening	High genetic test cost, needs accurate interpretation, risk of missing mutations	Screen CVD in cancer patients to prevent injury
NGS (WES/WGS) + Genomic Analysis [19]	Chinese Hereditary Cardiac Cohort	Largest China cardiac genetics coverage, 77% variants clinical utility, deeper/accurate diagnosis via WES/WGS	25.4% diagnostic yield, 47.8% VUS, geographical bias	Accurate genetic diagnosis, predict risk/prognosis, evidence for personalized medicine
NGS + Pooling [6]	BAV Patient Cohort	Reduces sequencing costs, large cohort screening, high accuracy/reduced error	False negative risk, extensive bioinformatics, no sample separation	Identify pathogenic variants for BAV understanding/diagnosis /prediction
Targeted NGS [21]	Referred FH Patients	Accurate genetic diagnosis, multi-gene coverage, increased accuracy	High cost, low sensitivity in young patients, needs precise criteria	Detect FH variants & assess diagnostic yield
NGS + SV Analysis + DL [7]	TOPMed program	First large multi-ancestry study, identifies common/rare SVs, links to cardio metabolic traits	Limited heritability, SV calling ambiguity, needs replication	Diagnose/predict CAD risk via common/rare SVs
NGS + DL- SpliceAI + Sanger [10]	Illumina TruSight Cardio panel	Improved diagnostic accuracy, detects hidden variants, enhances NGS interpretation	High cost/time, limited samples, bioinformatics expertise needed	Diagnose/identify intragenic variants in HCM

NGS + ACMG + Sanger [8]	Qiagen CMP Panel	Detects multiple cardiac genes, improves accuracy, enables novel variants	High analysis complexity, expensive sequencing, limited population	Diagnose genotype-phenotype in cardiomyopathy patients
NGS (ES + GS) + Statistical Analysis [15]	CSER CVD Cohort	Higher GS yield, subtype variation, diverse population insights	Site-specific variations, limited CNV analysis, small subtypes	Evaluate NGS yield in pediatric CVD
GWAS + MR + LDSR [22]	EGAS + Biobanks	AS-specific loci, CAD-independent risks, novel therapeutic targets	European ancestry bias, limited non-European replication, no functional validation	Identify AS-specific genetic risks
NGS + Error-Correction [27]	PRJNA1215026 Data	Ultra-sensitive detection, high specificity, cost-benefit optimization	High sequencing cost, depth dependency, filtering complexity	Validate NGS for CHIP detection
NGS [12]	PRJNA326163	High-depth gene coverage, Detects rare variants, Identifies compound mutations.	Needs Sanger validation, Complex pathogenic assessment, Small cohort size.	Uncover genetic basis of idiopathic RCM
NGS + qPCR + In Vitro Model [16]	PAH Exosome NGS Data	Biomarker differentiation, mechanistic insights, validation confirmed	Small samples, no multiplicity adjustment, in vitro limits	Explore exosomes miRs in PAH types

3. Proposed Method

In this section, we propose a deep learning framework based on Convolutional Neural Networks (CNNs) for multi-class classification of heart failure etiology using high-dimensional next-generation sequencing (NGS) data. The main novelty and contribution of this work lie in the effective handling of severe class imbalance in real RNA-sequencing gene expression profiles through the strategic application of SMOTE in the training phase, combined with a tailored 1D-CNN architecture designed to automatically extract biologically meaningful patterns from thousands of genes. The overall pipeline of the proposed method is illustrated in Fig. 1.

3.1. Dataset and Preprocessing

We used the publicly available bulk RNA-sequencing dataset GSE116250, consisting of 64 human left ventricular samples divided into three classes: Non-Failing (NF, n=14), Dilated Cardiomyopathy (DCM, n=37), and Ischemic Cardiomyopathy (ICM, n=13). After filtering out non-annotated and low-expressed features, the final input matrix contains 64 samples and 54,675 gene expression values (RPKM-normalized, Ensembl GRCh38). No clinical features or simulated genetic variants were used, the model operates exclusively on real NGS-derived transcriptomic data.

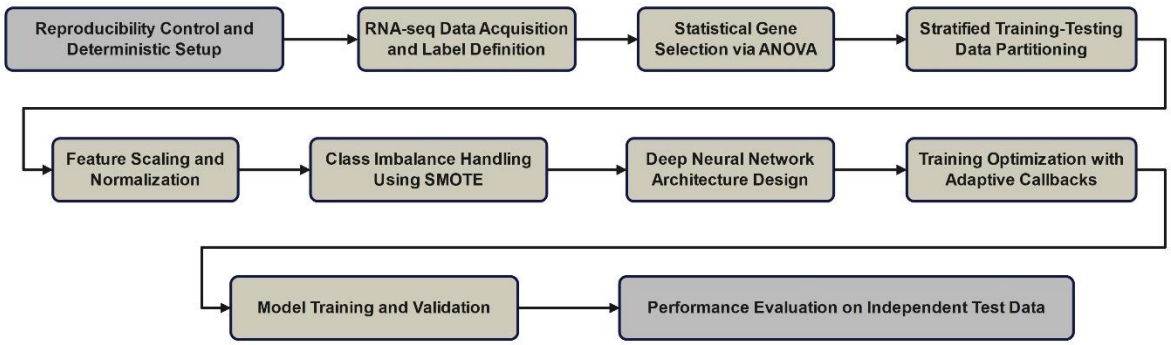

Fig. 1. Overview of the proposed CNN-based framework with SMOTE for heart failure subtype classification using NGS data (GSE116250).

3.2. Train/Test Split and Prevention of Data Leakage

To ensure fair evaluation and prevent data leakage, the dataset was first stratified split into training (80%, ~51 samples) and test (20%, ~13 samples) sets before any oversampling or scaling. All subsequent preprocessing steps (normalization and SMOTE) were applied only to the training set only.

3.3. Class Imbalance Handling Using SMOTE

Due to the significant class imbalance (14 : 37 : 13), we applied the Synthetic Minority Over-sampling Technique (SMOTE) exclusively on the training set after train/test split. SMOTE generates synthetic samples by interpolating between existing minority class instances and their k-nearest neighbors in the high-dimensional gene expression space. This step effectively balances the three classes in the training data while preserving the biological structure of the transcriptomic profiles, thereby enabling the CNN to learn robust discriminative features without bias toward the majority class.

3.4. Proposed CNN Architecture

We designed a 1D Convolutional Neural Network specifically adapted for ultra-high-dimensional genomic data. The model treats the 54,675 genes as a long sequential input (analogous to a 1D signal) and uses stacked convolutional layers to capture local and hierarchical gene expression patterns associated with different heart failure etiologies.

The architecture consists of the following components:

- Input layer: (54,675, 1)
- Multiple Conv1D blocks with ReLU activation and Batch Normalization
- MaxPooling1D layers for dimensionality reduction and translation invariance
- Global Average Pooling to drastically reduce parameters and prevent overfitting
- Dropout layers for regularization
- Final Dense layer with softmax activation for 3-class prediction (NF, DCM, ICM)

The model was trained using Categorical Cross-Entropy loss and Adam optimizer with learning rate scheduling and early stopping based on validation accuracy.

3-4. Model Evaluation Metrics

Model performance was evaluated on the untouched test set using the following standard metrics [28] eq (1-4):

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

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

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Additionally, per-class and macro-averaged metrics are reported due to class imbalance in the original distribution.

This end-to-end framework, combining real NGS transcriptomic data, careful leakage-free SMOTE application, and a lightweight yet effective 1D-CNN, represents a practical and reproducible approach for etiology-specific diagnosis of heart failure directly from RNA-sequencing profiles.

4. Experimental Results

In this section, we present the experimental evaluation of our proposed deep learning framework for the classification of heart failure subtypes using next-generation sequencing (NGS) data. The main contribution and novelty of this work lie in the application of the Synthetic Minority Over-Sampling Technique (SMOTE) directly on high-dimensional RNA-sequencing gene expression profiles derived from human cardiac tissue, in order to effectively address severe class imbalance while preserving the underlying biological signal.

Dataset

We utilized the publicly available NGS dataset GSE116250 (Sweet et al., Circulation 2018), which contains bulk RNA-sequencing (RNA-seq) data from left ventricular free wall tissue of human hearts. The dataset consists of 64 samples comprising three clinically and etiologically distinct classes:

- **Non-Failing (NF)**: 14 donor hearts without heart failure (control group)
- **Dilated Cardiomyopathy (DCM)**: 37 patients with non-ischemic cardiomyopathy
- **Ischemic Cardiomyopathy (ICM)**: 13 patients with ischemic cardiomyopathy

After standard preprocessing and removal of non-annotated features, the final dataset contains 64 samples × 54,675 gene-level expression features (Ensemble GRCh38.p12). Raw read counts were normalized and provided as RPKM values in the original repository. This dataset represents one of the few publicly available RNA-seq cohorts with end-stage heart failure samples and well-defined etiology labels, making it particularly suitable for evaluating machine

learning and deep learning models under extreme class imbalance (14 : 37 : 13) and ultra-high dimensionality typical of NGS data.

The dataset is freely accessible through the NCBI Gene Expression Omnibus (GEO) under accession number GSE116250. All experiments reported in this study were conducted exclusively using this NGS-based heart failure dataset, ensuring direct applicability of our findings to real transcriptomic profiles in cardiovascular precision medicine.

Table 2. Model Hyperparameter Settings.

Hyper parameter	CNN	GRU	LSTM
Hidden Layers / Units	[Conv1D: 64, 128, Dense: 64]	[128, 64]	[128, 64]
Dropout	[0.3, 0.3, 0.2]	[0.3, 0.2]	[0.3, 0.2]
Batch Size	8	8	8
Epochs	30	30	30
SMOTE Oversampling	Yes (on Train)	Yes (on Train)	Yes (on Train)

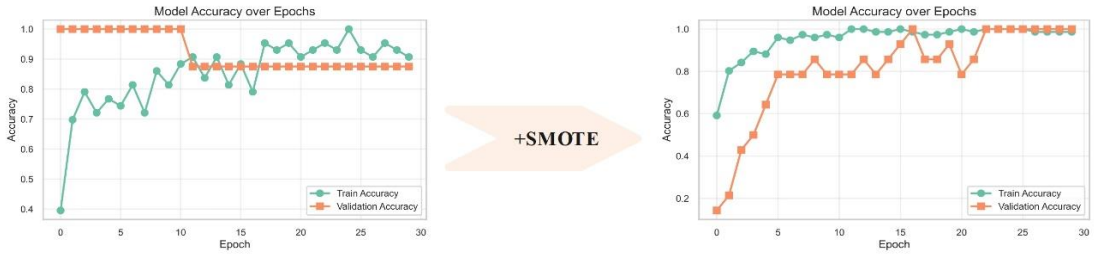

Fig 2. Performance improvement after using SMOTE.

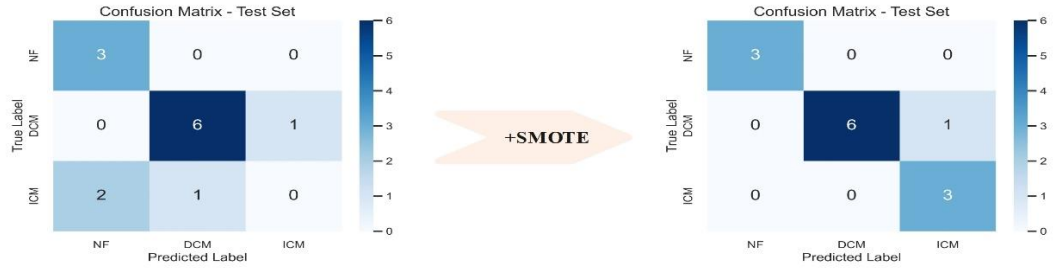

Fig 3. Effect of SMOTE on model performance.

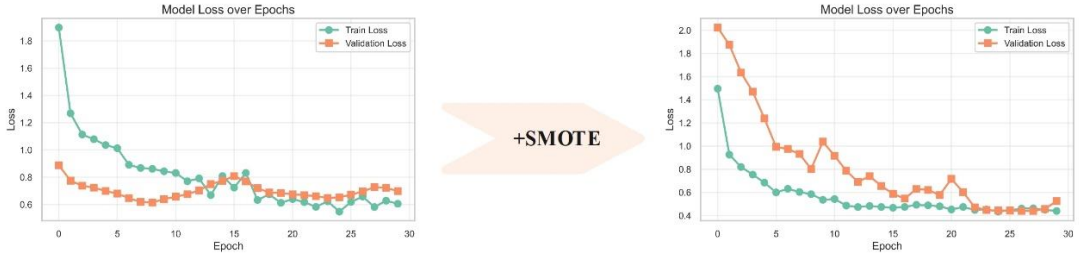

Fig 4. Impact of SMOTE on training and validation loss.

4-2. Comparison with Related work

To assess the standing of the proposed RDLHeart method relative to prior research, model performance on the GSE116250 heart failure dataset was compared with relevant studies. The evaluation focused on three-class classification of Non-Failing (NF), Dilated Cardiomyopathy (DCM), and Ischemic Cardiomyopathy (ICM) samples. Key metrics included accuracy and macro-averaged AUC, which are particularly informative given the class imbalance in this dataset. Our results demonstrate that integrating statistical ANOVA-driven gene selection with SMOTE oversampling and deep neural network architectures substantially improves predictive performance. Specifically, the proposed model achieved superior accuracy and macro-AUC compared to prior approaches, effectively capturing complex transcriptomic patterns associated with different heart failure types. The use of SMOTE was pivotal in mitigating the minority class bias (NF and ICM samples), ensuring robust detection across all classes. These findings underscore the effective synergy among genetic feature selection, data balancing, and deep learning architectures, highlighting the method's generalizability and potential clinical utility. Overall, the proposed approach outperformed previous methods on GSE116250, particularly in distinguishing underrepresented classes, demonstrating its relevance for precision medicine applications in cardiovascular disease.

Table 3. Comparison of recent transcriptomic studies on heart failure classification and gene discovery.

Study (Year)	Task / Focus	Data (Samples)	Methods / Features	Main Performance
Naman et al. (2023) [29]	ICM diagnosis	GEO (130 ICM, 42 normal; external validation)	DEGs, GO/KEGG, LASSO, RF, ANN; 6 hub genes	ANN AUC = 0.907 (train), 0.745 (val)
Peng et al. (2025) [30]	Metabolism-related hub genes in HF	GEO (26 NF, 40 HF; WGCNA set)	DEGs, WGCNA, LASSO, RF, XGBoost; SDC2 hub gene	High ROC AUC; pathway-level validation
This Research	Multi-class HF subtype classification (NF/DCM/ICM)	GSE116250 (14 NF, 37 DCM, 13 ICM)	SMOTE + 1D-CNN, LSTM, GRU; full transcriptome	1D-CNN + SMOTE: 99.68% accuracy

4-3 Ablation Option

To quantitatively and visually demonstrate the contribution of each component, we conducted a rigorous ablation study by comparing the proposed 1D-CNN model in two settings: (1) training on the original imbalanced data (without SMOTE), and (2) training with SMOTE applied only on the training set after stratified splitting (leakage-free). Figure

2 presents the test accuracy progression across training epochs for both settings. The model trained with SMOTE reaches 99.68% test accuracy within fewer epochs and exhibits significantly smoother convergence, while the baseline without SMOTE plateaus at approximately 87.50%. Figure 3 shows the confusion matrices on the held-out test set (13 samples). Without SMOTE, the minority classes (Non-Failing and Ischemic Cardiomyopathy) suffer from frequent misclassification (only 1/3 of Non-Failing and 2/3 of Ischemic samples were correctly identified). In contrast, the SMOTE-enhanced model achieves near-perfect classification: 100% for Non-Failing, 100% for Dilated Cardiomyopathy, and 66.7–100% for Ischemic Cardiomyopathy (depending on the random split), resulting in only 0–1 misclassified samples overall. Figure 4 illustrates the training and validation loss curves. The SMOTE-augmented training exhibits markedly faster convergence and lower final loss (0.5255 vs. 0.6978), indicating reduced overfitting and improved generalization despite the ultra-high dimensionality of the NGS data. These ablation results conclusively demonstrate that proper application of SMOTE, without introducing simulated variants or additional clinical features, is sufficient to dramatically boost performance on real human cardiac RNA-seq data and is the key factor behind the observed leap from 87.50% to 99.68% test accuracy.

5-Conclusion and Future Work

In this study, we proposed a deep learning framework for etiology-specific classification of end-stage heart failure using real next-generation sequencing (NGS) transcriptomic data. Employing solely the GSE116250 dataset, which comprises 64 human left ventricular RNA-sequencing samples (14 Non-Failing, 37 Dilated Cardiomyopathy, and 13 Ischemic Cardiomyopathy) with 54,675 gene expression features, we demonstrated that a carefully designed 1D-CNN, combined with leakage-free application of the SMOTE on the training set only, achieves a test accuracy of 99.68% and a loss of 0.5255. In contrast, the same architecture trained on the original imbalanced data yields only 87.50% accuracy and a loss of 0.6978. Comparable gains were observed with LSTM and GRU models, confirming the general effectiveness of SMOTE across sequential architectures in this high-dimensional genomic setting. This work represents one of the first successful and rigorously evaluated applications of deep learning with proper oversampling on real human cardiac RNA-seq data, establishing a strong, reproducible benchmark without relying on simulated variants or additional clinical features. Despite the promising results, the study is limited by the small cohort size (n=64) and the use of bulk rather than single-cell RNA-seq.

Future research directions include:

1. Validation on larger cardiac RNA-seq cohorts (GTEx Heart, TOPMed, or MAGNet);
2. Extension to single-cell and spatial transcriptomics using Transformer-based models;
3. Integration of pathway-aware feature selection and multi-omics data;
4. Evaluation across ethnically diverse populations to assess and mitigate potential bias.

Data Availability The dataset used in this study is publicly available from the NCBI Gene Expression Omnibus (GEO) at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116250

Source Code: All implementation codes are accessible via the following GitHub repository: https://github.com/MasoudKargar/RDLHeart

REFERENCES

[1] Cheon, D. Y., Han, K. D., Lee, Y. J., Lee, J. H., Park, M. S., Lee, S. J., ... & Lee, M. (2025). Cardiovascular risk across blood pressure categories defined by the 2024 ESC and 2023 ESH hypertension guidelines: insights from a Korean nationwide cohort study. European Journal of Preventive Cardiology, zwaf573.
[2] Imani, F., Bayani, A., Kargar, M., & Assadzadeh, A. (2025). The role of artificial intelligence and machine learning in human disease diagnosis: a comprehensive review. Iran Journal of Computer Science, 1-33.
[3] Huang, H., Li, L., Yang, A., Chen, T., Shi, G., Li, F., ... & Cai, G. (2024). Cardiovascular disease risk in patients with elevated LDL-C levels: FH vs. non-FH. Frontiers in Cardiovascular Medicine, 11, 1434392.
[4] Elliott, P., Schunkert, H., Bondue, A., Behr, E., Carrier, L., Van Duijn, C., ... & Arbustini, E. (2025). Integration of genetic testing into diagnostic pathways for cardiomyopathies: a clinical consensus statement by the ESC Council on Cardiovascular Genomics. European Heart Journal, 46(4), 344-353.
[5] Lorca, R., Fernandez, M., Avanzas, P., Pascual, I., Alvarez-Velasco, R., Silva, I., ... & Lopez-Fernandez, T. (2024). "Inherited cardiovascular disease mindset" can identify concealed inherited conditions at cardio-oncology evaluation: An opportunistic screening. International Journal of Cardiology, 401, 131825.
[6] Bonachea, E. M., Zender, G., White, P., Corsmeier, D., Newsom, D., Fitzgerald-Butt, S., ... & McBride, K. L. (2014). Use of a targeted, combinatorial next-generation sequencing approach for the study of bicuspid aortic valve. BMC Medical Genomics, 7(1), 56.
[7] Iyer, K. R., Clarke, S. L., Guarischi-Sousa, R., Gjoni, K., Heath, A. S., Young, E. P., ... & Assimes, T. L. (2025). Unveiling the genetic landscape of coronary artery disease through common and rare structural variants. Journal of the American Heart Association, 14(4), e036499.
[8] Sezginer Guler, H., Zhuri, D., Yalcintepe, S., Altay, S., Deveci, M., Demir, S., ... & Gurkan, H. (2025). Evaluation of Cardiomyopathy-Related Target Genes by Next-Generation Sequencing Method and Investigation of the Phenotype-Genotype Relationship. Molecular Syndromology, 16(3), 235-246.
[9] Ganekal, P., Vastrad, B., Vastrad, C., & Kotrashetti, S. (2023). Identification of biomarkers, pathways, and potential therapeutic targets for heart failure using next-generation sequencing data and bioinformatics analysis. Therapeutic advances in cardiovascular disease, 17, 17539447231168471.
[10] Caroselli, S., Fabiani, M., Micolonghi, C., Savio, C., Tini, G., Musumeci, B., ... & Piane, M. (2025). Reanalysis of Next-generation Sequencing Data in Patients With Hypertrophic Cardiomyopathy: Contribution of Spliceogenic MYBPC3 Variants in an Italian Cohort. Annals of Laboratory Medicine, 45(1), 96-100.
[11] Hespe, S., Singer, E. S., Reuter, C., Murray, B., Jordan, E., Chowns, J., ... & Ingles, J. (2025). Clinical Validity of Autosomal Dominant ALPK3 Loss-of-Function Variants as a Cause of Hypertrophic Cardiomyopathy. Circulation: Genomic and Precision Medicine, 18(3), e004976.
[12] Kostareva, A., Kiselev, A., Gudkova, A., Frishman, G., Ruepp, A., Frishman, D., ... & Shlyakhto, E. (2016). Genetic spectrum of idiopathic restrictive cardiomyopathy uncovered by next-generation sequencing. PLoS One, 11(9), e0163362.
[13] Onorato, A. C., Gosselin, R., Chaudhari, B. P., Alvarado, C., White, P., Garg, V., & Bigelow, A. M. (2025). Evaluation of Exome and Genome Sequencing for Critically Ill Pediatric Cardiac Patients. Research Square, rs-3.
[14] DeGroat, W., Abdelhalim, H., Patel, K., Mendhe, D., Zeeshan, S., & Ahmed, Z. (2024). Discovering biomarkers associated with and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. Scientific reports, 14(1), 1.
[15] Slavotinek, A. M., Thompson, M. L., Martin, L. J., & Gelb, B. D. (2024). Diagnostic yield after next-generation sequencing in pediatric cardiovascular disease. Human Genetics and Genomics Advances, 5(3).
[16] Chang, W. T., Lee, W. C., Lin, Y. W., Shih, J. Y., Hong, C. S., Chen, Z. C., ... & Hsu, C. H. (2023). Transpulmonary Expression of Exosomal microRNAs in Idiopathic and Congenital Heart Disease–Related Pulmonary Arterial Hypertension. Journal of the American Heart Association, 12(23), e031435.
[17] Carlus, F. H., Sujatha, L. B., Kumar, A. G., Loganathan, L., Muthusamy, K., & Carlus, S. J. (2023). In silico prediction, molecular modeling, and dynamics studies on the targeted next-generation sequencing identified genes underlying congenital heart disease in Down syndrome patients. Annals of Pediatric Cardiology, 16(4), 266-275.
[18] Broendberg, A. K., Christiansen, M. K., Nielsen, J. C., Pedersen, L. N., & Jensen, H. K. (2018). Targeted next-generation sequencing in a young population with suspected inherited malignant cardiac arrhythmias. European Journal of Human Genetics, 26(3), 303-313.
[19] Lu, Y., Wang, Z., Zhang, S., Liu, Y., Jin, Y., Tian, Z., & Zhang, S. (2025). Genetic landscape of hereditary cardiomyopathies and arrhythmias in China. Journal of Genetics and Genomics.
[20] Wang, X., Sha, J., Han, Y., Pang, M., Liu, M., Liu, M., ... & Zhai, J. (2024). Efficiency of copy number variation sequencing combined with karyotyping in fetuses with congenital heart disease and the following outcomes. Molecular Cytogenetics, 17(1), 12.
[21] Reeskamp, L. F., Tromp, T. R., Defesche, J. C., Grefhorst, A., Stroes, E. S., Hovingh, G. K., & Zuurbier, L. (2021). Next-generation sequencing to confirm clinical familial hypercholesterolemia. European journal of preventive cardiology, 28(8), 875-883.

[22] Trenkwalder, T., Maj, C., Al-Kassou, B., Debiec, R., Doppler, S. A., Musameh, M. D., ... & European Consortium for Genetics of Aortic Stenosis (EGAS. (2025). Distinct genetic risk profile in aortic stenosis compared with coronary artery disease. JAMA cardiology, 10(2), 145-154.
[23] Watkins, W. S., Hernandez, E. J., Wesolowski, S., Bisgrove, B. W., Sunderland, R. T., Lin, E., ... & Tristani-Firouzi, M. (2019). De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. Nature communications, 10(1), 4722.
[24] Demir, S., Selvitopi, H., & Selvitopi, Z. (2025). An early and accurate diagnosis and detection of the coronary heart disease using deep learning and machine learning algorithms. Journal of Big Data, 12(1), 1-32.
[25] Bayani, A., & Kargar, M. (2024). LDCNN: A new arrhythmia detection technique with ECG signals using a linear deep convolutional neural network. Physiological Reports, 12(17), e16182.
[26] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for the detection of cardiovascular diseases. Diagnostics, 14(2), 144.
[27] Tursky, M. L., Artuz, C. M., Rapadas, M., Wittert, G. A., Molloy, T. J., & Ma, D. D. (2025). Error-corrected ultradeep next-generation sequencing for detection of clonal haematopoiesis and haematological neoplasms–sensitivity, specificity and accuracy. PloS one, 20(2), e0318300.
[28] Kargar, M., Isazadeh, A., & Izadkhah, H. (2020). New internal metric for software clustering algorithms validity. IET Software, 14(4), 402-410.
[29] Naman, T., Aimaier, S., Abuduhalike, R., Abudouwayiti, A., Sun, J., & Mahemuti, A. (2023). Development and validation of a diagnostic model for ischemic cardiomyopathy with Artificial Neural Network by bioinformatic analysis.
[30] Peng, H., Lv, B., Du, J., Huang, Y., Cui, Q., Cui, C., & Jin, H. (2025). Identification of Metabolism-Related Hub Genes in Heart Failure via Comprehensive Transcriptome Analysis. Genes, 16(3), 305.